Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Improvements in Speech Synthesis

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Improvements in Speech Synthesis

COST 258: The Naturalness of Synthetic Speech

Edited by
E. Keller,
University of Lausanne, Switzerland
G. Bailly,
INPG, France
A. Monaghan,
Aculab plc, UK
J. Terken,
Technische Universiteit Eindhoven, The Netherlands
M. Huckvale,
University College London, UK

JOHN WILEY & SONS, LTD

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Copyright © 2002 by John Wiley & Sons, Ltd
Baffins Lane, Chichester,
West Sussex, PO19 1UD, England

National 01243 779777
International (+44) 1243 779777
e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk

Visit our Home Page on http://www.wiley.co.uk or http://www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency, 90 Tottenham Court Road, London, W1P 9HE, UK, without the permission in writing of the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the publication.

Neither the author(s) nor John Wiley and Sons Ltd accept any responsibility or liability for loss or damage occasioned to any person or property through using the material, instructions, methods or ideas contained herein, or acting or refraining from acting as a result of such use. The author(s) and Publisher expressly disclaim all implied warranties, including merchantability of fitness for any particular purpose.

Designations used by companies to distinguish their products are often claimed as trademarks. In all instances where John Wiley and Sons is aware of a claim, the product names appear in initial capital or capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 605 Third Avenue,
New York, NY 10158–0012, USA

WILEY-VCH Verlag GmbH
Pappelallee 3, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton,
Queensland 4064, Australia

John Wiley & Sons (Canada) Ltd, 22 Worcester Road
Rexdale, Ontario, M9W 1L1, Canada

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02–01,
Jin Xing Distripark, Singapore 129809

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0471 49985 4

Typeset in 10/12pt Times by Kolam Information Services Ltd, Pondicherry, India.
Printed and bound in Great Britain by Biddles Ltd, Guildford and King's Lynn.
This book is printed on acid-free paper responsibly manufactured from sustainable forestry, in which at least two trees are planted for each one used for paper production.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Contents

List of contributors **ix**

Preface **xiii**

Part I Issues in Signal Generation **1**

1 Towards Greater Naturalness: Future Directions of Research in Speech Synthesis 3
Eric Keller

2 Towards More Versatile Signal Generation Systems 18
Gérard Bailly

3 A Parametric Harmonic + Noise Model 22
Gérard Bailly

4 The COST 258 Signal Generation Test Array 39
Gérard Bailly

5 Concatenative Text-to-Speech Synthesis Based on Sinusoidal Modelling 52
Eduardo Rodríguez Banga, Carmen García Mateo and Xavier Fernández Salgado

6 Shape Invariant Pitch and Time-Scale Modification of Speech Based on a Harmonic Model 64
Darragh O'Brien and Alex Monaghan

7 Concatenative Speech Synthesis Using SRELP 76
Erhard Rank

Part II Issues in Prosody **87**

8 Prosody in Synthetic Speech: Problems, Solutions and Challenges 89
Alex Monaghan

9 State-of-the-Art Summary of European Synthetic Prosody R&D 93
Alex Monaghan

10 Modelling FO in Various Romance Languages: Implementation in Some TTS Systems 104
Philippe Martin

11 Acoustic Characterisation of the Tonic Syllable in Portuguese 120
João Paulo Ramos Teixeira and Diamantino R.S. Freitas

12 Prosodic Parameters of Synthetic Czech: Developing Rules for Duration and Intensity 129
Marie Dohalská, Jana Mejvaldová and Tomas Duběda

13 MFGI, a Linguistically Motivated Quantitative Model of German Prosody 134
Hansjörg Mixdorff
14 Improvements in Modelling the FO Contour for Different Types of Intonation Units in Slovene 144
Ales Dobnikar
15 Representing Speech Rhythm 154
Brigitte Zellner Keller and Eric Keller
16 Phonetic and Timing Considerations in a Swiss High German TTS System 165
Beat Siebenhaar, Brigitte Zellner Keller and Eric Keller
17 Corpus-based Development of Prosodic Models Across Six Languages 176
Justin Fackrell, Halewijn Vereecken, Cynthia Grover, Jean-Pierre Martens and Bert Van Coile
18 Vowel Reduction in German Read Speech 186
Christina Widera

Part III Issues in Styles of Speech 197

19 Variability and Speaking Styles in Speech Synthesis 199
Jacques Terken
20 An Auditory Analysis of the Prosody of Fast and Slow Speech Styles in English, Dutch and German 204
Alex Monaghan
21 Automatic Prosody Modelling of Galician and its Application to Spanish 218
Eduardo López Gonzalo, Juan M. Villar Navarro and Luis A. Hernández Gómez
22 Reduction and Assimilatory Processes in Conversational French Speech: Implications for Speech Synthesis 228
Danielle Duez
23 Acoustic Patterns of Emotions 237
Branka Zei Pollermann and Marc Archinard
24 The Role of Pitch and Tempo in Spanish Emotional Speech: Towards Concatenative Synthesis 246
Juan Manuel Montero Martinez, Juana M. Gutiérrez Arriola, Ricardo de Córdoba Herralde, Emilia Victoria Enríquez Carrasco and José Manuel Pardo Muñoz
25 Voice Quality and the Synthesis of Affect 252
Ailbhe Ní Chasaide and Christer Gobl
26 Prosodic Parameters of a 'Fun' Speaking Style 264
Kjell Gustafson and David House
27 Dynamics of the Glottal Source Signal: Implications for Naturalness in Speech Synthesis 273
Christer Gobl and Ailbhe Ní Chasaide
28 A Nonlinear Rhythmic Component in Various Styles of Speech 284
Brigitte Zellner Keller and Eric Keller

Part IV Issues in Segmentation and Mark-up **293**

29 Issues in Segmentation and Mark-up 295
Mark Huckvale
30 The Use and Potential of Extensible Mark-up (XML) in Speech Generation 297
Mark Huckvale
31 Mark-up for Speech Synthesis: A Review and Some Suggestions 307
Alex Monaghan
32 Automatic Analysis of Prosody for Multi-lingual Speech Corpora 320
Daniel Hirst
33 Automatic Speech Segmentation Based on Alignment with a Text-to-Speech System 328
Petr Horák
34 Using the COST 249 Reference Speech Recogniser for Automatic Speech Segmentation 339
Narada D. Warakagoda and Jon E. Natvig

Part V Future Challenges **349**

35 Future Challenges 351
Eric Keller
36 Towards Naturalness, or the Challenge of Subjectiveness 353
Geneviève Caelen-Haumont
37 Synthesis Within Multi-Modal Systems 363
Andrew Breen
38 A Multi-Modal Speech Synthesis Tool Applied to Audio-Visual Prosody 372
Jonas Beskow, Björn Granström and David House
39 Interface Design for Speech Synthesis Systems 383
Gudrun Flach

Index **391**

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

List of contributors

Marc Archinard
Geneva University Hospitals
Liaison Psychiatry
Boulevard de la Cluse 51
1205 Geneva, Switzerland

Gérard Bailly
Institut de la Communication Parlée
INPG
46 av. Felix Vialet
38031 Grenoble-cedex, France

Eduardo Rodríguez Banga
Signal Theory Group (GTS).
Dpto. Tecnologías de las
Comunicaciones.
ETSI Telecomunicación
Universidad de Vigo
36200 Vigo, Spain

Jonas Beskow
CTT/Dept. of Speech, Music and
Hearing
KTH
100 44 Stockholm, Sweden

Andrew Breen
Nuance Communications Inc.
The School of Information Systems
University of East Anglia
Norwich, NR4 7TJ, United Kingdom

Geneviève Caelen-Haumont
Laboratoire Parole et Langage
CNRS
Université de Provence
29 Av. Robert Schuman
13621 Aix en Provence, France

Ricardo de Córdoba Herralde
Universidad Politécnica de Madrid
ETSI Telecomunicación
Ciudad Universitaria s/n
28040 Madrid, Spain

Ales Dobnikar
Institute J. Stefan
Jamova 39
1000 Ljubljana, Slovenia

Marie Dohalská
Institute of Phonetics
Charles University, Prague
nam. Jana Palacha 2
116 38 Prague 1, Czech Republic

Tomas Dubeda
Institute of Phonetics
Charles University, Prague
nam. Jana Palacha 2
116 38 Prague 1, Czech Republic

Danielle Duez
Laboratoire Parole et Langage
CNRS
Université de Provence
29 Av. Robert Schuman
13621 Aix en Provence, France

Emilia Victoria Enríquez Carrasco
Facultad de Filología. UNED
C/ Senda del Rey 7
28040 Madrid, Spain

Justin Fackrell
Crichton's Close
Canongate
Edinburgh EH8 8DT
UK

Xavier Fernández Salgado
Signal Theory Group (GTS)
Dpto. Tecnologías de las Comunicaciones
ETSI Telecomunicación
Universidad de Vigo
36200 Vigo, Spain

Gudrun Flach
Dresden University of Technology
Laboratory of Acoustics and Speech Communication
Mommsenstr. 13
01069 Dresden, Germany

Diamantino R.S. Freitas
Fac. de Eng. da Universidade do Porto
Rua Dr Roberto Frias
4200 Porto, Portugal

Carmen García Mateo
Signal Theory Group (GTS)
Dpto. Tecnologías de las Comunicaciones
ETSI Telecomunicación
Universidad de Vigo
36200 Vigo, Spain

Christer Gobl
Centre for Language and Communication Studies
Arts Building,
Trinity College
Dublin 2, Ireland

Björn Granström
CTT/Dept. of Speech, Music and Hearing
KTH
100 44 Stockholm,
Sweden

Cynthia Grover
Belgacom Towers
Koning Albert II Iaan 27
1030 Brussels, Belgium

Kjell Gustafson
CTT/Dept. of Speech, Music and Hearing
KTH
100 44 Stockholm, Sweden

Juana M. Gutiérrez Arriola
Universidad Politécnica de Madrid
ETSI Telecomunicación
Ciudad Universitaria s/n
28040 Madrid, Spain

Luis A. Hernández Gómez
ETSI Telecommunicación
Ciudad Universitaria s/n
28040 Madrid, Spain

Daniel Hirst
Laboratoire Parole et Langage
CNRS
Université de Provence
29 Av. Robert Schuman
13621 Aix en Provence, France

Petr Horák
Institute of Radio Engineering and Electronics
Academy of Sciences of the Czech Republic
Chaberská 57
182 51 Praha 8 – Kobylisy,
Czech Republic

David House
CTT/Dept. of Speech, Music and Hearing
KTH
100 44 Stockholm, Sweden

Mark Huckvale
Phonetics and Linguistics
University College London
Gower Street
London WC1E 6BT,
United Kingdom

Eric Keller
LAIP-IMM-Lettres
Université de Lausanne
1015 Lausanne, Switzerland

Eduardo López Gonzalo
ETSI Telecommunicación
Ciudad Universitaria s/n
28040 Madrid, Spain

Jean-Pierre Martens
ELIS
Ghent University
Sint-Pietersnieuwstraat 41
9000 Gent, Belgium

Philippe Martin
University of Toronto
77A Lowther Ave
Toronto, ONT
Canada M5R IC9

Jana Mejvaldová
Institute of Phonetics
Charles University, Prague
nam. Jana Palacha 2
116 38 Prague 1, Czech Republic

Hansjörg Mixdorff
Dresden University of Technology
Hilbertstr. 21
12307 Berlin, Germany

Alex Monaghan
Aculab plc
Lakeside
Bramley Road
Mount Farm
Milton Keynes MK1 1PT,
United Kingdom

Juan Manuel Montero Martínez
Universidad Politécnica de Madrid
ETSI Telecomunicación
Ciudad Universitaria s/n
28040 Madrid, Spain

Jon E. Natvig
Telenor Research and Development
P.O. Box 83
2027 Kjeller, Norway

Ailbhe Ní Chasaide
Phonetics and Speech Laboratory
Centre for Language and
Communication Studies
Trinity College
Dublin 2, Ireland

Darragh O'Brien
11 Lorcan Villas
Santry
Dublin 9, Ireland

José Manuel Pardo Muñoz
Universidad Politécnica de Madrid
ETSI Telecomunicación
Ciudad Universitaria s/n
28040 Madrid, Spain

Erhard Rank
Institute of Communications
and Radio-frequency Engineering
Vienna University of Technology
Gusshausstrasse 25/E389
1040 Vienna, Austria

Beat Siebenhaar
LAIP-IMM-Lettres
Université de Lausanne
1015 Lausanne, Switzerland

João Paulo Ramos Teixeira
ESTG-IPB
Campus de Santa Apolónia
Apartado 38
5301–854 Bragança, Portugal

Jacques Terken
Technische Universiteit Eindhoven
IPO, Center for User-System
Interaction
P.O. Box 513
5600 MB Eindhoven,
The Netherlands

Bert Van Coile
L&H
FLV 50
8900 Ieper, Belgium

Halewijn Vereecken
Collegiebaan 29/11
9230 Wetteren, Belgium

Juan M. Villar Navarro
ETSI Telecomunicación
Ciudad Universitaria s/n
28040 Madrid, Spain

Narada D. Warakagoda
Telenor Research and Development
P.O. Box 83
2027 Kjeller, Norway

Christina Widera
Institut für
Kommunikationsforschung und
Phonetik
Universität of Bonn
Poppelsdorfer Allee 47
53115 Bonn, Germany

Branka Zei Pollermann
Geneva University Hospitals
Liaison Psychiatry
Boulevard de la Cluse 51
1205 Geneva, Switzerland

Brigitte Zellner Keller
LAIP-IMM-Lettres
Université de Lausanne
1015 Lausanne, Switzerland

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Preface

Making machines speak like humans is a dream that is slowly coming to fruition. When the first automatic computer voices emerged from their laboratories twenty years ago, their robotic sound quality severly curtailed their general use. But now after a long period of maturation, synthetic speech is beginning to reach an initial level of acceptability. Some systems are so good that one even wonders if the recording was authentic or manufactured.

The effort to get to this point has been considerable. A variety of quite different technologies had to be developed, perfected and examined in depth, requiring skills and interdisciplinary efforts in mathematics, signal processing, linguistics, statistics, phonetics and several other fields. The current compendium in research on speech synthesis is quite representative of this effort, in that it presents work in signal processing as well as in linguistics and the phonetic sciences, performed with the explicit goal of arriving at a greater degree of *naturalness* in synthesised speech.

But more than just describing the status quo, the current volume points the way to the future. The researchers assembled here generally concur that the current, increasingly healthy state of speech synthesis is by no means the *end* of a technological development, much rather that it is an excellent *starting point*. A great deal more work is still needed to bring about much greater variety and flexibility to our synthetic voices, so that they can be used in a much wider set of everyday applications. That is what the current volume traces out in some detail.

Work in signal processing is perhaps the most crucial for the further success of speech synthesis, since it lays the theoretical and technological foundation for developments to come. But right behind follows more extensive research on prosody and styles of speech, work which will trace out the types of voices that will be appropriate to a variety of contexts. And finally, work on the increasingly standardised user interfaces in the form of system options and text mark-up is making it possible to open speech synthesis to a wide variety of non-specialist users.

The research published here emerges from the four-year European COST 258 project which has served primarily to assemble the authors of this volume in a set of twice-yearly meetings from 1997 to 2001. The value of these meetings can hardly be underestimated. 'Trial balloons' could be launched within an encouraging smaller circle, well before they were presented to highly critical international congresses. Informal off-podium contacts furnished crucial information on what works and does not work in speech synthesis. And many fruitful associations between research teams were formed and strengthened in this context. This is the rich texture of scientific and human interactions from which progress has emerged and future realisations are likely to grow. As chairman and secretary of this COST

project, we wish to thank all our colleagues for the exceptional experience that has made this volume possible.

Eric Keller and Brigitte Zellner Keller
University of Lausanne, Switzerland
October, 2001

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Part I

Issues in Signal Generation

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

1

Towards Greater Naturalness

Future Directions of Research in Speech Synthesis

Eric Keller
Laboratoire d'analyse informatique de la parole (LAIP)
IMM-Lettres, University of Lausanne, 1015 Lausanne, Switzerland
Eric.Keller@imm.unil.ch

Introduction

In the past ten years, many speech synthesis systems have shown remarkable improvements in quality. Instead of monotonous, incoherent and mechanical-sounding speech utterances, these systems produce output that sounds relatively close to human speech. To the ear, two elements contributing to the improvement stand out, improvements in signal quality, on the one hand, and improvements in coherence and naturalness, on the other. These elements reflect, in fact, two major technological changes. The improvements in signal quality of good contemporary systems are mainly due to the use and improved control over concatenative speech technology, while the greater coherence and naturalness of synthetic speech are primarily a function of much improved prosodic modelling.

However, as good as some of the best systems sound today, few listeners are fooled into believing that they hear human speakers. Even when the simulation is very good, it is still not perfect – no matter how one wishes to look at the issue. Given the massive research and financial investment from which speech synthesis has profited over the years, this general observation evokes some exasperation. The holy grail of 'true naturalness' in synthetic speech seems so near, and yet so elusive. What in the world could still be missing?

As so often, the answer is complex. The present volume introduces and discusses a great variety of issues affecting naturalness in synthetic speech. In fact, at one level or another, it is probably true that most research in speech synthesis today deals with this very issue. To start the discussion, this article presents a personal view of recent encouraging developments and continued frustrating limitations of

current systems. This in turn will lead to a description of the research challenges to be confronted over the coming years.

Current Status

Signal Quality and the Move to Time-Domain Concatenative Speech Synthesis

The first generation of speech synthesis devices capable of unlimited speech (KlattTalk, DEC-Talk, or early InfoVox synthesisers) used a technology called 'formant synthesis' (Klatt, 1989; Klatt and Klatt, 1990; Styger and Keller, 1994). While speech produced by formant synthesis produced the classic 'robotic' style of speech, formant synthesis was also a remarkable technological development that has had some long-lasting effects. In this approach, *voiced* speech sounds are created much as one would create a sculpture from stone or wood: a complex waveform of harmonic frequencies is created first, and 'the parts that are too much', i.e. non-formant frequencies, are suppressed by filtering. For *unvoiced* or *partially voiced* sounds, various types of noise are created, or are mixed in with the voiced signal. In formant synthesis, speech sounds are thus created entirely from equations. Although obviously modelled on actual speakers, a formant synthesiser is not tied to a single voice. It can be induced to produce a great variety of voices (male, female, young, old, hoarse, etc).

However, this approach also posed several difficulties, the main one being that of excessive complexity. Although theoretically capable of producing close to human-like speech under the best of circumstances (YorkTalk a–c, Webpage), these devices must be fed a complex and coherent set of parameters every 2–10 ms. Speech degrades rapidly if the coherence between the parameters is disrupted. Some coherence constraints are given by mathematical relations resulting from vocal tract size relationships, and can be enforced automatically via algorithms developed by Stevens and his colleagues (Stevens, 1998). But others are language- and speaker-specific and are more difficult to identify, implement, and enforce automatically. For this reason, really good-sounding synthetic speech has, to my knowledge, never been produced entirely automatically with formant synthesis.

The apparent solution for these problems has been the general transition to 'time-domain concatenative speech synthesis' (TD-synthesis). In this approach, large databases are collected, and constituent speech portions (segments, syllables, words, and phrases) are identified. During the synthesis phase, designated signal portions (diphones, polyphones, or even whole phrases[1]) are retrieved from the database according to phonological selection criteria ('unit selection'), chained together ('concatenation'), and modified for timing and melody ('prosodic modification'). Because such speech portions are basically stored and minimally modified

[1] A diphone extends generally from the middle of one sound to the middle of the next. A polyphone can span larger groups of sounds, e.g., consonant clusters. Other frequent configurations are demi-syllables, tri-phones and 'largest possible sound sequences' (Bhaskararao, 1994). Another important configuration is the construction of carrier sentences with 'holes' for names and numbers, used in announcements for train and airline departures and arrivals.

segments of human speech, TD-generated speech consists by definition only of possible human speech sounds, which in addition preserve the personal characteristics of a specific speaker. This accounts, by and large, for the improved signal quality of current TD speech synthesis.

Prosodic Quality and the Move to Stochastic Models

The second major factor in recent improvements of speech synthesis quality has been the refinement of prosodic models (see Chapter 9 by Monaghan, this volume, plus further contributions found in the prosody section of this volume). Such models tend to fall into two categories: predominantly linguistic and predominantly empirical-statistic ('stochastic'). For many languages, early linguistically inspired models did not furnish satisfactory results, since they were incapable of providing credible predictive timing schemas or the full texture of a melodic line. The reasons for these insufficiencies are complex. Our own writings have criticised the exclusive dependence on phonosyntax for the prediction of major and minor phrase boundaries, the difficulty of recreating specific Hertz values for the fundamental frequency ('melody', abbr. F0) on the basis of distinctive features, and the strong dependence on the notion of 'accent' in languages like French where accents are not reliably defined (Zellner, 1996, 1998a; Keller *et al*., 1997).

As a consequence of these inadequacies, so-called 'stochastic' models have moved into the dominant position among high-quality speech synthesis devices. These generally implement either an array or a tree structure of predictive parameters and derive statistical predictors for timing and F0 from extensive database material. The prediction parameters do not change a great deal from language to language. They generally concern the position in the syllable, word and phrase, the sounds making up a syllable, the preceding and following sounds, and the syntactic and lexical status of the word (e.g., Keller and Zellner, 1996; Zellner Keller and Keller, in press). Models diverge primarily with respect to the quantitative approach employed (e.g., artificial neural network, classification and regression tree, sum-of-products model, general linear model; Campbell, 1992b; Riley, 1992; Keller and Zellner, 1996; Zellner Keller and Keller, Chapters 15 and 28, this volume), and the logic underlying the tree structure.

While stochastic models have brought remarkable improvements in the refinement of control over prosodic parameters, they have their own limitations and failures. One notable limit is rooted in the 'sparse data problem' (van Santen and Shih, 2000). That is, some of the predictive parameters occur a great deal less frequently than others, which makes it difficult to gather enough material to estimate their influence in an overall predictive scheme. Consequently a predicted melodic or timing parameter may be 'quite out of line' every once in a while. A second facet of the same sparse data problem is seen in parameter interactions. While the effects of most predictive parameters is approximatively cumulative, a few parameter combinations show unusually strong interaction effects. These are often difficult to estimate, since the contributing parameters are so rare and enter into interactions even less frequently. On the whole, 'sparse data' problems are solved in either a 'brute force' approach (gather more data, *much* more), by careful analysis of data (e.g., establish sound groups, rather than model sounds

individually), and/or by resorting to a set of supplementary rules that 'fix' some of the more obvious errors induced by stochastic modelling.

A further notable limit of stochastic models is their averaging tendency, well illustrated by the problem of modelling F0 at the end of sentences. In many languages, questions can end on either a higher or a lower F0 value than that used in a declarative sentence (as in 'is *that* what you mean?'). If high-F0 sentences are not rigorously, perhaps manually, separated from low-F0 sentences, the resulting statistical predictor value will tend towards a mid-F0 value, which is obviously wrong. A fairly obvious example was chosen here, but the problem is pervasive and must be guarded against throughout the modelling effort.

The Contribution of Timing

Another important contributor to greater prosodic quality has been the improvement of the prediction of timing. Whereas early timing models were based on simple average values for different types of phonetic segments, current synthesis systems tend to resort to fairly complex stochastic modelling of multiple levels of timing control (Campbell, 1992a, 1992b; Keller and Zellner, 1996; Zellner 1996, 1998a, b).

Developing timing control that is precise as well as adequate to all possible speech conditions is rather challenging. In our own adjustments of timing in a French synthesis system, we have found that changes in certain vowel durations as small as 2% can induce audible improvements or degradations in sound quality, particularly when judged over longer passages. Further notable improvements in the perceptual quality of prosody can be obtained by a careful analysis of links between timing and F0. Prosody only sounds 'just right' when F0 peaks occur at expected places in the vowel. Also of importance is the order and degree of interaction that is modelled between timing and F0. Although the question of whether timing or F0 modelling should come first has apparently never been investigated systematically, our own experiments have suggested that timing feeding into F0 gives considerably better results than the inverse (Zellner, 1998a; Keller *et al.*, 1997; Siebenhaar *et al.*, chapter 16, this volume). This modelling arrangement permits timing to influence a number of F0 parameters, including F0 peak width in slow and fast speech modes.

Upstream, timing is strongly influenced by phrasing, or the way an utterance is broken up into groups of words. Most traditional speech synthesis devices were primarily guided by phonosyntactic principles in this respect. However, in our laboratory, we have found that psycholinguistically driven dependency trees oriented towards actual human speech behaviour seem to perform better in timing than dependency trees derived from phonosyntactic principles (Zellner, 1997). That is, our timing improves if we attempt to model the way speakers tend to group words in their real-time speech behaviour. In our modelling of French timing, a relatively simple, psycholinguistically motivated phrasing ('chunking') principle has turned out to be a credible predictor of temporal structures even when varying speech rate (Keller *et al.*, 1993; Keller and Zellner, 1996). Recent research has shown that this is not a peculiarity of our work on French, because similar results have also been obtained with German (Siebenhaar *et al.*, chapter 16, this volume).

To sum up recent developments in signal quality and prosodic modelling, it can be said that a typical contemporary high-quality system tends to be a TD-synthesis system incorporating a series of fairly sophisticated stochastic models for timing and melody, and less frequently, one for amplitude. Not surprisingly, better quality has led to a much wider use of speech synthesis, which is illustrated in the next section.

Uses for High-Quality Speech Synthesis

Given the robot-like quality of early forms of speech synthesis, the traditional application for speech synthesis has been the simulation of a 'serious and responsible speaker' in various virtual environments (e.g., a reader for the visually handicapped, for remote reading of email, product descriptions, weather reports, stock market quotations, etc.). However, the quality of today's best synthesis systems broadens the possible applications of this technology. With sufficient naturalness, one can imagine automated news readers in virtual radio stations, salesmen in virtual stores, or speakers of extinct and recreated languages.

High-quality synthesis systems can also be used in places that were not considered before, such as assisting language teachers in certain language learning exercises. Passages can be presented as frequently as desired, and sound examples can be made up that could not be produced by a human being (e.g., speech with intonation, but no rhythm), permitting the training of prosodic and articulatory competence. Speech synthesisers can slow down stretches of speech to ease familiarisation and articulatory training with novel sound sequences (LAIPTTS a, b, Webpage[2]). Advanced learners can experiment with the accelerated reproduction speeds used by the visually handicapped for scanning texts (LAIPTTS c, d, Webpage). Another obvious second-language application area is listening comprehension, where a speech synthesis system acts as an 'indefatigable substitute native speaker' available 24 hours a day, anywhere in the world.

A high-quality speech synthesis could further be used for literacy training. Since illiteracy has stigmatising status in our societies, a computer can profit from the fact that it is *not* a human, and is thus likely to be perceived as non-judgemental and neutral by learners. In addition, speech synthesis could become a useful tool for linguistic and psycholinguistic experimentation. Knowledge from selected and diverse levels (phonetic, phonological, prosodic, lexical, etc.) can be simulated to verify the relevance of type of knowledge individually and interactively. Already now, speech synthesis systems can be used to experiment with rhythm and pitch patterns, the placement of major and minor phrase boundaries, and typical phonological patterns in a language (LAIPTTS e, f, i–l, Webpage). Finally, speech synthesis increasingly serves as a computer tool. Like dictionaries, grammars (correctors) and translation systems, speech synthesisers are finding a natural place on computers. Particularly when the language competence of a synthesis system begins to outstrip that of some of the better second language users, such systems become useful new adjunct tools.

[2] LAIPTTS is the speech synthesis system of the author's laboratory (LAIPTTS-F for French, LAIPTTS-D for German).

Limits of Current Systems

But rising expectations induced by a wider use of improved speech synthesis systems also serve to illustrate the failings and limitations of contemporary systems. Current top systems for the world's major languages not only tend to make some glaring errors, they are also severely limited with respect to styles of speech and number of voices. Typical contemporary systems offer perhaps a few voices, and they produce essentially a single style of speech (usually a neutral-sounding 'news-reading style'). Contrast that with a typical human community of speakers, which incorporates an enormous variety of voices and a considerable gamut of distinct speech styles, appropriate to the innumerable facets of human language interaction. While errors can ultimately be eliminated by better programming and the marking up of input text, insufficiencies in voice and style variety are much harder problems to solve.

This is best illustrated with a concrete example. When changing speech style, speakers tend to change timing. Since many timing changes are non-linear, they cannot be easily predicted from current models. Our own timing model for French, for example, is based on laboratory recordings of a native speaker of French, reading a long series of French sentences – in excess of 10 000 manually measured segments. Speech driven by this model is credible and can be useful for a variety of purposes. However, this timing style is quite different from that of a well-known French newscaster recorded in an actual TV newscast. Sound example TV_BerlinOrig.wav is a short portion taken from a French TV newscast of January 1998, and LAIPTTS h, Webpage, illustrates the reading of the same text with our speech synthesis system. Analysis of the example showed that the two renderings differ primarily with respect to timing, and that the newscaster's temporal structure could not easily be derived from our timing model.[3] Consequently, in order to produce a timing model for this newscaster, a large portion of the study underlying the original timing model would probably have to be redone (i.e., another 10 000 segments to measure, and another statistical model to build).

This raises the question of how many *speech styles* are required in the absolute. A consideration of the most common style-determining factors indicates that it must be quite a few (Table 1.1). The total derived from this list is 180 (4*5*3*3) theoretically possible styles. It is true that the Table 1.1 is only indicative: there is as yet no unanimity on the definition of 'style of speech' or its 'active parameters' (see the discussion of this issue by Terken, Chapter 19, this volume). Also some styles could probably be modelled as variants of other styles, and some parameter combinations are impossible or unlikely (a spelled, commanding presentation of questions, for example). While some initial steps towards expanded styles of speech are currently being pioneered (see the articles in this volume in Part III), it remains

[3] Interestingly, a speech stretch recreated on the basis of the natural timing measures, but implementing our own melodic model, was auditorily much closer to the original (LAIPTTS g, Webpage). This illustrates a number of points to us: first, that the modelling of timing and fundamental frequencies are largely independent of each other, second, that the modelling of timing should probably precede the modelling of F0 as we have argued, and third, that our stochastically derived F0 model is not unrealistic.

Table 1.1 Theoretically possible styles of speech

Parameter	Instantiations	N
Speech rate	spelled, deliberate, normal, fast	4
Type of speech	spontaneous, prepared oral, command, dialogue, multilogue, reading	5
Material-related	continuous text, lists, questions (perhaps more)	3
Dialect	(dependent on language and grain of analysis)	3

true that only very few of all possible human speech styles are supported by current speech synthesis systems.

Emotional and expressive speech constitutes another evident gap for current systems, despite a considerable theoretical effort currently directed at the question (Ní Chasaide and Gobl, Chapter 25, this volume; Zei and Archinard, Chapter 23, this volume; ISCA workshop, www.qub.ac.uk/en/isca/index.htm). The lack of general availability of emotional variables prevents systems from being put to use in animation, automatic dubbing, virtual theatre, etc. It may be asked how many voices would theoretically be desirable. Table 1.2 shows a list of factors that are known to, or can conceivably influence, voice quality. Again, this list is likely to be incomplete and not all theoretical combinations are possible (it is difficult to conceive of a toddler, speaking in commanding fashion on a satellite hook-up, for example). But even without entering into discussions of granularity of analysis and combinatorial possibility, it is evident that there is an enormous gap between the few synthetic voices available now, and the half million or so (10*5*11*6*6*7*4) theoretically possible voices listed in Table 1.2.

Table 1.2 Theoretically possible voices

Parameter	Instantiations	N
Age	infant, toddler, young child, older child, adolescent, young adult, middle-aged adult, mature adult, fit older adult, senescent adult	10
Gender	very male (long vocal tract), male (shorter vocal tract), difficult-to-tell (medium vocal tract), female (short vocal tract), very female (very short vocal tract)	5
Psychological disposition	sleepy-voiced, very calm, calm-and-in-control, alert, questioning, interested, commanding, alarmed, stressed, in distress, elated	11
Degree of formality	familiar, amicable, friendly, stand-offish, formal, distant	6
Size of audience	alone, one person, two persons, small group, large group, huge audience	6
Type of communication	visual – close up, visual – some distance, visual – great distance, visual – teleconferencing, audio – good connection, audio – bad connection, delayed feedback (satellite hook-ups)	7
Communicative context	totally quiet, some background noise, noisy, very noisy	4

Impediments to New Styles and New Voices

We must conclude from this that our current technology provides clearly too few styles of speech and too few voices and voice timbres. The reason behind this deficiency can be found in a central characteristic of TD-synthesis. It will be recalled that this type of synthesis is not much more than a smartly selected, adaptively chained and prosodically modified rendering of pre-recorded speech segments. By definition, any new segment appearing in the synthetic speech chain must initially be placed into the stimulus material, and must be recorded and stored away before it can be used.

It is this encoding requirement that limits the current availability of styles and voices. Every new style and every new voice must be stored away as a full sound database before it can be used, and a 'full sound database' is minimally constituted of all sound transitions of the language (diphones, polyphones, etc.). In French, there are some 2 000 possible diphones, in German there are around 7 500 diphones, if differences between accented/unaccented and long/short variants of vowels are taken into account. This leads to serious storage and workload problems. If a typical French diphone database is 5 Mb, DBs for 'just' 100 styles and 10 000 voices would require (100*10 000*5) 5 million Mb, or 5 000 Gb. For German, storage requirements would double. The work required to generate all these databases in the contemporary fashion is just as gargantuan. Under favourable circumstances, a well-equipped speech synthesis team can generate an entirely new voice or a new style in a few weeks. The processing of the database itself only takes a few minutes, through the use of automatic speech recognition and segmentation tools. Most of the encoding time goes into developing the initial stimulus material, and into training the automatic segmentation device.

And there in lies the problem. For many styles and voices, the preparation phase is likely to be much more work than supporters of this approach would like to admit. Consider, for example, that some speech rate manipulations give totally new sound transitions that must be foreseen as a full co-articulatory series in the stimulus materials (i.e., the transition in question should be furnished in all possible left and right phonological contexts). For example, there are the following features to consider:

- reductions, contractions and agglomerations. In rapidly pronounced French, for example, the sequence 'l'intention *d'*allumer' can be rendered as /*n*alyme/, or 'pe*nd*ant' can be pronounced /pã*nd*ã/ instead of /pã*nd*ã/ (Duez, Chapter 22, this volume). Detailed auditory and spectrographic analyses have shown that transitions involving partially reduced sequences like /nd/ cannot simply be approximated with fully reduced variants (e.g., /n/). In the context of a high-quality synthesis, the human ear can tell the difference (Local, 1994). Consequently, contextually complete series of stimuli must be foreseen for transitions involving /nd/ and similarly reduced sequences.
- systematic non-linguistic sounds produced in association with linguistic activity. For example, the glottal stop can be used systematically to ask for a turn (Local, 1997). Such uses of the glottal stop and other non-linguistic sounds are not generally encoded into contemporary synthesis databases, but must be planned for inclusion in the next generation of high-quality system databases.

- freely occurring variants: '*of the* time' can be pronounced /əvðətajm/, /əvətajm/, /əvðətajm/, or /ənətajm/ (Ogden *et al.*, 1999). These variants, of which there are quite a few in informal language, pose particular problems to automatic recognition systems due to the lack of a one-to-one correspondence between the articulation and the graphemic equivalent. Specific measures must be taken to accommodate this variation.
- dialectal variants of the sound inventory. Some dialectal variants of French, for example, systematically distinguish between the initial sound found in '*un* signe' (a sign) and '*in*signe' (badge), while other variants, such as the French spoken by most young Parisians, do not. Since this modifies the sound inventory, it also introduces major modifications into the initial stimulus material.

None of these problems is extraordinarily difficult to solve by itself. The problem is that special case handling must be programmed for many different phonetic contexts, and that such handling can change from style to style and from voice to voice. This brings about the true complexity of the problem, particularly in the context of full, high-quality databases for several hundred styles, several hundred languages, and many thousands of different voice timbres.

Automatic Processing as a Solution

Confronted with these problems, many researchers appear to place their full faith in automatic processing solutions. In many of the world's top laboratories, stimulus material is no longer being carefully prepared for a scripted recording session. Instead, hours of relatively naturally produced speech are recorded, segmented and analysed with automatic recognition algorithms. The results are down-streamed automatically into massive speech synthesis databases, before being used for speech output. This approach follows the argument that: 'If a child can learn speech by automatic extraction of speech features from the surrounding speech material, a well-constructed neural network or hidden Markov model should be able to do the same.'

The main problem with this approach is the cross-referencing problem. Natural language studies and psycholinguistic research indicate that in learning speech, humans cross-reference spoken material with semantic references. This takes the form of a complex set of relations between heard sound sequences, spoken sound sequences, structural regularities, semantic and pragmatic contexts, and a whole network of semantic references (see also the subjective dimension of speech described by Caelen-Haumont, Chapter 36, this volume). It is this complex network of relations that permits us to identify, analyse, and understand speech signal portions in reference to previously heard material and to the semantic reference itself. Even difficult-to-decode portions of speech, such as speech with dialectal variations, heavily slurred speech, or noise-overlaid signal portions can often be decoded in this fashion (see e.g., Greenberg, 1999).

This network of relationships is not only perceptual in nature. In speech production, we appear to access part of the same network to produce speech that transmits information faultlessly to listeners despite massive reductions in acoustic clarity, phonetic structure, and redundancy. Very informal forms of speech, for example, can remain perfectly understandable for initiated listeners, all the while

showing considerably obscured segmental and prosodic structure. For some strongly informal styles, we do not even know yet how to segment the speech material in systematic fashion, or how to model it prosodically.[4] The enormous network of relations rendering comprehension possible under such trying circumstances takes a human being twenty or more years to build, using the massive parallel processing capacity of the human brain.

Current automatic analysis systems are still far from that sort of processing capacity, or from such a sophisticated level of linguistic knowledge. Only relatively simple relationships can be learned automatically, and automatic recognition systems still derail much too easily, particularly on rapidly pronounced and informal segments of speech. This in turn retards the creation of databases for the full range of stylistic and vocal variations that we humans are familiar with.

Challenges and Promises

We are thus led to argue (a) that the dominant TD technology is too cumbersome for the task of providing a full range of styles and voices; and (b) that current automatic processing technology is not up to generating automatic databases for many of the styles and voices that would be desirable in a wider synthesis application context. Understandably, these positions may not be very popular in some quarters. They suggest that after a little spurt during which a few more mature adult voices and relatively formal styles will become available with the current technology, speech synthesis research will have to face up to some of the tough speech science problems that were temporarily left behind. The problem of excessive complexity, for example, will have to be solved with the combined tools of a deeper understanding of speech variability and more sophisticated modelling of various levels of speech generation. Advanced spectral synthesis techniques are also likely to be part of this effort, and this is what we turn to next.

Major Challenge One: Advanced Spectral Synthesis Techniques

'Reports of my death are greatly exaggerated,' said Mark Twain, and similarly, spectral synthesis methods were probably buried well before they were dead. To mention just a few teams who have remained active in this domain throughout the 1990s: Ken Stevens and his colleagues at MIT and John Local at the University of

[4] Sound example Walker and Local (Webpage) illustrates this problem. It is a stretch of informal conversational English between two UK university students, recorded under studio conditions. The transcription of the passage, agreed upon by two native-dialect listeners, is as follows: '*I'm* gonna save that and water my plant with it (1.2 s pause with in-breath), give some to Pip (0.8 s pause), 'cos we were trying, 'cos it says that it shouldn't have treated water.' The spectral structure of this passage is very poor, and we submit that current automatic recognition systems would have a very difficult time decoding this material. Yet the person supervising the recording reports that the two students never once showed any sign of not understanding each other. (Thanks to Gareth Walker and John Local, University of York, UK, for making the recording available.)

York (UK) have continued their remarkable investigations on formant synthesis (Local, 1994, 1997; Stevens, 1998). Some researchers, such as Professor Hoffmann's team in Dresden, have put formant synthesisers on ICs. Professor Vich's team in Prague has developed advanced LPC-based methods, LPC is also the basis of the SRELP algorithm for prosody manipulation, as an alternative to PSOLA technique, described by Erhard Rank in Chapter 7 of this volume. Professor Burileanu's team in Rumania, as well as others, have pursued solutions based on the CELP algorithm. Professor Kubin's team in Vienna (now Graz), Steve McLaughlin at Edinburgh and Donald Childers/José Principe at the University of Florida have developed synthesis structures based on the Non-linear Oscillator Model. And perhaps most prominent has been the work on harmonics-and-noise modelling (HNM) (Stylianou, 1996; and articles by Bailly, Banga, O'Brien and colleagues in this volume). HNM provides acoustic results that are particularly pleasing, and the key speech transform function, the harmonics+noise representation, is relatively easy to understand and to manipulate.[5]

For a simple analysis–re-synthesis cycle, the algorithm proceeds basically as follows (precise implementations vary): narrow-band spectra are obtained at regular intervals in the speech signal, amplitudes and frequencies of the harmonic frequencies are identified, irregular and unaccounted-for frequency (noise) components are identified, time, frequency and amplitude modifications of the stored values are performed as desired, and the modified spectral representations of the harmonic and noise components are inverted into temporal representations and added linearly. When all steps are performed correctly (no mean task), the resulting output is essentially 'transparent', i.e., indistinguishable from normal speech. In the framework of the COST 258 signal generation test array (Bailly, Chapter 4, this volume), several such systems have been compared on a simple F0-modification task (www.icp.inpg.fr/cost258/evaluation/server/cost258_coders.html). The results for the HNM system developed by Eduardo Banga of Vigo in Spain are given in sound examples Vigo (a–f). Using this technology, it is possible to perform the same functions as those performed by TD-synthesis, at the same or better levels of sound quality.

Crucially, voice and timbre modifications are also under programmer control, which opens the door to the substantial new territory of voice/timbre modifications, and promises to drastically reduce the need for separate DBs for different voices.[6] In addition, the HNM (or similar) spectral transforms can be rendered storage-efficient. Finally, speed penalties that have long disadvantaged spectral techniques with respect to TD techniques have recently been overcome through the combination of efficient algorithms and the use of faster processor speeds. Advanced HNM algorithms can, for example, output speech synthesis in real time on computers equipped with 300+ MHz processors.

[5] A new European project has recently been launched to undertake further research in the area of non-linear speech processing (COST 277).

[6] It is not clear yet if just *any* voice could be generated from a single DB at the requisite quality level. At current levels of research, it appears that at least initially, it may be preferable to create DBs for 'families' of voices.

Major Challenge Two: The Modelling of Style and Voice

But building satisfactory spectral algorithms is only the beginning, and the work required to implement a full range of style or voice modulations with such algorithms is likely to be daunting. Sophisticated voice and timbre models will have to be constructed to enforce 'voice credibility' over voice/timbre modifications. These models will store voice and timbre information abstractly, rather than explicitly as in TD-synthesis, in the form of underlying parameters and inter-parameter constraints.

To handle informal styles of speech in addition to more formal styles, and to handle the full range of dialectal variation in addition to a chosen norm, a set of complex language use, dialectal and sociolinguistic models must be developed. Like the voice/timbre models, the style models will represent their information in abstract, underlying and inter-parameter constraint form. Only when the structural components of such models are known, will it become possible to employ automatic recognition paradigms to look in detail for the features that the model expects.[7] Voice/timbre models as well as language use, dialectal and sociolinguistic models will have to be created with the aid of a great deal of experimentation, and on the basis of much traditional empirical scientific research.

In the long run, complete synthesis systems will have to be driven by empirically-based models that encode the admirable complexity of our human communication apparatus. This will involve clarifying the theoretical status of a great number of parameters that remain unclear or questionable in current models. Concretely, we must learn to predict style-, voice- and dialect-induced variations both at the detailed phonetic and prosodic levels before we can expect our synthesis systems to provide natural-sounding speech in a much larger variety of settings.

But the long-awaited pay-off will surely come. The considerable effort delineated here will gradually begin to let us create virtual speech on a par with the impressive visual virtual worlds that exist already. While these results are unlikely to be 'just around the corner', they are the logical outcomes of the considerable further research effort described here.

A New Research Tool: Speech Synthesis as a Test of Linguistic Modelling

A final development to be touched upon here is the use of speech synthesis as a scientific tool with considerable impact. In fact, speech synthesis is likely to help advance the described research effort more rapidly than traditional tools would.

[7] The careful reader will have noticed that we are *not* suggesting that the positive developments of the last decade be simply discarded. Statistical and neural network approaches will remain our main tools for discovering structure and parameter loading coefficients. Diphone, polyphone, etc. databases will remain key storage tools for much of our linguistic knowledge. And automatic segmentation systems will certainly continue to prove their usefulness in large-scale empirical investigations. We *are* saying, however, that TD-synthesis is not up to the challenge of future needs of speech synthesis, and that automatic segmentation techniques need sophisticated theoretical guidance and programming to remain useful for building the next generation of speech synthesis systems.

This is because modelling results are much more compelling when they are presented in the form of audible speech than in the form of tabular comparisons or statistical evaluations. In fact, it is possible to envision speech synthesis becoming elevated to the status of an obligatory test for future models of language structure, language use, dialectal variation, sociolinguistic parametrisation, as well as timbre and voice quality. The logic is simple: if our linguistic, sociolinguistic and psycholinguistic theories are solid, it should be possible to demonstrate their contribution to the greater quality of synthesised speech. If the models are 'not so hot', we should be able to hear that as well.

The general availability of such a test should be welcome news. We have long waited for a better means of challenging a language-science model than saying that 'my *p*-values are better than yours' or 'my informant can say what your model doesn't allow'. Starting immediately, a language model can be run through its paces with many different styles, stimulus materials, speech rates, and voices. It can be caused to fail, and it can be tested under rigorous controls. This will permit even external scientific observers to validate the output of our linguistic models. After a century of sometimes wild theoretical speculation and experimentation, linguistic modelling may well take another step towards becoming an externally accountable science, and that despite its enormous complexity. Synthesis can serve to verify analysis.

Conclusion

Current speech synthesis is at the threshold of some vibrant new developments. Over the past ten years, improved prosodic models and concatenative techniques have shown that high-quality speech synthesis *is* possible. As the coming decade pushes current technology to its limits, systematic research on novel signal generation techniques and more sophisticated phonetic and prosodic models will open the doors towards even greater naturalness of synthetic speech appropriate to a much greater variety of uses. Much work on style, voice, language and dialect modelling waits in the wings, but in contrast to the somewhat cerebral rewards of traditional forms of speech science, much of the hard work in speech synthesis is sure to be rewarded by pleasing and quite audible improvements in speech quality.

Acknowledgements

Grateful acknowledgement is made to the Office Fédéral de l'Education (Berne, Switzerland) for supporting this research through its funding in association with Swiss participation in COST 258, and to the University of Lausanne for funding a research leave for the author, hosted in Spring 2000 at the University of York. Thanks are extended to Brigitte Zellner Keller, Erhard Rank, Mark Huckvale and Alex Monaghan for their helpful comments.

References

Bhaskararao, P. (1994). Subphonemic segment inventories for concatenative speech synthesis. In E. Keller (ed.). *Fundamentals in Speech Synthesis and Speech Recognition* (pp. 69–85). Wiley.

Campbell, W.N. (1992a). *Multi-level Timing in Speech*. PhD thesis, University of Sussex.

Campbell, W.N. (1992b). Syllable-based segmental duration. In G. Bailly *et al*. (eds), *Talking Machines: Theories, Models, and Designs* (pp. 211–224). Elsevier Science Publishers.

Campbell, W.N. (1996). CHATR: A high-definition speech resequencing system. *Proceedings 3rd ASA/ASJ Joint Meeting* (pp. 1223–1228). Honolulu, Hawaii.

Greenberg, S. (1999). Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, *29*, 159–176.

Keller, E. (1997). Simplification of TTS architecture vs. operational quality. *Proceedings of EUROSPEECH '97*. Paper 735. Rhodes, Greece.

Keller, E. and Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics, 17*, 53–75. University of York. (available at www.unil.ch/imm/docs/LAIP/pdf.files/ Keller-Zellner-96-YorkPprs.pdf).

Keller, E., Zellner, B., and Werner, S. (1997). Improvements in prosodic processing for speech synthesis. *Proceedings of Speech Technology in the Public Telephone Network: Where are we Today?* (pp. 73–76) Rhodes, Greece.

Keller, E., Zellner, B., Werner, S., and Blanchoud, N. (1993). The prediction of prosodic timing: Rules for final syllable lengthening in French. *Proceedings ESCA Workshop on Prosody* (pp. 212–215). Lund, Sweden.

Klatt, D.W. (1989). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, *82*, 737–793.

Klatt, D.H. and Klatt, L.C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, *87*, 820–857.

LAIPTTS (a–l). LAIPTTS_a_VersaillesSlow.wav., LAIPTTS_b_VersaillesFast.wav, LAIPTTS_c_VersaillesAcc.wav, LAIPTTS_d_VersaillesHghAcc.wav, LAIPTTS_e_Rhythm_fluent.wav, LAIPTTS_f_Rhythm_disfluent.wav, LAIPTTS_g_BerlinDefault.wav, LAIPTTS_h_BerlinAdjusted.wav, LAIPTTS_i_bonjour.wav . . . _l_bonjour.wav. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm

Local, J. (1994). Phonological structure, parametric phonetic interpretation and natural-sounding synthesis. In E. Keller (ed.), *Fundamentals in Speech Synthesis and Speech Recognition* (pp. 253–270). Wiley.

Local, J. (1997). What some more prosody and better signal quality can do for speech synthesis. *Proceedings of Speech Technology in the Public Telephone Network: Where are we Today?* (pp. 77–84). Rhodes, Greece.

Ogden, R., Local, J., and Carter, P. (1999). Temporal interpretation in ProSynth, a prosodic speech synthesis system. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A.C. Bailey (eds), *Proceedings of the XIVth International Congress of Phonetic Sciences*, vol. 2 (pp. 1059–1062). University of California, Berkeley, CA.

Riley, M. (1992). Tree-based modelling of segmental durations. In G. Bailly *et al*., (eds), *Talking Machines: Theories, Models, and Designs* (pp. 265–273). Elsevier Science Publishers.

Stevens, K.N. (1998). *Acoustic Phonetics*. The MIT Press.

Styger, T. and Keller, E. (1994). Formant synthesis. In E. Keller (ed.), *Fundamentals in Speech Synthesis and Speech Recognition* (pp. 109–128). Wiley.

Stylianou, Y. (1996). *Harmonic Plus Noise Models for Speech, Combined with Statistical Methods for Speech and Speaker Modification*. PhD Thesis, École Nationale des Télécommunications, Paris.

van Santen, J.P.H. and Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *JASA*, *107*, 1012–1026.

Vigo (a–f). Vigo_a_LesGarsScientDesRondins_neutral.wav, Vigo_b_LesGarsScientDesRondins_question.wav, Vigo_c_LesGarsScientDesRondins_slow.wav, Vigo_d_LesGarsScientDesRondins_surprise.wav, Vigo_e_LesGarsScientDesRondins_incredul.wav, Vigo_f_LesGars ScientDesRondins_itsEvident.wav. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm.

Walker, G. and Local, J. Walker_Local_InformalEnglish.wav. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258 volume.htm.

YorkTalk (a–c). YorkTalk_sudden.wav, YorkTalk_yellow.wav, YorkTalk_c_NonSegm.wav. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée, 1*, 7–23.

Zellner, B. (1997). Fluidité en synthèse de la parole. In E. Keller and B. Zellner (eds), *Les Défis actuels en synthèse de la parole. Études des Lettres*, *3* (pp. 47–78). Université de Lausanne.

Zellner, B. (1998a). *Caractérisation et prédiction du débit de parole en français. Une étude de cas*. Unpublished PhD thesis. Faculté des Lettres, Université de Lausanne. (Available at www.unil.ch/imm/docs/LAIP/ps.files/ DissertationBZ.ps).

Zellner, B. (1998b). Temporal structures for fast and slow speech rate. *ESCA/COCOSDA Third International Workshop on Speech Synthesis* (pp. 143–146). Jenolan Caves, Australia.

Zellner Keller, B. and Keller, E. (in press). The chaotic nature of speech rhythm: Hints for fluency in the language acquisition process. In Ph. Delcloque and V.M. Holland (eds) *Speech Technology in Language Learning*: *Recognition, Synthesis, Visualisation, Talking Heads and Integration*, Swets and Zeitlinger.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

2

Towards More Versatile Signal Generation Systems

Gérard Bailly
Institut de la Communication Parlée – UMR-CNRS 5009
INPG and Université Stendhal, 46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
bailly@icp.grenet.fr

Introduction

Reproducing most of the variability observed in natural speech signals is the main challenge for speech synthesis. This variability is highly contextual and is continuously monitored in speaker/listener interaction (Lindblom, 1987) in order to guarantee optimal communication with minimal articulatory effort for the speaker and cognitive load for the listener. The variability is thus governed by the structure of the language (morphophonology, syntax, etc.), the codes of social interaction (prosodic modalities, attitudes, etc.) as well as individual anatomical, physiological and psychological characteristics. Models of signal variability –and this includes prosodic signals – should thus generate an optimal signal given a set of desired features. Whereas concatenation-based synthesisers use these features directly for selecting appropriate segments, rule-based synthesisers require fuzzier[1] coarticulation models that relate these features to spectro-temporal cues using various data-driven least-square approximations. In either case, these systems have to use signal processing or more explicit signal representation in order to extract the relevant spectro-temporal cues. We thus need accurate signal analysis tools not only to be able to modify the prosody of natural speech signals but also to be able to characterise and label these signals appropriately.

Physical interpretability vs. estimation accuracy

For historical and practical reasons, complex models of the spectro-temporal organisation of speech signals have been developed and used mostly by rule-based

[1] More and more fuzzy as we consider interaction of multiple sources of variability. It is clear, for example, that spectral tilt results from a complex interaction between intonation, voice quality and vocal effort (d'Alessandro and Doval, 1998) and that syllabic structure has an effect on patterns of excitation (Ogden *et al.*, 2000).

synthesisers. The speech quality reached by a pure concatenation of natural speech segments (Black and Taylor, 1994; Campbell, 1997) is so high that complex coding techniques have been mostly used for the compression of segment dictionaries.

Physical interpretability

Complex speech production models such as formant or articulatory synthesis provide all spectro-temporal dimensions necessary and sufficient to characterise and manipulate speech signals. However, most parameters are difficult to estimate from the speech signal (articulatory parameters, formant frequencies and bandwidths, source parameters, etc.). Part of this problem is due to the large number of parameters (typically a few dozen) that have an influence on the entire spectrum: parameters are often estimated independently and consequently the analysis solution is not unique[2] and depends mainly on different estimation methods used.

If physical interpretability was a key issue for the development of early rule-based synthesisers where knowledge was mainly declarative, sub-symbolic processing systems (hidden Markov models, neural networks, regression trees, multilinear regression models, etc.) now succeed in producing a dynamically-varying parametric representation from symbolic input given input/output exemplars. Moreover, early rule-based synthesisers used simplified models to describe the dynamics of the parameters such as targets connected by interpolation functions or fed into passive filters, whereas more complex dynamics and phase relations have to be generated for speech to sound natural.

Characterising speech signals

One of the main strengths of formant or articulatory synthesis lies in providing a *constant* number of *coherent*[3] spectro-temporal parameters suitable for any sub-symbolic processing system that maps parameters to features (for feature extraction or parameter generation) or for spectro-temporal smoothing as required for segment inventory normalisation (Dutoit and Leich, 1993). Obviously traditional coders used in speech synthesis such as TD-PSOLA or RELP are not well suited to these requirements.

An important class of coders – *spectral models*, such as the ones described and evaluated in this section – avoid the oversimplified characterisation of speech signals in the time domain. One advantage of spectral processing is that it tolerates phase distortion, while glottal flow models often used to characterise the voice source (see, for example, Fant *et al.*, 1985) are very sensitive to the temporal shape of the signal waveform. Moreover spectral parameters are more closely related to perceived speech quality than time-domain parameters. The vast majority of these coders have been developed for speech coding as a means to bridge the gap (in

[2] For example, spectral slope can be modelled by source parameters as well as by formant bandwidths.

[3] Coherence here concerns mainly sensitivity to perturbations: small changes in the input parameters should produce small changes in spectro-temporal characteristics and vice versa.

terms of bandwidth) between waveform coders and LPC vocoders. For these coders, the emphasis has been on the perceptual transparency of the analysis-synthesis process, with no particular attention to the interpretability or transparency of the intermediate parametric representation.

Towards more 'ecological' signal generation systems

Contrary to articulatory or terminal-analogue synthesis that guarantees that almost all the synthetic signals could have been produced by a human being (or at least by a vocal tract), the coherence of the input parameters guarantees the naturalness of synthetic speech produced by *phenomenological models* (Dutoit, 1997, p. 193) such as the spectral models mentioned above. The resulting speech quality depends strongly on the *intrinsic* limitations imposed by the model of the speech signal and on the *extrinsic* control model. Evaluation of signal generation systems can thus divided into two main issues: (a) the intrinsic ability of the analysis-synthesis process to preserve subtle (but perceptually relevant) spectro-temporal characteristics of a large range of natural speech signals; and (b) the ability of the analysis scheme to deliver a parametric representation of speech that lends itself to an extrinsic control model. Assuming that most spectral vocoders provide toll-quality output for any speech signal, the evaluation proposed in this part concerns the second point and compares the performance of various signal generation systems on *independent* variation of prosodic parameters without any system-specific model of the interactions between parameters.

Part of this interaction should of course be modelled by an extrinsic control about which we are still largely ignorant. Emerging research fields tackled in Part III will oblige researchers to model the complex interactions at the acoustic level between intonation, voice quality and segmental aspects: these interactions are far beyond the simple superposition of independent contributions.

References

d'Alessandro, C. and Doval, B. (1998). Experiments in voice quality modification of natural speech signals: The spectral approach. *Proceedings of the International Workshop on Speech Synthesis* (pp. 277–282). Jenolan Caves, Australia.

Black, A.W. and Taylor, P. (1994). CHATR: A generic speech synthesis system. *COLING-94*, Vol. II, 983–986.

Campbell, W.N. (1997). Synthesizing spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi (eds), *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 165–186). Springer Verlag.

Dutoit, T. (1997). *An Introduction to Text-to-speech Synthesis*. Kluwer Academics.

Dutoit, T. and Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, *13*, 435–440.

Fant, G., Liljencrants, J., and Lin, Q. (1985). A Four Parameter Model of the Glottal Flow. Technical Report 4. Speech Transmission Laboratory, Department of Speech Communication and Music Acoustics, KTH.

Lindblom, B. (1987). Adaptive variability and absolute constancy in speech signals: Two themes in the quest for phonetic invariance. *Proceedings of the XIth International Congress of Phonetic Sciences*, Vol. 3 (pp. 9–18). Tallin, Estonia.

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovičová, J., and Heid, S. (2000). ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Computer Speech and Language*, *14*, 177–210.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

3

A Parametric Harmonic + Noise Model

Gérard Bailly
Institut de la Communication Parlée –UMR-CNRS 5009
INPG and Université Stendhal, 46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
bailly@icp.grenet.fr

Introduction

Most current text-to-speech systems (TTS) use concatenative synthesis where segments of natural speech are manipulated by analysis-synthesis techniques in such a way that the resulting synthetic signal conforms to a given computed prosodic description. Since most prosodic descriptions include melody, segment duration and energy, such coders should allow at least these modifications. However, the modifications are often accompanied by distortions in other spatio-temporal dimensions that do not necessarily reflect covariations observed in natural speech. Contrary to synthesis-by-rule systems where such observed covariations may be described and implemented (Gobl and Ní Chasaide, 1992), coders should intrinsically exhibit properties that guarantee an optimal extrapolation of temporal/spectral behaviour given only a reference sample. One of these desired properties is shape invariance in the time domain (McAulay and Quatieri, 1986; Quatieri and McAulay, 1992). Shape invariance means maintaining the signal shape in the vicinity of vocal tract excitation (pitch marks). PSOLA techniques achieve this by centring short-term signals on pitch marks.

Although TD-PSOLA-based coders (Hamon *et al*., 1989; Charpentier and Moulines, 1990; Dutoit and Leich, 1993) and cepstral vocoders are preferred in most TTS systems and outperform vocal tract synthesisers driven by synthesis-by-rule systems, they still do not produce adequate covariation, particularly for large prosodic modifications. They also do not allow accurate and flexible control of covariation: the covariation depends on speech styles, and shape invariance is only a first approximation – a minimum common denominator – of what occurs in natural speech.

Sinusoidal models can maintain shape invariance by preserving the phase and amplitude spectra at excitation instants. Valid covariation of these spectra according to prosodic variations may be added to better approximate natural

speech. Modelling this covariation is one of the possible improvements in the naturalness of synthetic speech envisaged by COST 258. This chapter describes a parametric HNM suitable for building such comprehensive models.

Sinusoidal models

McAulay and Quatieri

In 1986 McAulay and Quatieri (McAulay and Quatieri, 1986; Quatieri and McAulay, 1986) proposed a sinusoidal analysis-synthesis model that is based on amplitudes, frequencies, and phases of component sine waves. The speech signal s(t) is decomposed into L(t) sinusoids at time t:

$$\mathrm{s(t)} = \sum_{l=1}^{L(\mathrm{t})} A_l(t)\mathcal{R}(e^{j\psi_l(t)}),$$

where $A_l(t)$ and $\psi_l(t)$ are the amplitude and phase of the lth sinewave along the frequency track $\omega_l(t)$. These tracks are determined using a birth–death frequency tracker that associates the set of $\omega_l(t)$ with FFT peaks. The problem is that the FFT spectrum is often spoiled by spurious peaks that 'come and go due to the effects of side-lobe interaction' (McAulay and Quatieri, 1986, p. 748). We will come back to this problem later.

Serra

The residual of the above analysis/synthesis sinusoidal model has a large energy, especially in unvoiced sounds. Furthermore, the sinusoidal model is not well suited to the lengthening of these sounds, which results – as in TD-PSOLA techniques – in a periodic modulation of the original noise structure. A phase randomisation technique may be applied (Macon, 1996) to overcome this problem. Contrary to Almeida and Silva (1984), Serra (1989; Serra and Smith, 1990) considers the residual as a stochastic signal whose spectrum should be modelled globally.

This stochastic signal includes aspiration, plosion and friction noise, but also modelling errors partly due to the procedure for extracting sinusoidal parameters.

Stylianou et al.

Stylianou *et al.* (Laroche *et al.*, 1993; Stylianou, 1996) do not use Serra's birth–death frequency tracker. Given the fundamental frequency of the speech signal, they select harmonic peaks and use the notion of maximal voicing frequency (MVF). Above the MVF, the residual is considered as being stochastic, and below the MVF as a modelling error.

This assumption is, however, unrealistic. The aspiration and friction noise may cover the entire speech spectrum even in the case of voiced sounds. Before examining a more realistic decomposition on p. 000, we will first discuss the sinusoidal analysis scheme.

The sinusoidal analysis

Most sinusoidal analysis procedures rely on an initial FFT. Sinusoidal parameters are often estimated using frequencies, amplitudes and phases of the FFT peaks. The values of the parameters obtained by this method are not directly related to $A_l(t)$ and $\varphi_l(t)$ This is mainly because of the windowing and energy leaks due to the discrete nature of the computed spectrum. Chapter 2 of Serra's thesis is dedicated to the optimal choice of FFT length, hop size and window (see also Harris, 1978) or more recently (Puckette and Brown, 1998). This method produces large modelling errors – especially for sounds with few harmonics[1] – that most sinusoidal models filter out (Stylianou, 1996) in order to interpret the residual as a stochastic component.

George and Smith (1997) propose an analysis by synthesis method (ABS) for the sinusoidal model based on an iterative estimation and subtraction of elementary sinusoids. The parameters of each sinusoid are estimated by minimisation of a linear least-squares approximation over candidate frequencies. The original ABS algorithm iteratively selects each candidate frequency in the vicinity of the most prominent peak of the FFT of the residual signal.

We improved the algorithm (PS-ABS for Pitch-Synchronous ABS) by (a) forcing $\omega_l(t)$ to be a multiple of the local pitch period ω_0; (b) iteratively estimating the parameters using a time window centred on a pitch mark and exactly equal to the two adjacent pitch periods; and (c) compensating for the mean amplitude change in the analysis window.

The average modelling error on the fully harmonic synthetic signals provided by d'Alessandro *et al.* (1998; Yegnanarayana *et al.*, 1998) is -33 dB for PS-ABS.

We will evaluate below the ability of the proposed PS-ABS method to produce a residual signal that can be interpreted as the real stochastic contribution of noise sources to the observed signal.

Deterministic/stochastic decomposition

Using an extension of continuous spectral interpolation (Papoulis, 1986) to the discrete domain, d'Alessandro and colleagues have proposed an iterative procedure for the initial separation of the deterministic and stochastic components (d'Alessandro *et al.*, 1995 and 1998). The principle is quite simple: each frequency is initially attributed to either component. Then one component is iteratively interpolated by alternating between time and frequency domains where domain-specific constraints are applied: in the time domain, the signal is truncated and in the frequency domain, the spectrum is imposed on the frequency bands originally attributed to the interpolated component. These time/frequency constraints are applied at each iteration and convergence is obtained after a few iterations (see Figure 3.1). Our implementation of this original algorithm is called YAD in the following.

[1] Of course FFT-based methods may give low modelling errors for complex sounds, but the estimated sinusoidal parameters do not reflect the true sinusoidal content.

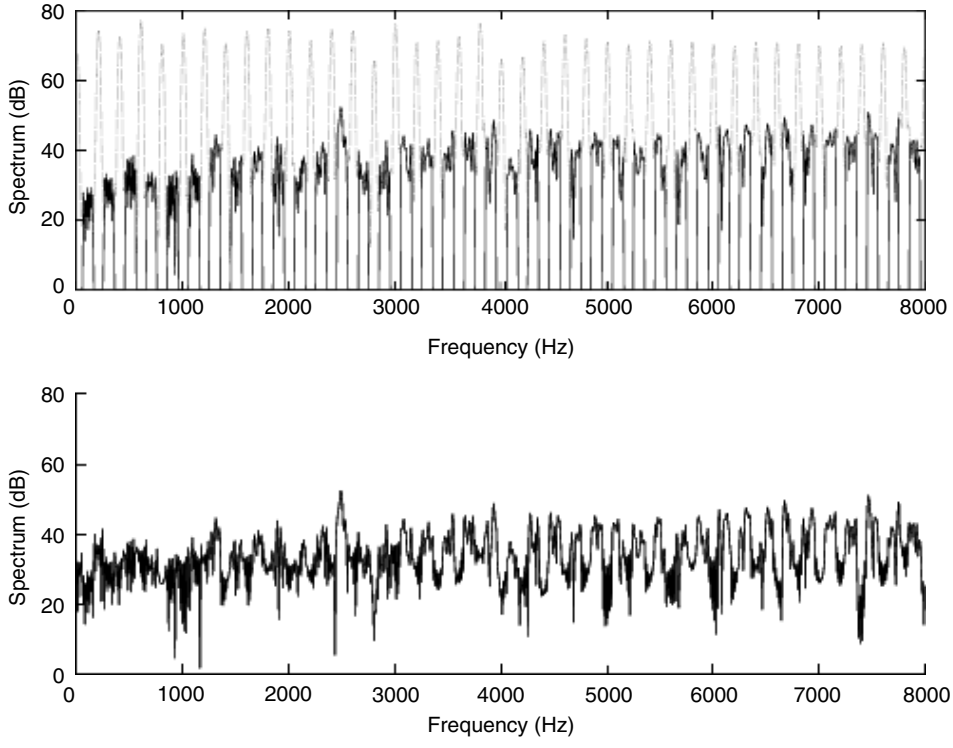


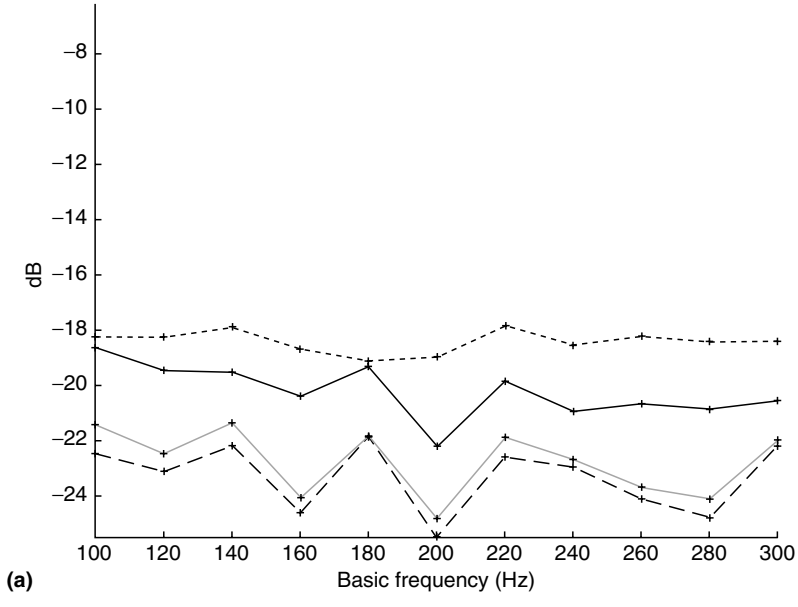
Figure 3.1 Interpolation of the aperiodic component of the LP residual of a frame of a synthetic &WID[;&WID]; – F0: 200 Hz – Sampling frequency: 16 kHz. Top: the FFT spectrum before extrapolation with the original spectrum with dotted lines. Bottom: after extrapolation

This initial procedure has been extended by Ahn and Holmes (1997) by a joint estimation that alternates between deterministic/stochastic interpolation. Our implementation is called AH in the following.

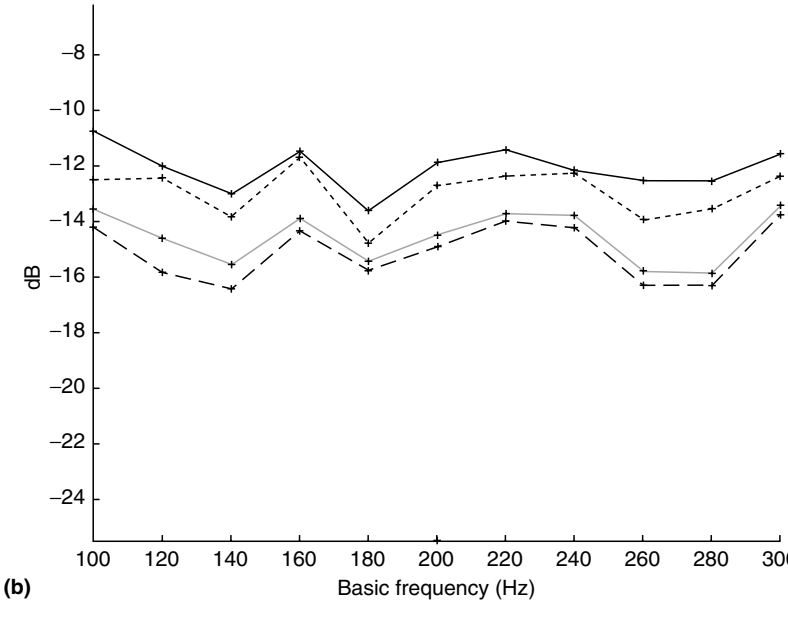
These two decomposition procedures were compared to the PS-ABS proposed above using synthetic stimuli used by d'Alessandro *et al.* (d'Alessandro *et al*., 1998; Yegnanarayana *et al*., 1998). We also assessed our current implementation of their algorithm. The results are summarised in Figure 3.2. They show that the YAD and AH perform equally well and slightly better than the original YAD implementation. This is probably due to the stop conditions: we stop the convergence when successive interpolated aperiodic components differ by less than 0.1 dB. The average number of iterations for YAD is, however, 18.1 compared to 2.96 for AH. The estimation errors for PS-ABS are always 4dB higher.

We further compared the decomposition procedures using natural VFV nonsense stimuli, where F is a voiced fricative (see Figure 3.3). When comparing YAD, AH and PS-ABS, the average differences between V's and F's HNR (cf. Table 3.1) were 18, 18.8 and 17.5 respectively.

For now the AH method seems to be the quickest and the most reliable method for the decomposition of harmonic/aperiodic components of speech (see Figure 3.4).


dB
−8
−10
−12
−14
−16
−18
−20
−22
−24
100
120
140
160
180
200
220
240
260
280
300
Basic frequency (Hz)

(a)


dB
−8
−10
−12
−14
−16
−18
−20
−22
−24
100
120
140
160
180
200
220
240
260
280
300
Basic frequency (Hz)

(b)

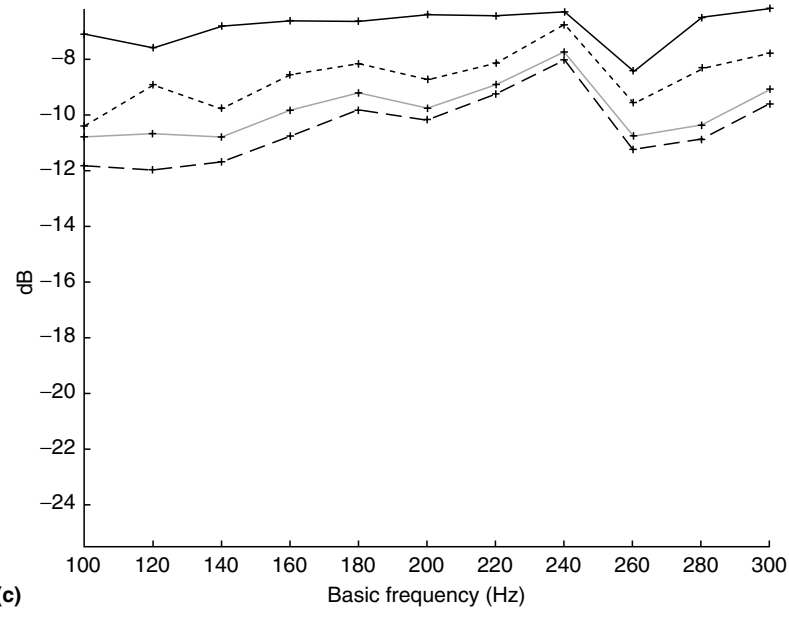


Figure 3.2 Recovering a known deterministic component using four different algorithms: PS-ABS (solid), YAD (dashed), AH (dotted). The original YAD results have been added (dash dot). The figures show the relative error of the deterministic component at different F0 values for three increasing aperiodic/deterministic ratio: (a) 20 dB, (b) 10 dB and (c) 5 dB

Table 3.1 Comparing harmonic to aperiodic ratio (HNR) at the target of different sounds

Phonemes	Number of targets	HNR (dB)		
		YAD	AH	PS-ABS
a	24	24.53	26.91	24.04
i	24	27.89	30.79	26.22
u	24	29.66	32.73	24.13
y	24	29.09	31.76	21.52
v	16	15.51	18.03	11.96
z	16	6.36	8.07	3.26
ʒ	16	7.49	9.12	4.22

Sinusoidal modification and synthesis

Synthesis

Most sinusoidal synthesis methods make use of the polynomial sinusoidal synthesis described by McAulay and Quatieri (1986, p. 750). The phase $\psi_l(t)$ is

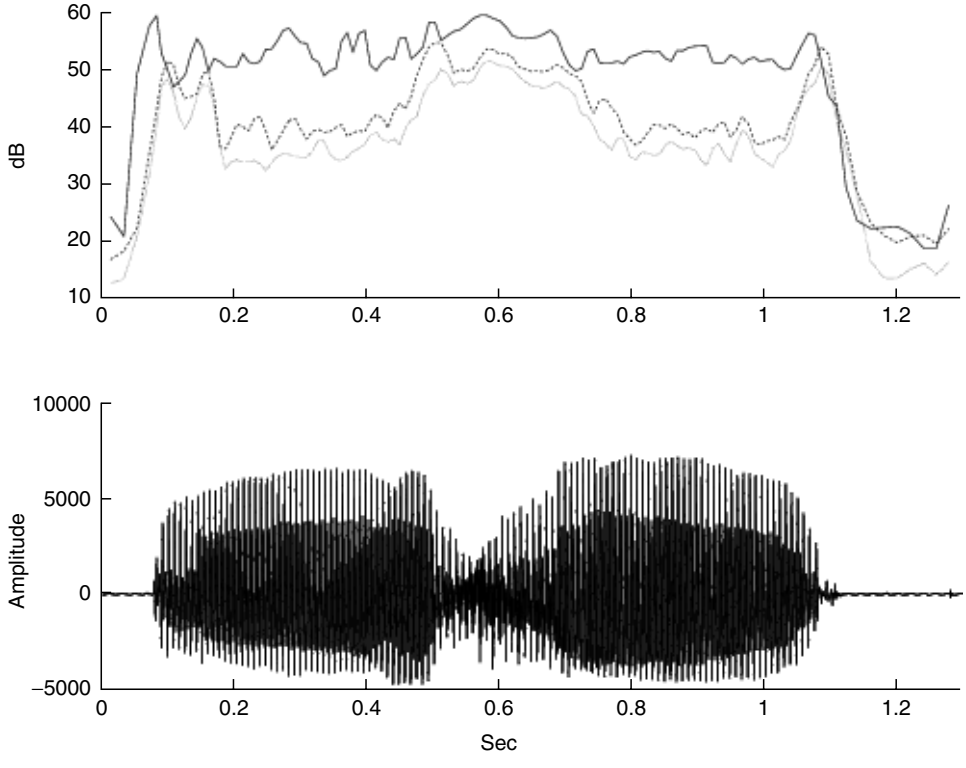


Figure 3.3 Energy of the aperiodic signal decomposed by different algorithms (same conventions as in Figure 3.2)

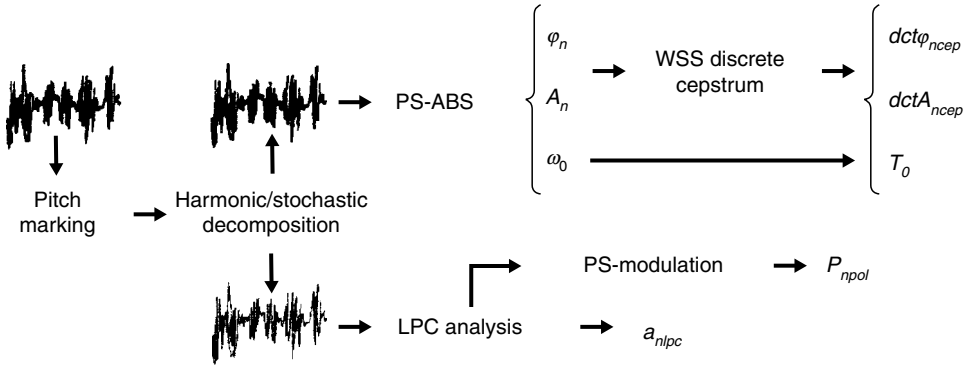


Figure 3.4 The proposed analysis scheme

interpolated between two successive frames n and $n+1$ characterised by $(\omega_l^n, \omega_l^{n+1}, \varphi_l^n, \varphi_l^{n+1})$ with a 3rd order polynomial $\psi_1(\mathrm{t}) = \mathrm{a} + \mathrm{bt} + \mathrm{ct}^2 + \mathrm{dt}^3$ where $0 < t < \Delta T$ with

$$\begin{cases} a = \varphi_1^n \\ b = \omega_1^n \\ \begin{bmatrix} \mathrm{c} \\ \mathrm{d} \end{bmatrix} = \begin{bmatrix} \frac{3}{\Delta T^2} & \frac{-1}{\Delta T} \\ \frac{-2}{\Delta T^3} & \frac{1}{\Delta T^2} \end{bmatrix} \cdot \begin{bmatrix} \varphi_1^{n+1} - \varphi_1^n - \omega_1^n \Delta T + 2\pi M \\ \omega_1^{n+1} - \omega_1^n \end{bmatrix} \\ M = E\left[\frac{1}{2\pi} - (\varphi_1^n + \omega_1^n \Delta T - \varphi_1^{n+1}) + \frac{\Delta T}{2}(\omega_1^{n+1} - \omega_1^n)\right] \end{cases}$$

Time-scale modification

For this purpose, systems avoid a pitch-synchronous analysis and synthesis scheme and introduce a higher-order polynomial interpolation (Pollard *et al.*, 1996; Macon, 1996). However, in the context of concatenative synthesis, it seems reasonable to assume access to individual pitch cycles. In this case, the polynomial sinusoidal synthesis described above has the intrinsic ability to interpolate between periods (see, for example, Figure 3.5).

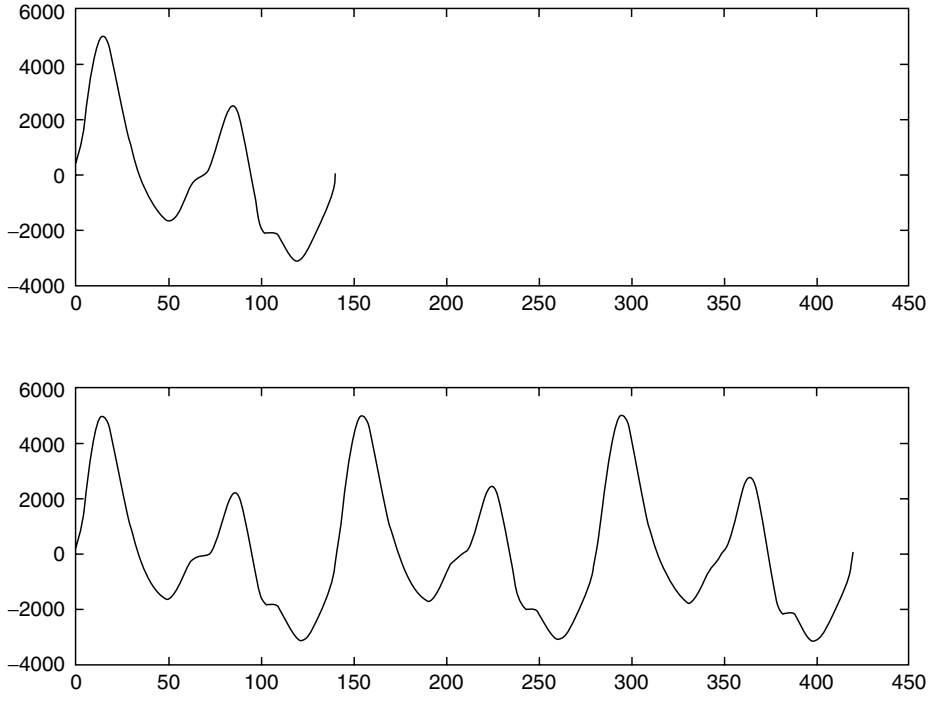


Figure 3.5 Intrinsic ability of the polynomial sinusoidal synthesis to interpolate periods. Top: synthesised period of length T = 140 samples. Bottom: same sinusoidal parameters but with T = 420 samples

Instead of a crude duplication of pitch-synchronous short-term signals, such an analysis/synthesis technique offers a precise estimation of the spectral characteristics of every pitch period and a clean and smooth time expansion.

Pitch-scale modification

Figure 3.6 shows PS-ABS estimations of amplitude and phase spectra for a synthetic vowel produced by exciting an LPC filter with a train of pulses at different F0 values. Changing the fundamental frequency of the speech signal while maintaining shape invariance and the spectral envelope consists thus of re-sampling the envelope at new harmonics.

Spectral interpolation

This can be achieved by interpolation (e.g. cubic splines have been used in Figure 3.6) or by estimating a model of the envelope. Stylianou (1996) uses, for example, a Discrete Cepstrum Transform (DCT) of the envelope, a procedure introduced by Galas and Rodet (1991), which has the advantage of characterising spectra with a constant number of parameters. Such a parametric representation simplifies later

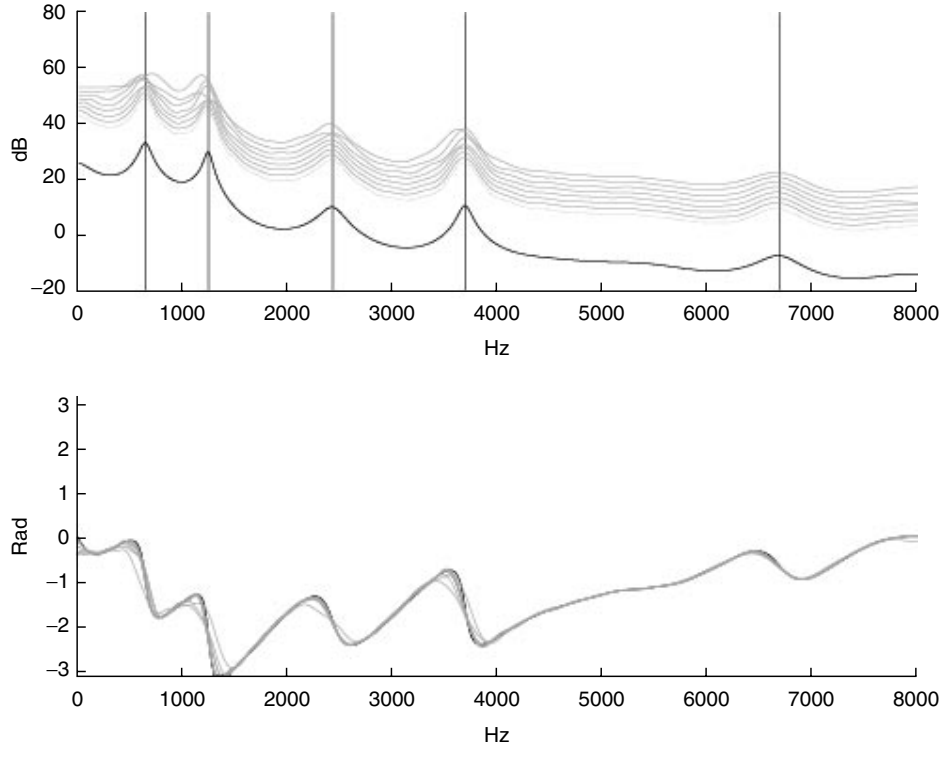


Figure 3.6 Amplitude and phase spectra for a synthetic [a]; produced by an LPC filter excited by a train of pulses at F0 ranging from 51 to 244 Hz. Amplitude spectrum lowers linearly with log(F0)

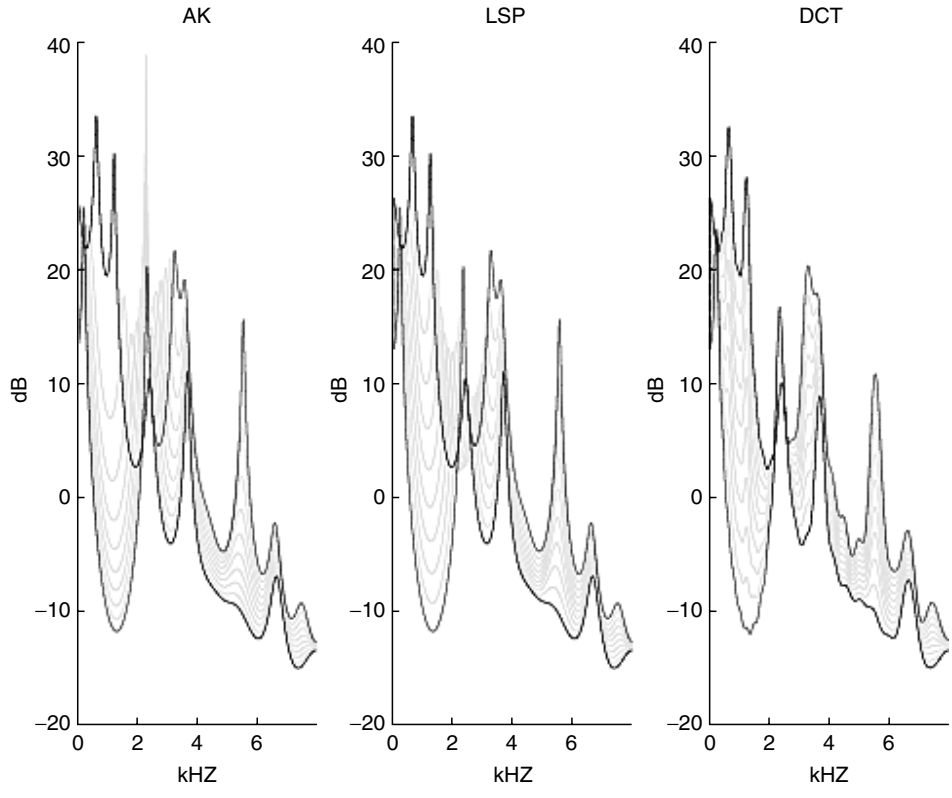


Figure 3.7 Interpolating between two spectra (here [a] and [i]) using three different models of the spectral envelope). From left to right: the linear prediction coefficients, the line spectrum pairs, the proposed DCT

spectral control and smoothing. Figure 3.7 shows the effect of different representations of the spectral envelope on interpolated spectra: the DCT produces a linear interpolation between spectra, whereas Line Spectrum Pairs (LSP) exhibit a more realistic interpolation between resonances (see Figure 3.8).

Discrete Cepstrum

Stylianou *et al*. use a constrained DCT operating on a logarithmic scale: cepstral amplitudes are weighted in order to favour a smooth interpolation. We added a weighted spectrum slope constraint (Klatt, 1982) that relaxes least-square approximation in the vicinity of valleys in the amplitude spectrum. Formants are better modelled and estimation of phases at harmonics with low amplitudes is relaxed. The DCT is applied to both the phase and amplitude spectra. The phase spectrum should of course be unwrapped before applying the DCT (see, for example, Stylianou, 1996; Macon, 1996).

Figure 3.9 shows an example of the estimation of the spectral envelope by a weighted DCT applied to the ABS spectrum.

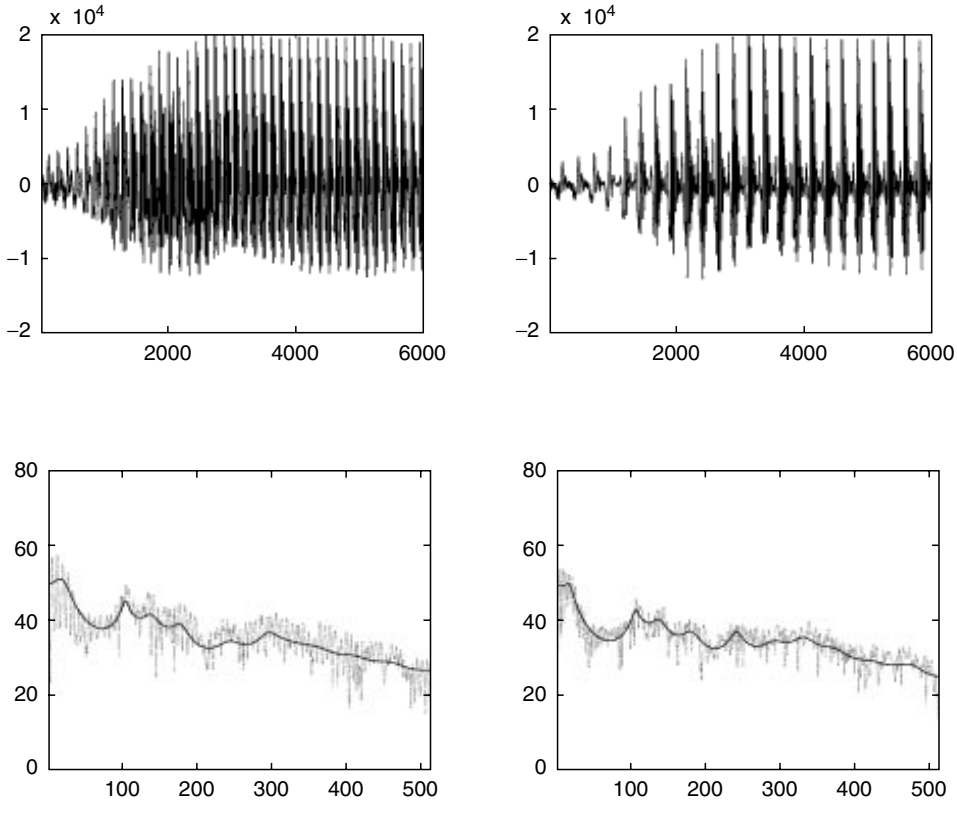
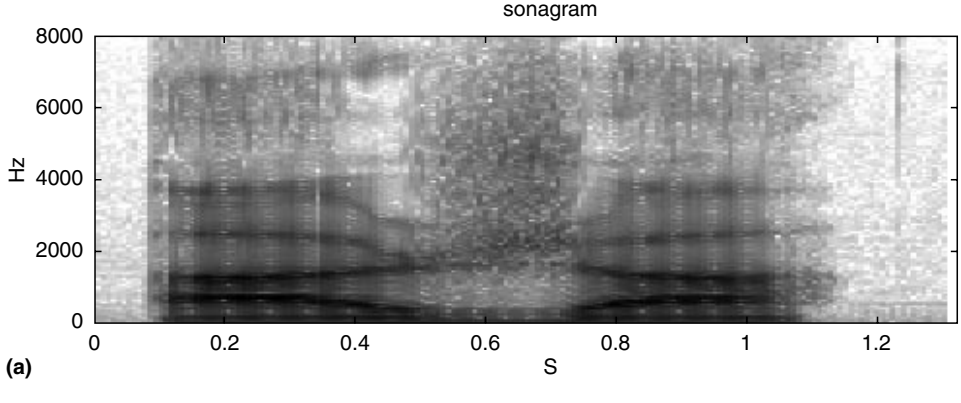


Figure 3.8 Modification of the deterministic component by sinusoidal analysis/synthesis. Left: part of the original signal and its FFT superposed with the LPC spectrum. Right: same for the resynthesis with a pitch scale of 0.6



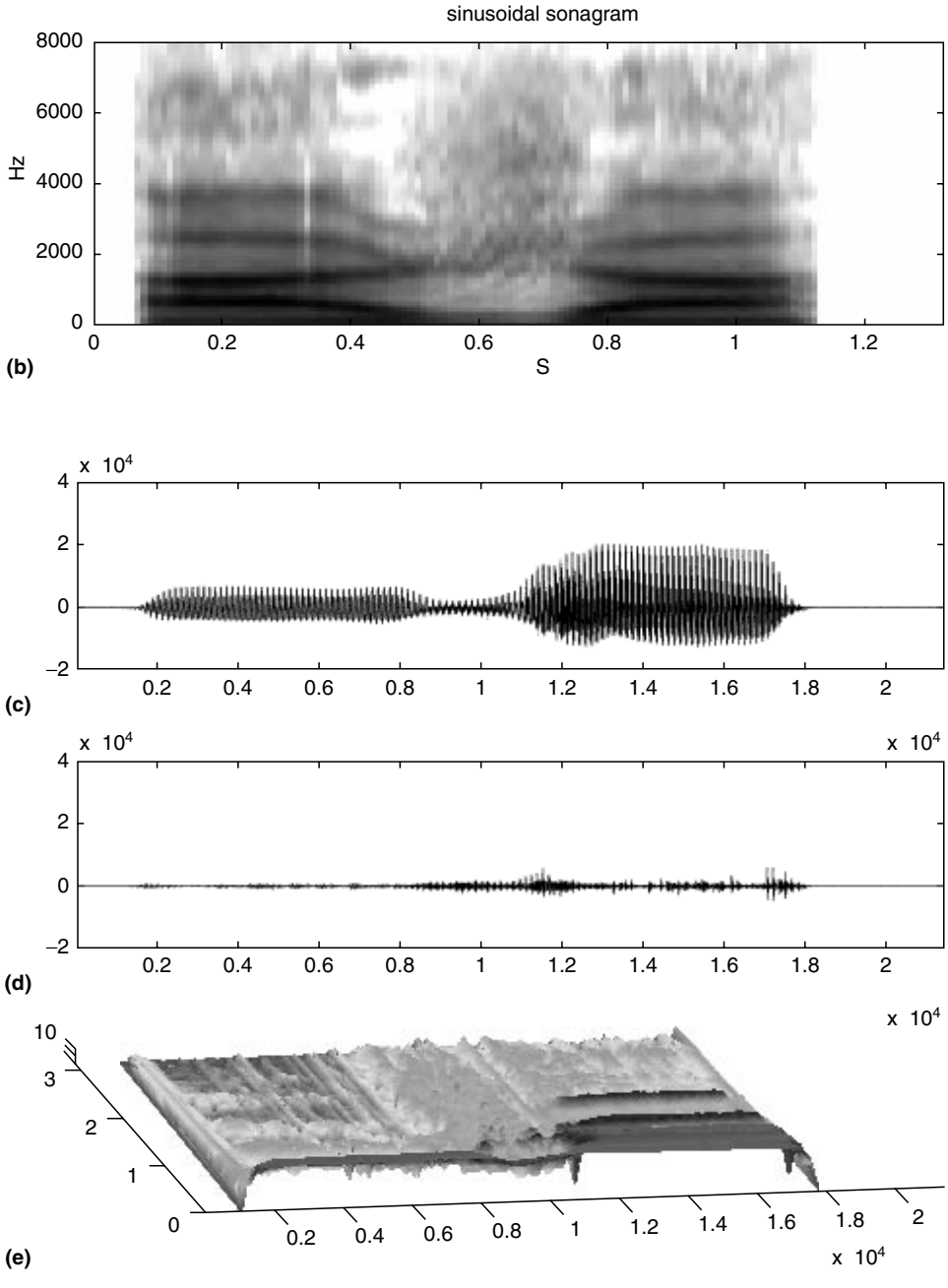


Figure 3.9 PS-ABS results. (a) sonagram of the original nonsense word /uʒa/. (b) amplitude spectrum estimated by and interpolated using weighted spectrum slope Discrete Cepstrum. (c) a nonsense word /uʒa/. (d) residual of the deterministic signal, (e) estimated amplitude spectrum

Stochastic analysis and synthesis

Formant waveforms

Richard and d'Alessandro (1997) proposed an analysis-modification-synthesis technique for stochastic signals. A multi-band analysis is performed where each bandpass signal is considered as a series of overlapping Formant Waveforms (FW) (Rodet, 1980). Figure 3.10 shows how the temporal modulation of each bandpass signal is preserved. We improved the analysis procedure by estimating all parameters in the time domain by least-square and optimisation procedures.

Modulated LPC

Here we compare results with the modulated output of a white noise excited LPC. The analysis is performed pitch-synchronously using random pitchmarks in the unvoiced portions of the signal. The energy pattern M(t) of the LPC residual within each period is modelled as a polynomial $M(t) = P(t/T0)$ with $t \in [0,T0]$ and is estimated using the modulus of the Hilbert transform of the signal. In order to preserve continuity between adjacent periods, the polynomial fit is performed on a window centred in the middle of the period and equals to 1.2 times the length of the period.

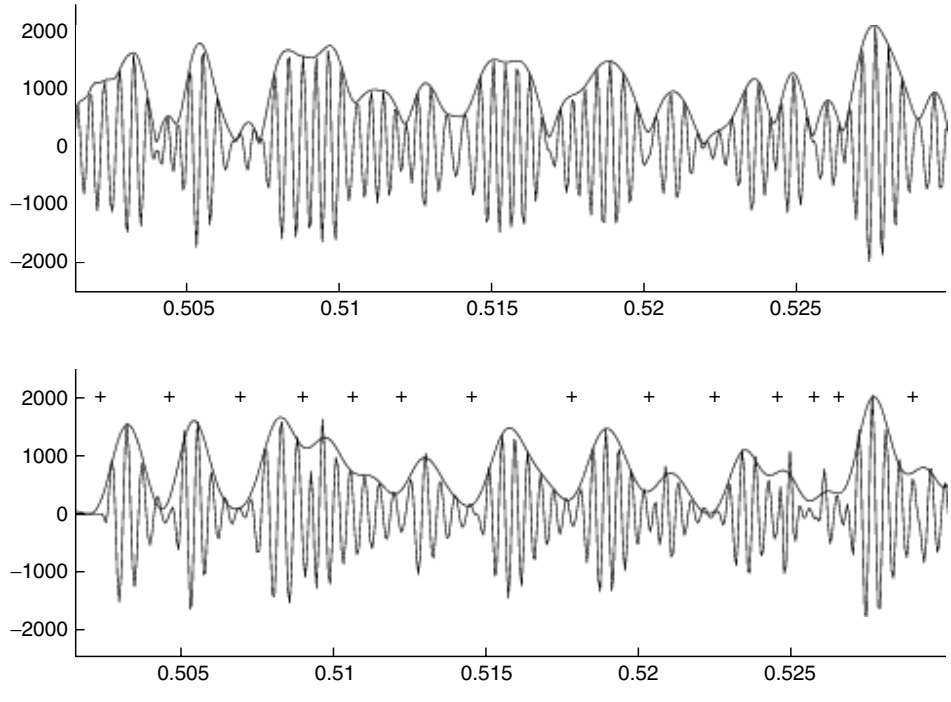


Figure 3.10 Top: original sample of a frequency band (1769–2803 Hz) with the modulus of the Hilbert transform superposed. Bottom: copy synthesis using FW (excitation times are marked with crosses)

Perceptual evaluation

We processed the stochastic components of VFV stimuli, where F is either a voiced fricative (those used in the evaluation of the H+N decomposition) or an unvoiced one. The stochastic components were estimated by the AH procedure (see Figure 3.11). We compared the two analysis-synthesis techniques for stochastic signals described above by simply adding the re-synthesised stochastic waveforms back to the harmonic component (see Figure 3.12).

Ten listeners participated in a preference test including the natural original. The original was preferred 80% and 71% of the time when compared to FW and

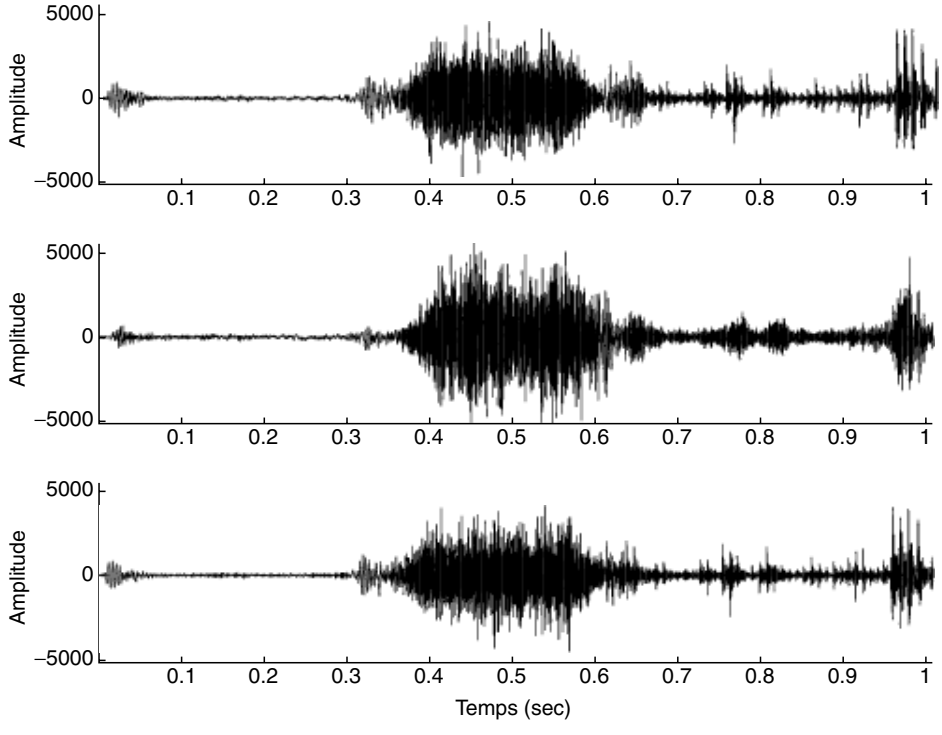


Figure 3.11 Top: original stochastic component of a nonsense word [uʒa]; Middle: copy synthesis using modulated LPC. Bottom: using FW

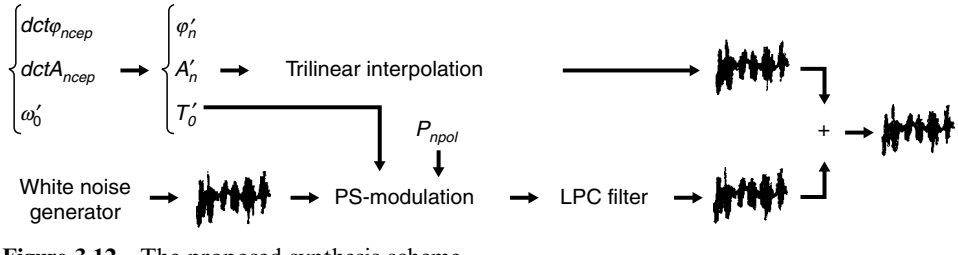


Figure 3.12 The proposed synthesis scheme

modulated LPC respectively. These results show that the quality of the copy synthesis is in both cases of good quality. Modulated LPC is preferred 67% of the time when compared to FW: this score is mainly explained by the unvoiced fricatives. This could be due to an insufficient number of subbands (we used 7 for an 8 kHz bandwidth). Modulated LPC has two further advantages: it produces fewer parameters (a constant number of parameters for each period), and is easier to synchronise with the harmonic signal. This synchronisation is highly important when manipulating the pitch period in voiced signals: Hermes (1991) showed that a synchronisation that does not mimic the physical process will result in a streaming effect. The FW representation is, however, more flexible and versatile and should be of most interest when studying voice styles.

Conclusion

We presented an accurate and flexible analysis-modification–synthesis system suitable for speech coding and synthesis. It uses a stochastic/deterministic decomposition and provides an entirely parametric representation for both components. Each period is characterised by a constant number of parameters. Despite the addition of stylisation procedures, this system achieves results on the COST 258 signal generation test array (Bailly, Chapter 4, this volume) comparable to more standard HNMs. The parametric representation offers increased flexibility for testing spectral smoothing or voice transformation procedures, and even for studying and modelling different styles of speech.

Acknowledgements

Besides COST 258 this work has been supported by ARC-B3 initiated by AUPELF-UREF. We thank Yannis Stylianou, Eric Moulines and Gaël Richard for their help and Christophe d'Alessandro for providing us with the synthetic vowels used in his papers.

References

Ahn, R. and Holmes, W.H. (1997). An accurate pitch detection method for speech using harmonic-plus-noise decomposition. *Proceedings of the International Congress of Speech Processing* (pp. 55–59). Seoul, Korea.

d'Alessandro, C., Darsinos, V., and Yegnanarayana, B. (1998). Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Transactions on Speech and Audio Processing*, *6*, 12–23.

d'Alessandro, C., Yegnanarayana, B., and Darsinos, V. (1995). Decomposition of speech signals into deterministic and stochastic components. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 760–763). Detroit, USA.

Almeida, L.B. and Silva, F.M. (1984). Variable-frequency synthesis: An improved harmonic coding scheme. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 27.5.1–4). San Diego, USA.

Charpentier, F. and Moulines, E. (1990). Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Communication*, *9*, 453–467.

Dutoit, T. and Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, *13*, 435–440.

Galas, T. and Rodet, X. (1991). Generalized functional approximation for source-filter system modeling. *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 3 (pp. 1085–1088). Genoa, Italy.

George, E.B. and Smith, M.J.T. (1997). Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Transactions on Speech and Audio Processing*, *5*, 389–406.

Gobl, C. and Ní Chasaide, A. (1992). Acoustic characteristics of voice quality. *Speech Communication*, *11*, 481–490.

Hamon, C., Moulines, E., and Charpentier, F. (1989). A diphone synthesis system based on time domain prosodic modification of speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, (pp. 238–241). Glasgow, Scotland.

Harris, F.J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings IEEE*, *66*, 51–83.

Hermes, D.J. (1991). Synthesis of breathy vowels: Some research methods. *Speech Communication*, *10*, 497–502.

Klatt, D.H. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1278–1281). Paris, France.

Laroche, J., Stylianou, Y., and Moulines, E. (1993). HNS: Speech modification based on a harmonic + noise model. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 550–553). Minneapolis, USA.

Macon. M.W. (1996). Speech synthesis based on sinusoidal modeling. Unpublished PhD thesis, Georgia Institute of Technology.

McAulay, R.J. and Quatieri, T.F. (1986). Speech analysis-synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-34*, *4*, 744–754.

Papoulis, A. (1986). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill.

Pollard, M.P., Cheetham, B.M.G., Goodyear, C.C., Edgington, M.D., and Lowry, A. (1996). Enhanced shape-invariant pitch and time-scale modification for concatenative speech synthesis. *Proceedings of the International Conference on Speech and Language Processing* (pp. 1433–1436). Philadelphia, USA.

Puckette, M.S. and Brown, J.C. (1998). Accuracy of frequency estimates using the phase vocoder. *IEEE Transactions on Speech and Audio Processing*, *6*, 166–176.

Quatieri, T.F. and McAulay, R.J. (1986). Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-34*, *4*, 1449–1464.

Quatieri, T.F. and McAulay, R.J. (1989). Phase coherence in speech reconstruction for enhancement and coding applications. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1 (pp. 207–210). Glasgow, Scotland.

Quatieri, T.F. and McAulay, R.J. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, *40(3)*, 497–510.

Richard, G. and d'Alessandro, C. (1997). Modification of the aperiodic component of speech signals for synthesis. In J.P.H. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg (eds), *Progress in Speech Synthesis* (pp. 41–56). Springer Verlag.

Rodet, X. (1980). Time-domain formant wave function synthesis. *Computer Music Journal*, *8(3)*, 9–14.

Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, CA.

Serra, X. and Smith, J. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, *14(4)*, 12–24.

Stylianou, Y. (1996). *Harmonic Plus noise models for speech, combined with statistical methods, for speech and speaker modification*. PhD thesis, École Nationale des Télécommunications, Paris.

Yegnanarayana, B., d'Alessandro, C., and Darsinos, V. (1998). An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Transactions on Speech and Audio Processing*, *6(1)*, 1–11.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

4

The COST 258 Signal Generation Test Array

Gérard Bailly
Institut de la Communication Parlée, UMR-CNRS 5009
INPG and Université Stendhal, 46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
bailly@icp.inpg.fr

Introduction

Speech synthesis systems aim at computing signals from a symbolic input ranging from a simple raw text to more structured documents, including abstract linguistic or phonological representations such as are available in a concept-to-speech system. Various representations of the desired utterance are built during processing. All these speech synthesis systems, however, use at least a module to convert a phonemic string into an acoustic signal, some characteristics of which have also been computed beforehand. Such characteristics range from nothing – as in hard concatenative synthesis (Black and Taylor, 1994; Campbell, 1997) – to detailed temporal and spectral specifications – as in formant or articulatory synthesis (Local, 1994), but most speech synthesis systems compute at least basic prosodic characteristics, such as the melody and the segmental durations the synthetic output should have.

Analysis-Modification-Synthesis Sytems (AMSS) (see Figure 4.1) produce intermediate representations of signals that include these characteristics. In concatenative synthesis, the analysis phase is often performed off-line and the resulting signal representation is stored for retrieval at synthesis time. In synthesis-by-rule, rules infer regularities from the analysis of large corpora and re-build the signal representation at run-time.

A key problem in speech synthesis is the modification phase, where the original representation of signals is modified in order to take into account the desired prosodic characteristics. These prosodic characteristics should ideally be reflected by covariations between parameters in the entire representation, e.g. variation of the open quotient of the voiced source and of formants according to F_0 and intensity, formant transitions according to duration changes etc. Contrary to synthesis-by-rule systems, where such observed covariations may be described and implemented (Gobl and Chasaide, 1992), the ideal AMSS for concatenative systems

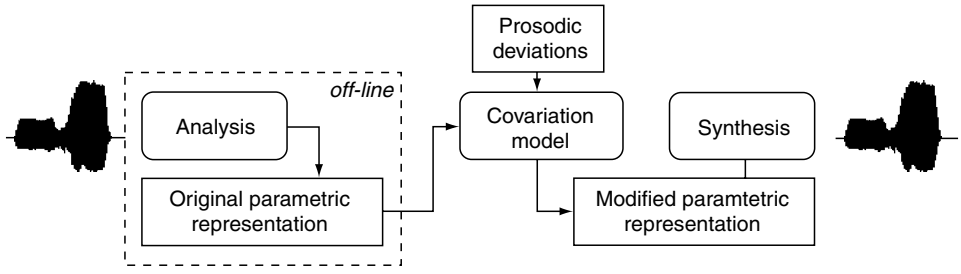


Figure 4.1 Block diagram of an AMSS: the analysis phase is often performed off-line. The original parametric representations are stored or used to infer rules that will re-build the parametric representation at run-time. Prosodic changes modify the original parametric representation of the speech signal, optimally taking covariation into account

exhibit intrinsic properties – e.g. shape invariance in the time domain (McAulay and Quatieri, 1986; Quatieri and McAulay, 1992) – that guarantee an optimal extrapolation of temporal/spectral behaviour from a reference sample. Systems with a large inventory of speech tokens replace this requirement by careful labelling and a selection algorithm that minimises distortion.

The aim of the COST 258 signal generation test array is to provide benchmarking resources and methodologies for assessing all types of AMSS. The benchmark consists in comparing the performance of AMSS on tasks of increasing difficulty: from the control of a single prosodic parameter of a single sound to the intonation of a whole utterance. The key idea is to provide reference AMSS, including the coder that is assumed to produce the most natural-sounding output: a human being. The desired prosodic characteristics are thus extracted from human utterances and given as prosodic targets to the coder under test. A server has been established to provide reference resources (signals, prosodic description of signals) and systems to (1) speech researchers, for evaluating their work with reference systems; and (2) Text-to-Speech developers, for comparing and assessing competing AMSS. The server may be accessed at the following address: http://www.icp.inpg.fr/cost258/evaluation/server/cost258_coders.

Evaluating AMSS: An Overview

The increasing importance of the evaluation/assessment process in speech synthesis research is evident: the Third International Workshop on Speech Synthesis in Jenolan Caves, Australia, had a special session dedicated to Multi-Lingual Text-to-Speech Synthesis Evaluation, and in the same year there was the First International Conference on Language Resources and Evaluation (LREC) in Grenada, Spain. In June 2000 the second LREC Conference was held in Athens, Greece. In Europe, several large-scale projects have had working groups on speech output evaluation including the EC-Esprit SAM project and the Expert Advisory Group on Language Engineering and Standards (EAGLES). The EAGLES handbook already provides a good overview of existing evaluation tasks and techniques which are described according to a taxonomy of six parameters: subjective vs. objective measurement, judgement vs. functional testing, global vs. analytic assessment,

black box vs. glass box approach, laboratory vs. field tests, linguistic vs. acoustic. We will discuss the evaluation of AMSS along some relevant parameters of this taxonomy.

Global vs. Analytic Assessment

The recent literature has been marked by the introduction of important AMSS, such as the emergence of TD-PSOLA (Hamon *et al.*, 1989; Charpentier and Moulines, 1990) and the MBROLA project (Dutoit and Leich, 1993), the sinusoidal model (Almeida and Silva, 1984; McAulay and Quatieri, 1989; Quatieri and McAulay, 1992), and the Harmonic + Noise models (Serra, 1989; Stylianou, 1996; Macon, 1996). The assessment of these AMSS is often done via 'informal' listening tests involving pitch or duration-manipulated signals, comparing the proposed algorithm to a reference in preference tests. These informal experiments are often not reproducible, use *ad hoc* stimuli[1] and compare the proposed AMSS with the authors' own implementation of the reference coder (they often use a system referenced as TDPSOLA, although not implemented by Moulines' team). Furthermore, such a global assessment procedure provides the developer or the reader with poor diagnostic information. In addition, how can we ensure that these time-consuming tests (performed in a given laboratory with a reduced set of items and a given number of AMSS) are incremental, providing end-users with increasingly complete data on a system's performance?

Black Box vs. Glass Box Approach

Many evaluations published to date either involve complete systems (often identified anonymously by the synthesis technique used, as in Sonntag *et al.*, 1999) or compare AMSS within the same speech synthesis system (Stylianou, 1998; Syrdal *et al.*, 1998). Since natural speech – or at least natural prosody – is often not included, the test only determines which AMSS is the most suitable according to the whole text-to-speech process. Moreover, the AMSS under test do not always share the same properties: TD-PSOLA, for example, is very sensitive to phase mismatch across boundaries and cannot smooth spectral discontinuities.

Judgement vs. Functional Testing

Pitch or duration manipulations are usually limited to simple multiplication/division of the speech rate or register, and do not reflect the usual task performed by AMSS of producing synthetic stimuli with natural intonation and rhythm. Manipulating the register and speech rate is quite different from a linear scaling of prosodic parameters. Listeners are thus not presented with plausible stimuli and judgements can be greatly affected by such unrealistic stimuli. The danger is thus

[1] Some authors (see, for example, Veldhuis and Yé, 1996) publishing in *Speech Communication* may nevertheless give access to the stimuli via a very useful server http://www.elsevier.nl:80/inca/publications/store/5/0/5/5/9/7 so that listeners may at least make their own judgement.

to move towards an aesthetic judgement that does not involve any reference to naturalness, i.e. that does not consider the stimuli to have been produced by a biological organism.

Discussion

We think that it would be valuable to construct a check list of formal properties that should be satisfied by any AMSS that claims to manipulate basic prosodic parameters, and extend this list to properties – such as smoothing abilities, generation of vocal fry, etc. – that could be relevant in the end user's choice. Relevant functional tests, judgement tests, objective procedures and resources should be proposed and developed to verify each property.

These tests should concentrate on the evaluation of AMSS independently of the application that would employ selected properties or qualities of a given AMSS: coding and speech synthesis systems using minimal modifications would require transparent analysis-resynthesis of natural samples whereas multi-style rule-based synthesis systems would require highly flexible and intelligible signal representation (Murray *et al*., 1996). These tests should include a natural reference and compete against it in order to fulfil one of the major goals of speech synthesis, which is the scientific goal of COST 258: improving the naturalness of synthetic speech.

The COST 258 proposal

We here propose to evaluate each AMSS on its performance of an appropriate prosodic *transplantation*, i.e. performing the task of modifying the prosodic characteristics of a source signal in order that the resulting synthetic signal has the same prosodic characteristics as a target signal. We test here not only the ability of AMSS to manipulate prosody but to answer questions such as:

- Does it perform the task in an appropriate way?
- Since manipulating some prosodic parameters such as pitch or duration modifies the timbre of sounds, is the resulting timbre acceptable or more precisely close to the timbre that could have been produced by the reference speaker if faced with the same phonological task?

This suggests that AMSS should be compared against a natural reference, in order to answer the questions above and to determine if the current description of prosodic tasks is sufficient to realise specific mappings and adequately carry the intended linguistic and paralinguistic information.

Description of tasks

The COST 258 server provides both source and target signals organised in various tasks designed to test various abilities of each AMSS. The first version of the server includes four basic tasks:

- *pitch control*: a speaker recorded the ten French vowels at different heights within his normal register.

- *duration control*: most AMSS have difficulty in stretching noise: a speaker recorded short and long versions of the six French fricatives in isolation and with a neutral vocalic substrate.
- *intonation*: AMSS should be able to control melody and segmental durations independently: a speaker recorded six versions of the same sentence with different intonation contours: a flat reference and five different modalities and prosodic attitudes (Morlec *et al.*, 2001).
- *emotion*: we extend the previous task to emotional prosody in order to test if prosodic descriptors of the available signals are sufficient to perform the same task for different emotions.

In the near future, a female voice will be added and a task to assess smoothing abilities will be included. AMSS are normally language-independent and can process any speech signal given an adequate prosodic description that could perhaps be enriched to take account of specific time/frequency characteristics of particular sounds (see below). Priority is not therefore given to a multi-lingual extension of the resources.

Physical resources

The server supplies each signal with basic prosodic descriptors (see Figure 4.2). These descriptors are stored as text files:

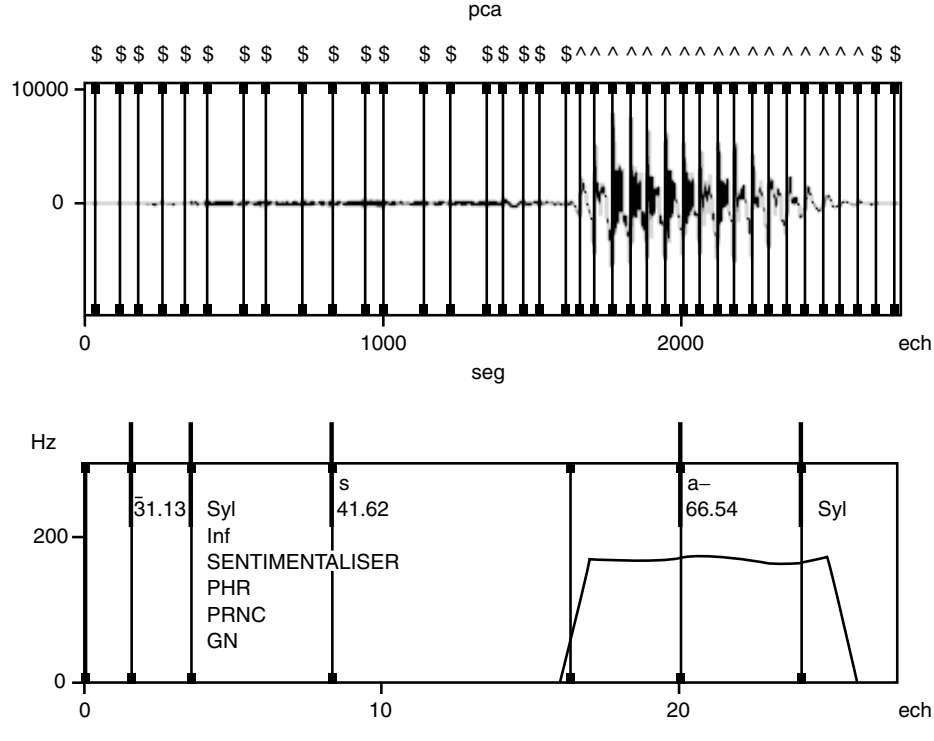


Figure 4.2 Prosodic descriptors of a sample signal. Top: pitch marks; Bottom: segmentation

- Segmentation files (extension .seg) contain the segment boundaries. Short-term energy of the signal (dB) at segment 'centres' is also available.
- Pitch mark files (extension .pca) contain onset landmarks for each period (marked by ^). Melody can thus be easily computed as the inverse of the series of periods. Additional landmarks have been inserted: burst onsets (marked by !) and random landmarks in unvoiced segments or silences (marked by $).

All signals are sampled at 16 kHz and time landmarks are given in number of samples. All time landmarks have been checked by hand.[2]

The Rules: Performing the Tasks

Each AMSS referenced in the server has fulfilled various tasks all consisting in transplanting prosody of various target samples onto a source sample (identified in all tasks with a filename ending with NT). In order to perform these various transplantation tasks, an AMSS can only use the source signal together with the source and target prosodic descriptors.

A discussion list will be launched in order to discuss what additional prosodic descriptors that can be semi-automatically determined should be added to the resources.

Evaluation Procedures

Besides providing reference resources to AMSS developers, the server will also gather and propose basic methodologies to evaluate the performance of each AMSS. In the vast majority of cases, it is difficult or impossible to perform mechanical evaluations of speech synthesis, and humans must be called upon in order to evaluate synthetic speech. There are two main reasons for this: (1) humans are able to produce judgements without any explicit reference and there is little hope of knowing exactly how human listeners process speech stimuli and compare two realisations of the same linguistic message; (2) speech processing is the result of a complex mediation between top-down processes (*a priori* knowledge of the language, the speaker or the speaking device, the situation and conditions of the communication, etc.) and signal-dependent information (speech quality, prosody, etc.). In the case of synthetic speech, the contribution of top-down processes to the overall judgement is expected to be important and no quantitative model can currently take into account this contribution in the psycho-acoustic models of speech perception developed so far.

However, the two objections made above are almost irrelevant for the COST 258 server: all tests are made with an actual reference and all stimuli have to conform to prosodic requirements so that no major qualitative differences are expected to arise.

[2] Please report any mistakes to the author (bailly@icp.inpg.fr).

Objective vs. Subjective Evaluation

Replacing time-consuming experimental work with an objective measurement of an estimated perceptual discrepancy between a signal and a reference thus seems reasonable but should be confirmed by examining the correlation with subjective quality (– see, for example, the effort in predicting boundary discontinuities (Klabbers and Veldhuis, 1998).

Currently there is no objective measure which correlates very well with human judgements. One reason for this is that a single frame only makes a small contribution to an objective measure but may contain an error which renders an entire utterance unacceptable or unintelligible for a human listener. The objective evaluation of prosody is particularly problematic, since precision at some points is crucial but at others is unimportant. Furthermore, whereas objective measures deliver time-varying information, human judgements consider the entire stimulus. Although gating experiments or online measures (Hansen and Kollmeier, 1999) may give some time-varying information, no comprehensive model of perceptual integration is available that can directly make the comparison of these time-varying scores possible.

On the other hand, subjective tests use few stimuli – typically a few sentences – and are difficult to replicate. Listeners may be influenced by factors other than signal quality especially when the level of quality is high. They are particularly sensitive to the phonetic structure of the stimuli and may not be able to judge the speech quality for foreign sounds. Listeners are also unable to process 'speech-like' stimuli.

Distortion Measures

Several distortion measures have been proposed in the literature that are supposed to correlate with speech quality (Quackenbush *et al.*, 1988). Each measure focuses on certain important temporal and spectral aspects of the speech waveform and it is very difficult to choose a measure that perfectly mimics the global judgement of listeners. Some measures take into account the importance of spectral masses and neglect or minimise the importance of distortions occurring in spectral bands with minimal energy (Klatt, 1982). Other measures include a speech production model, such as the stylisation of the spectrum by LPC.

Instead of choosing a single objective measure to evaluate spectral distortion, we chose here to compute several distortion measures and select a compact representation of the results that enhances the differences among the AMSS made available.

Following proposals made by Hansen and Pellom (1998) for evaluating speech enhancement algorithms, we used three measures: the Log-Likelihood ratio measure (LLR), the Log-Area-Ratio measure (LAR), and the weighted spectral slope measure (WSS) (Klatt, 1982). The Itakura-Saito distortion (IS) and the segmental signal-to-noise ratio (SNR) used by Hansen and Pellom were discarded since the temporal organisation of these distortion measures was difficult to interpret.

We will not evaluate the temporal distortion separately since the task already includes timing constraints – which can also be enriched – and temporal distortions will be taken into account in the frame-by-frame comparison process.

Evaluation

As emphasised by Hansen and Pellom (1998), the impact of noise on degraded speech quality is non-uniform. Similarly, an objective speech quality measure computes a level of distortion on a frame-by-frame basis. The effect of modelling noise on the performance of a particular AMSS is thus expected to be time-varying (see Figure 4.3). Although it is desirable to characterise each AMSS by its performance on each individual segment of speech, we performed a first experiment using the average and standard deviation of distortion measures for each task performed by each AMSS and evaluated by the three measures LAR, LLR and WSS, excluding comparison with reference frames with an energy below 30 dB.

Each AMSS is thus characterised by a set of 90 average distortions (3 distortion measures $\times$ 15 tasks $\times$ 2 characteristics (mean, std)). Different versions of 5 systems (TDPICP, c1, c2, c3, c4) were tested: 4 initial versions (TDPICP0,[3] c1_0, c2_0, c3_0, c4_0) processed the benchmark. The first results were presented at the Cost258 Budapest meeting in September 1997. After a careful examination of the results, improved versions of three systems (c1_0, c2_0, c4_0) were also tested.

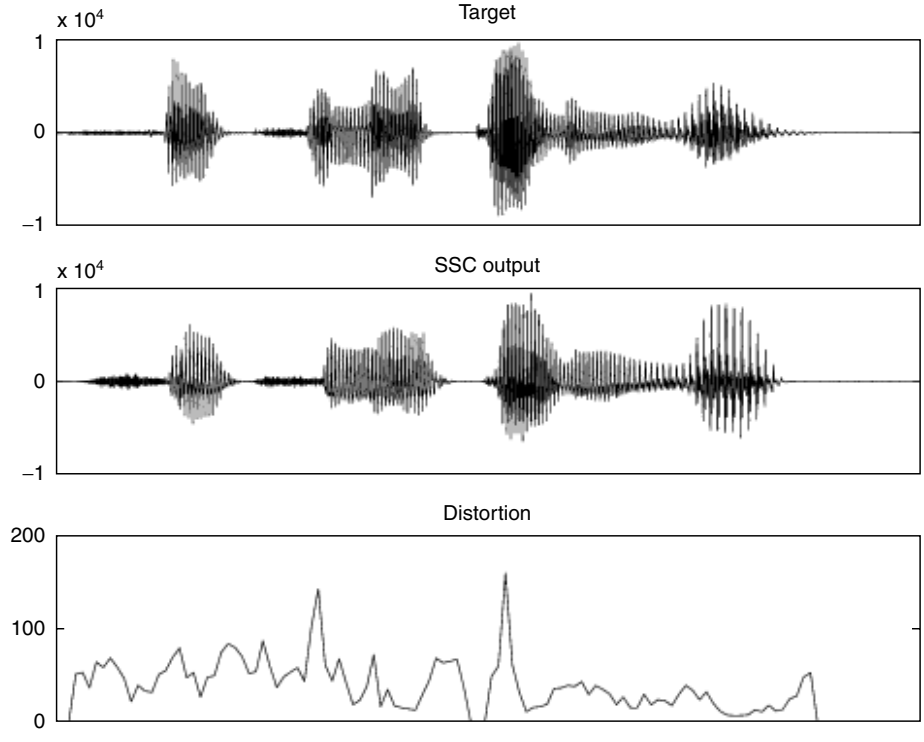


Figure 4.3 Variable impact of modelling error on speech quality. WSS quality measure versus time is shown below the analysed speech signal

[3] This robust implementation of TDPSOLA is described in (Bailly *et al.*, 1992). It mainly differs from Charpentier and Moulines (1990) in its windowing strategy that guarantees a perfect reconstruction in the absence of prosodic modifications.

We added four reference 'systems': the natural target (ORIGIN) and the target degraded by three noise levels (10 dB, 20 dB and 30 dB).

In order to produce a representation that reflects the global distance of each coder from the ORIGIN and maximises the difference among the AMSS, this set of 9×90 average distortions was projected onto the first factorial plane (see Figure 4.4) using a normalised principal component analysis procedure. The first, second and third components explain respectively 79.3%, 12.2% and 5.4% of the total variance in Figure 4.4.

Comments

We also projected the mean characteristics obtained by the systems on each of the four tasks (VO, FD, EM, AT) considering the others null. Globally, all AMSS correspond to a SNR of 20 dB. All improved versions resulted in bringing systems closer to the target. This improvement is quite substantial for systems c1 and c2, and demonstrates at least that the server provides the AMSS developers with useful diagnostic tools. Finally, two systems (c1_1, c2_1) seem to outperform the reference TDPSOLA analysis-modification-synthesis system.

The relative placement of the noisy signals (10 dB, 20 dB, 30 dB) and of the tasks (VO, FD, EM, AT) shows that the first principal component (PC) correlates with the SNR whereas the second PC correlates with the ratio between voicing/noise distortion – explained by the fact that FD and VO are placed at the extreme and that a 10 dB SNR has a lower ordinate than the higher SNRs. Distortion measures used here are in fact very sensitive to formant mismatches and when they are drowned in noise, the measures increase very rapidly. We would thus expect that systems c2_0 and c3_0 had an inadequate processing of unvoiced sounds, which is known to be true.

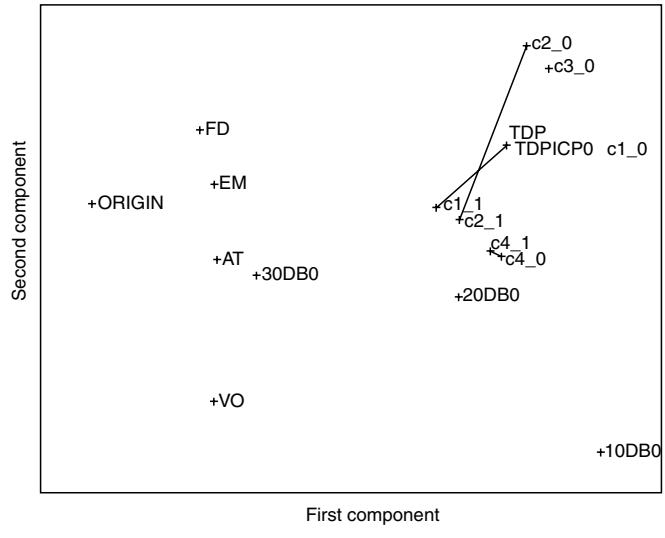
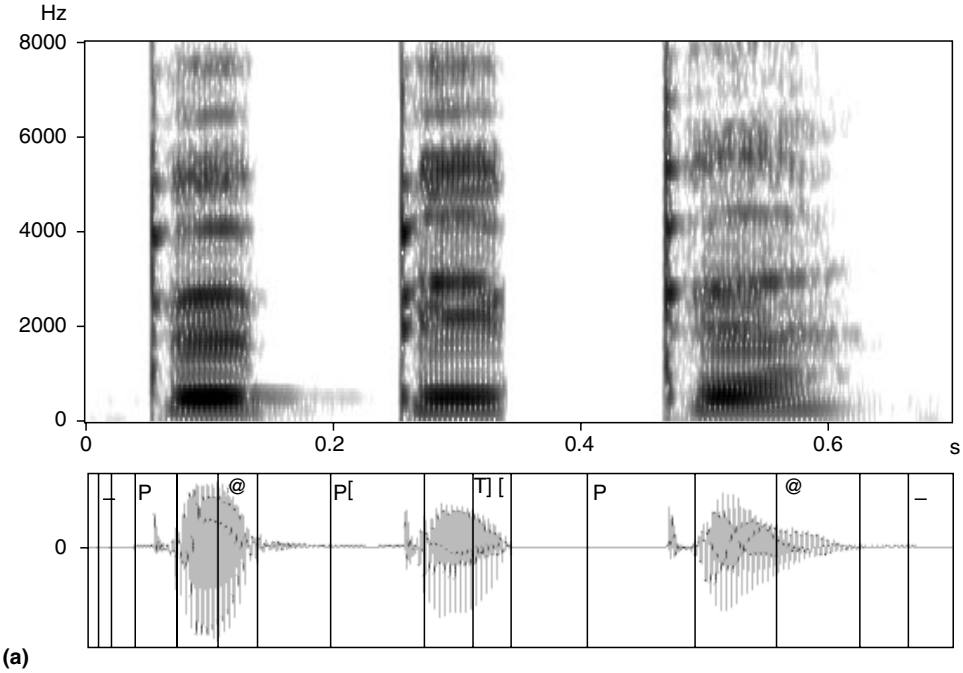
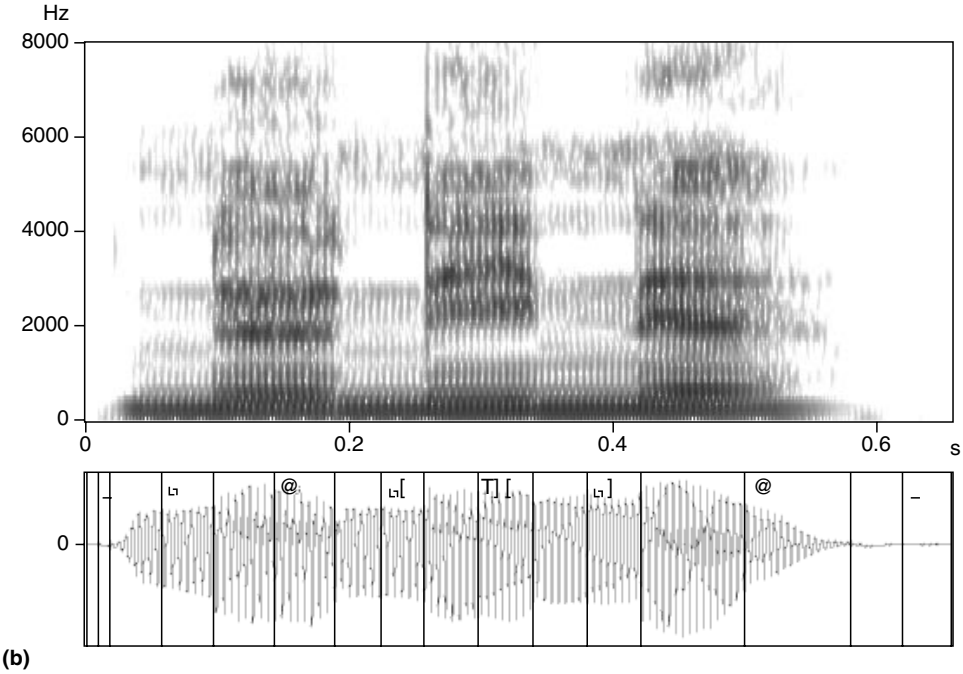


Figure 4.4 Projection of each AMSS on the first factorial plane. Four references have been added: the natural target and the target degraded by 10, 20 and 30 dB noise. c1_1, c2_1, c4_1 are improved version of respectively c1_0, c2_0, c4_0 made after a first objective evaluation


Hz
8000
6000
4000
2000
0
0
0.2
0.4
0.6
s
_
P
@
P[
T] [
P
@
_
0

(a)


Hz
8000
6000
4000
2000
0
0
0.2
0.4
0.6
s
_
ɹ
@
ɹ[
T] [
ɹ]
@
_
0

(b)

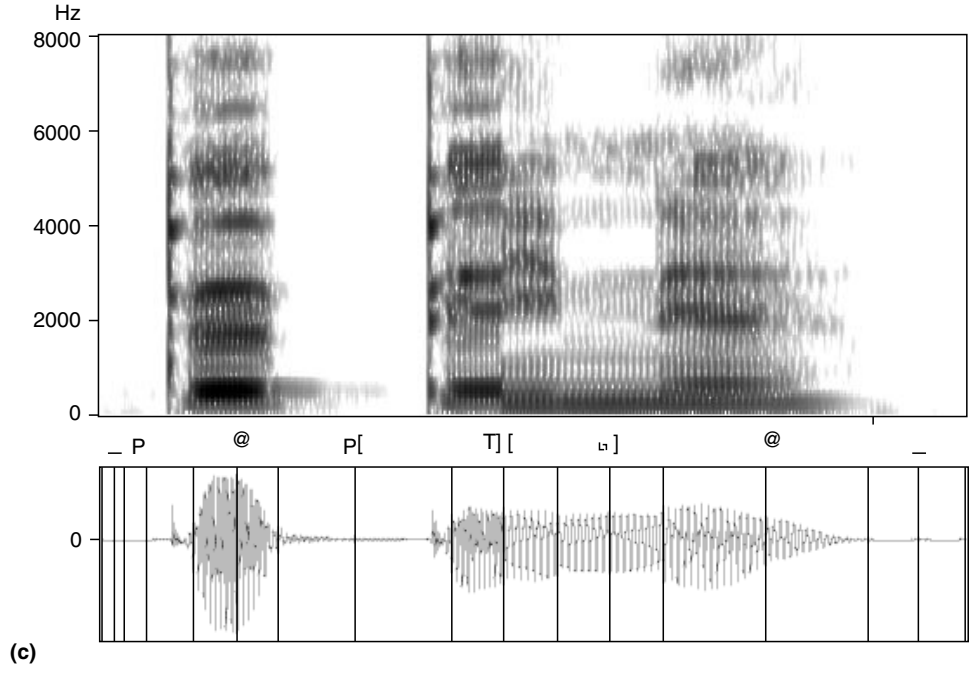


Figure 4.5 Testing the smoothing abilities of AMSS. (a) and (b) the two source signals [pəpipə] and [nəninə] (c) the hard concatenation of two signals at the second vocalic nuclei with an important spectral jump due to the nasalised vowel that AMSS will have to smooth

Conclusion

The Cost 258 signal generation test array should become a helpful tool for AMSS developers and TTS designers. It provides AMSS developers with the resources and methodologies needed to evaluate their work against various tasks and results obtained by reference AMSS.[4] It provides TTS designers with a benchmark to characterise and select the AMSS which exhibits the desired properties with the best performance.

The Cost 258 signal generation test array aims to develop a check list of the formal properties that should be satisfied by any AMSS, and extend this list to any parameter that could be relevant in the end user's choice. Relevant functional tests should be proposed and developed to verify each property. The server will grow in the near future in two main directions: we will incorporate new voices for each task – especially female voices – and new tasks. The first new task will be launched to test smoothing abilities, and will consist in comparing a natural utterance with a synthetic replica built from two different source segments instead of one (see Figure 4.5).

[4] We expect to inherit very soon the results obtained by the reference TD-PSOLA implemented by Charpentier and Moulines (1990).

Acknowledgements

This work has been supported by Cost 258 and ARC-B3 initiated by AUPELF-UREF. We thank all researchers who processed the stimuli of the first version of this server, in particular Eduardo Rodriguez Banga, Darragh O'Brien, Alex Monaghan and Miguel Gascuena. A special thanks to Esther Klabbers and Erhard Rank.

References

Almeida, L.B. and Silva, F.M. (1984). Variable-frequency synthesis: An improved harmonic coding scheme. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pages 27.5.1–4.). San Diego, USA.

Bailly, G., Barbe, T., and Wang, H. (1992). Automatic labelling of large prosodic databases: Tools, methodology and links with a text-to-speech system. In G. Bailly and C. Benoît, (eds) *Talking Machines: Theories, Models and Designs* (pp. 323–333). Elsevier B.V.

Black, A.W. and Taylor, P. (1994). CHATR: A generic speech synthesis system. *COLING-94*, Vol. II, 983–986.

Campbell, W.N. (1997). Synthesizing spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi (eds), *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 165–186). Springer Verlag.

Charpentier, F. and Moulines, E. (1990). Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Communication*, *9*, 453–467.

Dutoit, T. and Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, *13*, 435–440.

Gobl, C. and Chasaide. N. (1992). Acoustic characteristics of voice quality. *Speech Communication*, *11*, 481–490.

Hamon, C., Moulines, E., and Charpentier, F. (1989). A diphone synthesis system based on time domain prosodic modification of speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 238–241.

Hansen, J.H.L. and Pellom, B.L. (1998). An effective quality evaluation protocol for speech enhancement algorithms. *Proceedings of the International Conference on Speech and Language Processing*, *6*, 2819–2822.

Hansen, M. and Kollmeier, B. (1999). Continuous assessment of time-varying speech quality. *Journal of the Acoustical Society of America*, *105*, 2888–2899.

Klabbers, E. and Veldhuis, R. (1998). On the reduction of concatenation artefacts in diphone synthesis. *Proceedings of the International Conference on Speech and Language Processing*, *5*, 1983–1986.

Klatt, D.H. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1278–1281). Paris, France.

Local, J. (1994). Phonological structure, parametric phonetic interpretation and natural-sounding synthesis. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition* (pp. 253–270). Wiley and Sons.

McAnley, R.J. and Quatieri, T.F. (1986). Speech analysis-synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *ASSP-34(4)*, 744–754.

Macon, M.W. (1996). unpublished PhD thesis, Georgia Institute of Technology.

Morlec, Y., Bailly, G., and Aubergé, V. (2001) Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication*, 33–4, 357–371.

Murray I.R., Arnott J.L., and Rohwer, E.A. (1996). Emotional stress in synthetic speech: Progress and future directions. *Speech Communication*, *20*, 85–91.

Quackenbush, S.R., Barnwell, T.P., and Clements, M.A. (1988). *Objective Measures of Speech Quality*. Prentice-Hall.

Quatieri, T.F. and McAulay, R.J. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, *40*, 497–510.

Serra X. (1989). *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, CA.

Sonntag, G.P., Portele, T., Haas, F., and Köhler, J. (1999). Comparative evaluation of six German TTS systems. *Proceedings of the European Conference on Speech Communication and Technology*, *1*, 251–254. Budapest.

Stylianou, Y. (1996). *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, École Nationale des Télécommunications, Paris.

Stylianou, Y. (1998). Concatenative speech synthesis using a harmonic plus noise model. *ESCA/COCOSDA Workshop on Speech Synthesis* (pp. 261–266). Jenolan Caves, Australia.

Syrdal, A.K, Möhler, G., Dusterhoff, K., Conkie, A., and Black, A.W. (1998). Three methods of intonation modeling. *ESCA/COCOSDA Workshop on Speech Synthesis* (pp. 305–310). Jenolan Caves, Australia.

Veldhuis, R. and Yé, H. (1996). Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time Fourier transform. *Speech Communication*, *18*, 257–279.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

5

Concatenative Text-to-Speech Synthesis Based on Sinusoidal Modelling

Eduardo Rodríguez Banga, Carmen García Mateo and Xavier Fernández Salgado

Signal Theory Group (GTS), Dpto. Tecnologías de las Comunicaciones, ETSI Telecomunicación, Campus Universitario, Universidad de Vigo, 36200 Vigo, Spain
erbanga@gts.tsc.uvigo.es

Introduction

Text-to-speech systems based on concatenative synthesis are nowadays widely employed. These systems require an algorithm that allows concatenating the speech units and modifying their prosodic parameters to the desired values. Among these algorithms, TD-PSOLA (Moulines and Charpentier, 1990) is the best known due to its simplicity and the high quality of the resulting synthetic speech. This algorithm makes use of the classic overlap-add technique and a set of pitch marks that is employed to align the speech segments before summing them. Since it is a time-domain algorithm, it does not permit modifying the spectral characteristics of the speech directly and, consequently, its main drawback is said to be the lack of flexibility. For instance, the restricted range for time and pitch scaling has been widely discussed in the literature.

During the past few years, an alternative technique has become increasingly important: sinusoidal modelling. It is a more complex algorithm and computationally more expensive, but very flexible. The basic idea is to model every significant spectral component as a sinusoid. This is not a new idea, because in the previous decades some algorithms based on sinusoidal modelling had been proposed. Nevertheless, when used for time and pitch scaling, the quality of the synthetic speech obtained with most of these techniques was reverberant because of an inadequate phase modelling. In Quatieri and McAulay (1992), a sinusoidal technique is presented that allows pitch and time scaling without the reverberant effect of previous models. In the following we will refer to this method as the Shape-Invariant Sinusoidal Model (SISM). The term 'shape-invariant' refers to maintaining most of the temporal structure of the speech in spite of pitch or duration modifications.

In this chapter we present our work in the field of concatenative synthesis by means of sinusoidal modelling. The SISM provides quite good results when applied to a continuous speech signal but, when applied to text-to-speech synthesis, some problems appear. The main source of difficulties resides in the lack of continuity in speech units that were extracted from different contexts. In order to solve these problems, and based on the SISM, we have proposed (Banga *et al.*, 1997) a Pitch-Synchronous Shape-Invariant Sinusoidal Model (PSSM) which has now been further improved. The PSSM makes use of a set of pitch marks that are employed to carry out a pitch-synchronous analysis and as reference points when modifying the prosodic parameters of the speech or when concatenating speech units.

The outline of this chapter is as follows: in the next section, we briefly outline the principles of the Shape-Invariant Sinusoidal Model; second, we describe the basis and implementation of the Pitch Synchronous Sinusoidal Model and we present some results; finally, we discuss the application of the PSSM to a concatenative text-to-speech system, and offer some conclusions and some guidelines for further work.

The Shape-Invariant Sinusoidal Model (SISM)

This algorithm was originally proposed (Quatieri and McAulay, 1992) for time-scale and pitch modification of speech. This method works on a frame by frame basis, modelling the speech as the response of a time-varying linear system, $h(t)$, which models the response of the vocal tract and the glottis, to an excitation signal. Both the excitation signal, $e(t)$, and the speech signal, $s(t)$, are represented by a sum of sinusoids, that is:

$$e(t) = \sum_{j=1}^{J(t)} a_j(t) \cos[\Omega_j(t)] \tag{1}$$

$$s(t) = \sum_{j=1}^{J(t)} A_j(t) \cos[\theta_j(t)] \tag{2}$$

where $J(t)$ denotes the number of significant spectral peaks in the short-time spectrum of the speech signal, and where $a_j(t)$, $A_j(t)$ and $\Omega_j(t)$, $\theta_j(t)$ denote the amplitudes and instantaneous phases of the sinusoidal components. The amplitudes and instantaneous phases of the excitation and the speech signal are related by the following expressions:

$$A_j(t) = a_j(t) M_j(t) \tag{3}$$

$$\theta_j(t) = \Omega_j(t) + \psi_j(t) \tag{4}$$

where $M_j(t)$ and $\Psi_j(t)$ represent the magnitude and the phase of the transfer function of the linear system at the frequency of the j-th spectral component.

The excitation phase is supposed to be linear. In analogy with the classic model which considers that during voiced speech the excitation signal is a periodic pulse train, a parameter called 'pitch pulse onset time', t_0, is defined (McAulay and

Quatieri, 1986a). This parameter represents the time at which all the excitation components are in phase. Assuming that the j-th peak frequency, ω_j, is nearly constant over the duration of a speech frame, the resulting expression for the excitation phases is:

$$\Omega_j(t) = (t - t_0)\omega_j \tag{5}$$

In accordance with the expressions (4) and (5), the system phase, $\Psi_j(t)$, can be estimated as the difference between the measured phases at the spectral peaks and the excitation phase:

$$\psi_j(t) = \theta_j(t) - (t - t_0)\omega_j \tag{6}$$

In the Shape-Invariant Sinusoidal Model, duration modifications are obtained by time scaling the excitation amplitudes and the magnitude and the phase envelope of the linear system. Pitch modifications can be achieved by scaling the frequencies of the spectral peaks to the desired values, estimating the magnitude and the phase of the linear system at those new frequencies, and taking into account that the new pitch pulse onset times are placed in accordance with the new pitch period. The main achievement of the SISM is that it basically maintains the phase relations among the different sinusoidal components. As a result, the modified speech waveform is quite similar to the original and, consequently, it does not sound reverberant.

Since unvoiced sounds must be kept unmodified under pitch modifications, McAulay and Quatieri have also proposed a method to estimate a cut-off frequency (McAulay and Quatieri, 1990) above which the spectral components are considered unvoiced and left unmodified. According to our experience, this voicing estimation, as any other estimation, is tied to errors that may result in voicing some originally unvoiced segments. Although this effect is nearly imperceptible for moderated pitch modifications, it could be particularly important when large changes are carried out. Fortunately, this fact will not represent a severe limitation in text-to-speech synthesis, because we have some prior knowledge about the voiced or unvoiced nature of the sounds we are processing.

The Pitch-Synchronous Shape-Invariant Sinusoidal Model

Basis of the model

The previous model offers quite good results when applied to continuous speech for modification of the duration or the pitch. Nevertheless, we have observed (Banga *et al.*, 1997) that the estimated positions of the pitch pulse onset times (relative to a period) show some variability, apart from some clear errors, that may distort the periodicity of the speech signal.

The problem of the variability in the location of the pitch pulse onset times becomes more important in concatenative synthesis. In this case, we have to concatenate speech units that were extracted from different words in different contexts. As a result, the waveforms of the common allophone (the allophone at which the units are pasted) may be quite different and the relative position (within the pitch period) of the pitch pulse onset times may vary. If this circumstance is not taken

into account, alterations of the periodicity may appear at junctions between speech units, seriously affecting the synthetic speech quality. An interesting interpretation arises from considering the pitch pulse onset time as a concatenation point between speech units. When the relative positions of the pitch pulse onset times in the common allophone are not very similar, the periodicity of the speech is broken at junctions. Therefore, it is necessary to define a more stable reference or, alternatively, a previous alignment of the speech units.

With the TD-PSOLA procedure in mind, we decided to employ a set of pitch marks instead of the pitch pulse onset times. These pitch marks are placed pitch-synchronously on voiced segments and at a constant rate on unvoiced segments. On a stationary segment, the pitch marks, t_m, are located at a constant distance, t_d, from the authentic pitch pulse onset time, the glottal closure instant (GCI), T_0. Therefore,

$$t_m = T_0 + t_d \tag{7}$$

By substitution in equation (6), we obtain that the phase of the l-th spectral component at $t = t_m$ is given by

$$\theta_j(t_m) = \psi_j(t_m) + \omega_j t_d \tag{8}$$

i.e., apart from a linear phase term, it is equal to the system phase. Assuming local stationarity, the difference between the glottal closure instant and t_d is maintained along consecutive periods. Thus, the linear phase component is equivalent to a time shift, which is irrelevant from a perceptual point of view. We can also assume that the system phase is slowly varying, so the system phases at consecutive pitch pulse onset times (or pitch marks) will be quite similar. This last assumption is illustrated in Figure 5.1, where we can observe the spectral envelope and the phase response at four consecutive frames of the sound [a].

The previous considerations can be extended to the case of concatenating two segments of a same allophone that belong to different units obtained from different words. They will be especially valid in the central periods of the allophone, where the coarticulation effect is minimised although, of course, it will also depend on the variability of the speaker's voice, i.e., on the similarity of the different recordings of the allophones. From equation (9) we can also conclude that any set of time marks placed at a pitch rate can be used as pitch marks, with independence of their location within the pitch period (a difference with TD-PSOLA). Nevertheless, it is crucial to follow a consistent criterion to establish the position of the pitch marks.

Prosodic modification of speech signals

The PSSM has been successfully applied to prosodic modifications of continuous speech signals sampled at 16 kHz. In order to reduce the number of parameters of the model, we have assumed that, during voiced segments, the frequencies of the spectral components are harmonically related. During unvoiced segments a constant low pitch (100 Hz) is employed.

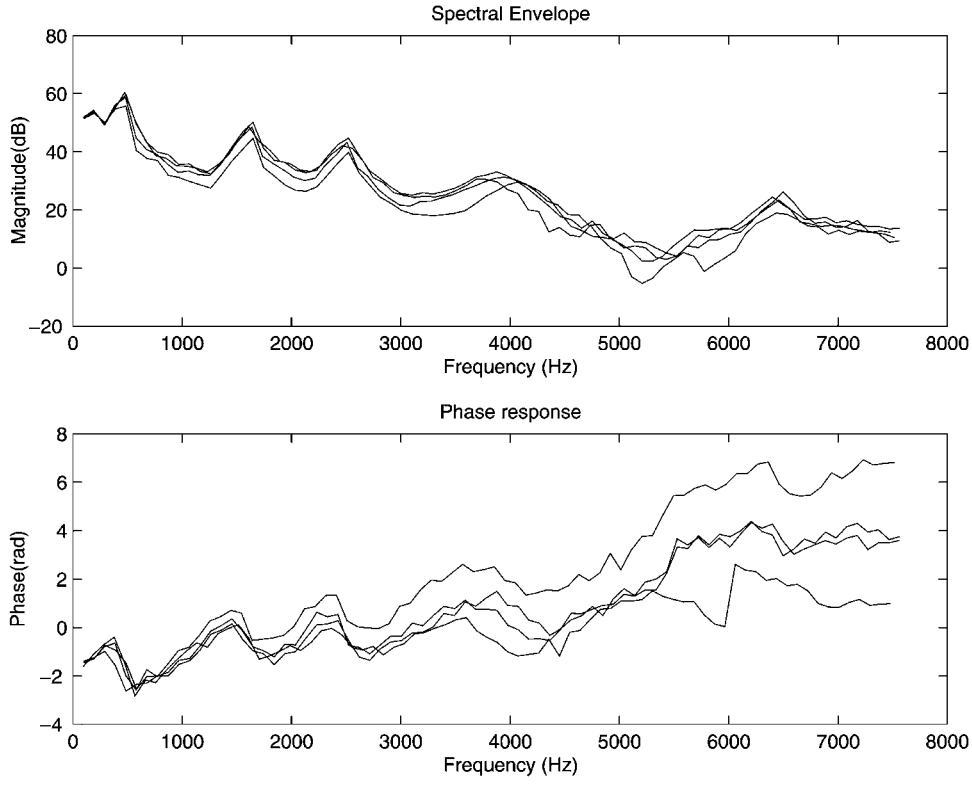


Figure 5.1 Spectral magnitude and phase response at four consecutive frames of sound [a]

Analysis

Pitch marks are placed at a pitch rate during voiced segments and a constant rate (10 ms) during unvoiced segments. A Hamming window (20–30 ms length) is centered at every pitch mark to obtain the different speech frames. The local pitch is simply calculated as the difference between consecutive pitch marks. An FFT of every frame is computed. The complex amplitudes (magnitude and phase) of the spectral components are determined by sampling the short-time spectrum at the pitch harmonics. As a result of the pitch-synchronous analysis, the system phase at the frequencies of the pitch harmonics is considered to be equal to the measured phases of the spectral components (apart from a nearly constant linear phase term). Finally, the value of the pitch period and the complex amplitudes of the pitch harmonics are stored.

Synthesis

The synthesis stage is mainly based on the Shape-Invariant Sinusoidal model. A sequence of pitch marks (or pitch pulse onset times) is generated taking into account the desired pitch and duration. These new pitch marks are employed to obtain the new excitation phases. Duration modifications affect the excitation amplitudes and

the magnitude and the phase of the linear system, which are time-scaled. With respect to pitch modifications, the magnitude of the linear system is estimated at the new frequencies by linear interpolation of the absolute value of the complex amplitudes (in a logarithmic scale), while the phase response is obtained by linear interpolation of the real and imaginary parts. As an example, in Figure 5.2 we can observe the estimated magnitudes and unwrapped phases for a pitch-scaling factor of 1.9.

Finally, the speech signal is generated as a sum of sinusoids in accordance with equation (2). Linear interpolation is employed for the magnitudes and a 'maximally smooth' third-order polynomial for the instantaneous phases (McAulay and Quatieri, 1986b). During voiced segments, the instantaneous frequencies (the first derivative of the instantaneous phases) are practically linear.

Unvoiced sounds are synthesised in the same manner as voiced sounds. Nevertheless, during unvoiced segments there is no pitch scaling and the phases, $\theta_j(t_m)$, are considered random in the interval $(-\pi, \pi)$.

In order to prevent these segments from periodicities that may appear when lengthening this type of sounds, we decided to subdivide each synthesis frame into several subframes and to randomise the phase at each subframe. This technique (Macon and Clements, 1997) was proposed in order to eliminate tonal artefacts in the ABS/OLA sinusoidal scheme. This method increases the bandwidth of each

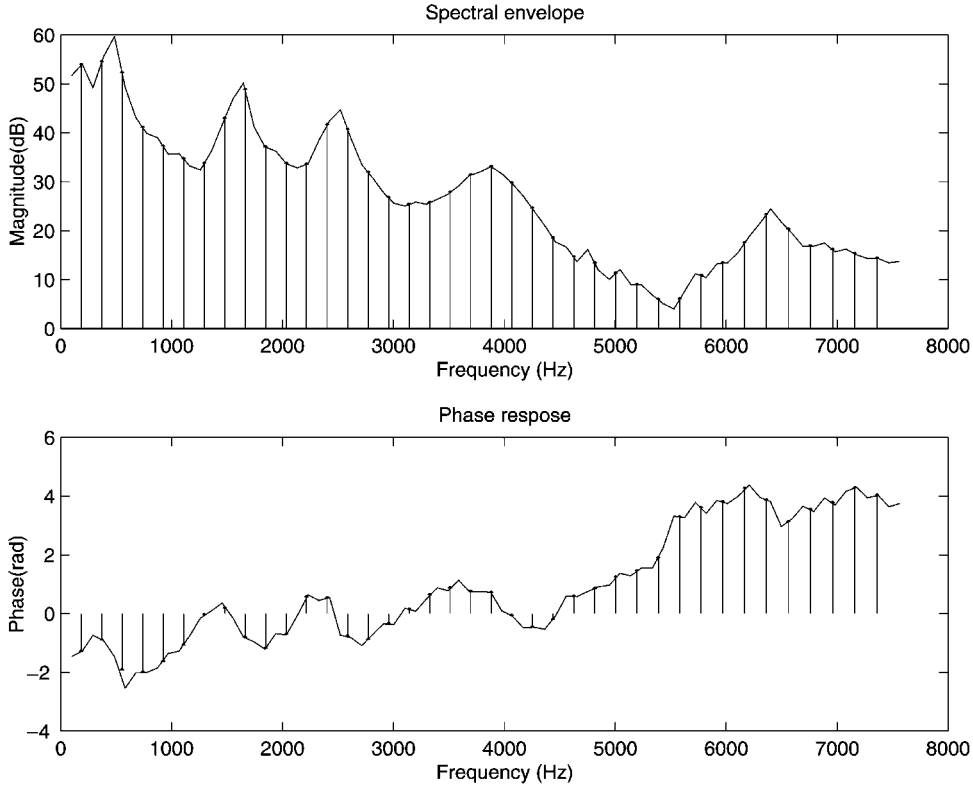


Figure 5.2 Estimated amplitudes and phases for a pitch scaling factor of 1.9

component, smoothing the short-time spectrum. The effect of phase randomisation on the instantaneous frequency is illustrated in Figure 5.3, where the instantaneous phase, the instantaneous frequency and the resulting synthetic waveform of a spectral component are represented. We can observe the fluctuations in the instantaneous frequency that increase the bandwidth of the spectral component. In spite of the sudden frequency changes at junctions between subframes (marked with dashed lines), the instantaneous phase and the synthetic signal are continuous.

Voiced fricative sounds are considered to be composed by a low-frequency periodic component and a high-frequency unvoiced component. In order to separate both contributions, a cut-off frequency is used. Several techniques can be used to estimate that cut-off frequency. Nevertheless, in some applications like text-to-speech we have some prior knowledge about the sounds we are processing and an empirical limit can be established (which may depend on the sound).

Finally, whenever pitch scaling occurs, signal energy must be adjusted to compensate for the increase or decrease in the number of pitch harmonics. Obviously, the model also allows modifying the amplitudes of the spectral components separately, and this is one of the most promising characteristics of sinusoidal modelling. Nevertheless, it is necessary to be very careful with this kind of spectral manipulation, since inadequate changes lead to very annoying effects.

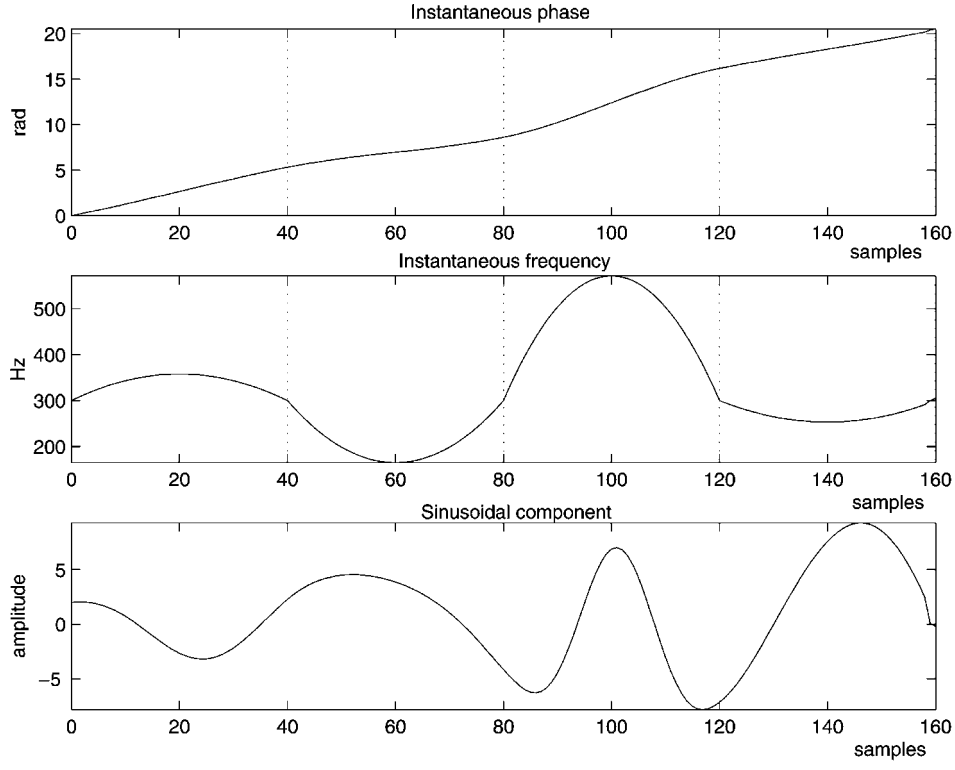


Figure 5.3 Effect of randomising the phases every subframe on the instantaneous phase and instantaneous frequency

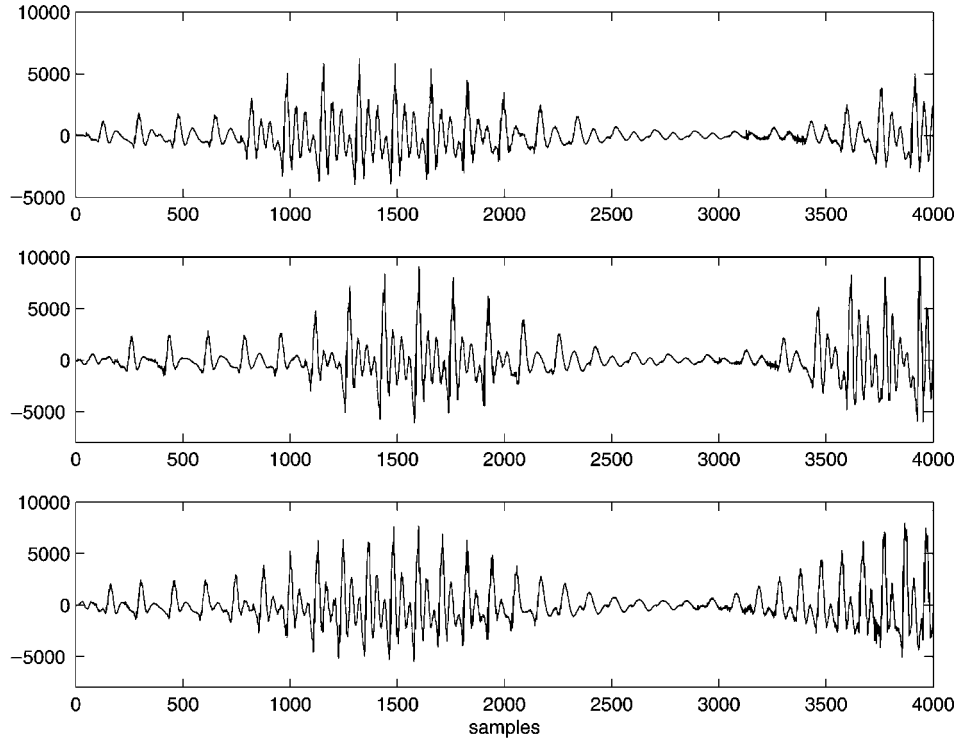


Figure 5.4 Original signal (upper plot) and two examples of synthetic signals after prosodic modification

As an example of the performance of the PSSM, in Figure 5.4, three speech segments are displayed. These waveforms correspond to an original speech signal and two synthetic versions of that signal whose prosodic parameters have been modified. In spite of the modifications, we can observe that the temporal structure of the original speech signal is basically maintained and, as a result, the synthetic signals do not present reverberant effects.

Concatenative Synthesis

In this section we discuss the application of the previous model to a text-to-speech system based on speech unit concatenation. We focus the description on our TTS system for Galician and Spanish that employs about 1200 speech units (diphones and triphones mainly) per available voice. These speech units were extracted from nonsense words that were recorded by two professional speakers (a male and a female). The sampling frequency was 16 kHz and the whole set of speech units was manually labelled. In order to determine the set of pitch marks for the speech unit database, we employed a pitch determination algorithm combined with the prior knowledge of the sound provided by the phonetic labels. During voiced segments, pitch marks were mainly placed at the local maxima (in absolute value) of the pitch periods, and during unvoiced segments they were placed every 10 ms.

The next step was a pitch-synchronous analysis of the speech unit database. Every speech frame was parameterised by the fundamental frequency and the magnitudes and the phases of the pitch harmonics. During unvoiced sounds, a fixed low pitch (100 Hz) was employed. It is important to note that, as a consequence of the pitch-synchronous analysis, the phases of the pitch harmonics are a good estimation of the system phase at those frequencies.

The synthesis stage is carried out as described in the previous section. It is necessary to emphasise that, in this model, no speech frame is eliminated or repeated. All the original speech frames are time-scaled by a factor that is a function of the original and desired durations. It is an open discussion whether or not this factor should be constant for every frame of a particular sound, that is, whether or not stationary and transition frames should be equally lengthened or shortened. At this time, with the exception of plosive sounds, we are using a constant factor.

In a concatenative TTS it is also necessary to ensure smooth transitions from one speech unit to another. It is especially important to maintain pitch continuity at junctions and smooth spectral changes. Since, in this model, the fundamental frequency is a parameter that can be finely controlled, no residual periodicity appears in the synthetic signal. With respect to spectral transitions between speech units, the

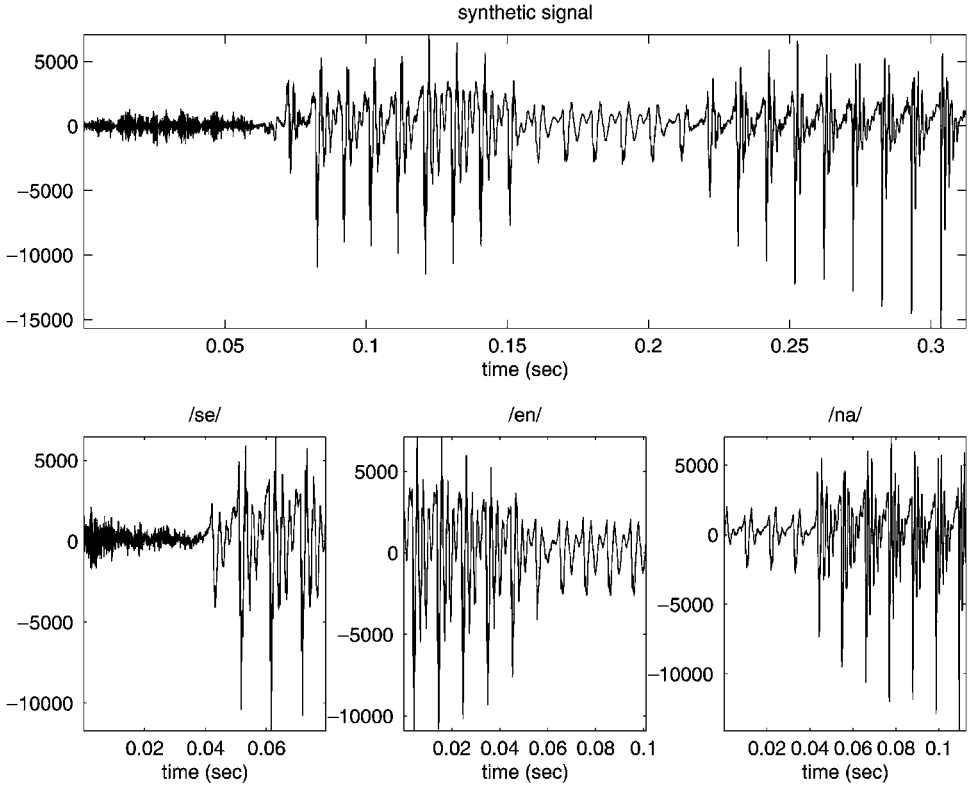


Figure 5.5 Synthetic speech signal and the corresponding original diphones

linear interpolation of the amplitudes normally provides sufficiently smooth transitions. Obviously, the longer the junction frame, the smoother the transition. So, if necessary, we can increase the factor of duration modification in this frame and reduce that factor in the other frames of the sound. Finally, another important point is to prevent our system from sudden energy jumps. This task is easily accomplished by means of a previous energy normalisation of the speech units, and by the frame-to-frame linear interpolation of the amplitudes of the pitch harmonics.

As an example of the performance of our algorithm, a segment of a synthetic speech signal (male voice) is shown in Figure 5.5, as well as the three diphones employed in the generation of that segment. We can easily observe that, in spite of pitch and duration modifications, the synthetic signal resembles the waveform of the original diphones. Comparing the diphones /se/ and /en/, we notice that the waveforms of the segments corresponding to the common phoneme [e] are slightly different. Nevertheless, even in this case, the proposed sinusoidal model provides smooth transitions between speech units, and no discontinuity or periodicity breakage appears in the waveform at junctions.

In order to show the capability of smoothing spectral transitions, a synthetic speech signal (female voice) and its narrowband spectrogram is represented in Figure 5.6. We can observe that the synthetic signal comes from the junction of

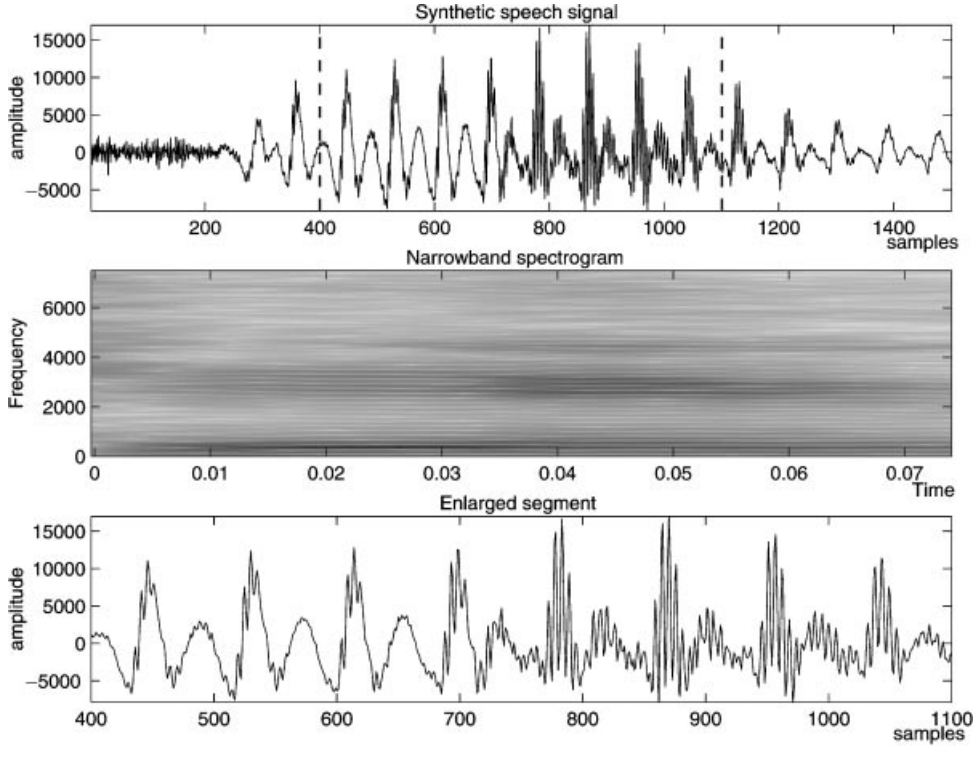


Figure 5.6 Female synthetic speech signal and its narrowband spectrogram. The speech segment between dashed lines of the upper plot has been enlarged in the bottom plot

two speech units where the common allophone has different characteristics in the time and frequency domains. In the area around the junction (shown enlarged, between the dashed lines), there is a pitch period that seems to have characteristics of contributions from the two realisations of the allophone. This is the junction frame. In the spectrogram there is no pitch discontinuity and hardly any spectral mismatch is noticed. As we have already mentioned, if a smoother transition were needed, we could use a longer junction frame. As a result, we would obtain more pitch periods with mixed characteristics.

Conclusion

In this chapter we have discussed the application of a sinusoidal algorithm to concatenative synthesis. The PSSM is capable of providing high-quality synthetic speech. It is also a very flexible method, because it allows modifying any spectral characteristic of the speech. For instance, it could be used to manipulate the spectral envelope of the speech signal. Further research is needed in this field, since inappropriate spectral manipulations can result in very annoying effects in the synthetic speech.

A formal comparison to other prosodic modification algorithms (TD-PSOLA, HNM, linear prediction models) is currently being developed in the framework of the COST 258 Signal Test Array. A detailed description of the evaluation procedure and some interesting results can be found in this volume and in Bailly *et al.* (2000). Some sound examples can be found at the web page of the COST258 Signal Test Array (http://www.icp.inpg.fr/cost258/evaluation/server/cost258_coders.html), where our system is denoted as PSSVGO, and at our own demonstration page (http://www.gts.tsc.uvigo.es/~erbanga /edemo.html).

Acknowledgements

This work has been partially supported by the 'Centro Ramón Piñeiro (Xunta de Galicia)', the European COST Action 258 'The naturalness of synthetic speech' and the Spanish CICYT under the projects 1FD97–0077–C02–C01, TIC1999–1116 and TIC2000–1005–C03–02.

References

Bailly, G., Banga, E.R., Monaghan, A., and Rank, E. (2000). The COST258 signal generation test array. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Vol.2. (pp. 651–654). Athens, Greece.

Banga, E. R., García-Mateo, C., and Fernández-Salgado, X. (1997). Shape-invariant prosodic modification algorithm for concatenative text-to-speech synthesis. *Proceedings of the 5th European Conference on Speech Communication and Technology* (pp. 545–548). Rhodes, Greece.

Macon, M. and Clements, M. (1997). Sinusoidal modeling and modification of unvoiced speech. *IEEE Transactions on Speech and Audio Processing*, *5*, 557–560.

McAulay, R.J. and Quatieri, T.F. (1986a). Phase modelling and its application to sinusoidal transform coding. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1713–1715). Tokyo, Japan.

McAulay, R.J. and Quatieri, T.F. (1986b). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *34*, 744–754.

McAulay, R.J. and Quatieri, T.F. (1990). Pitch estimation and voicing detection based on a sinusoidal model. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 249–252). Albuquerque, USA.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.

Quatieri, T.F. and McAulay, R.J. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, *40*, 497–510.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

6

Shape Invariant Pitch and Time-Scale Modification of Speech Based on a Harmonic Model

Darragh O'Brien and Alex Monaghan
Sun Microsystems Inc. and Aculab plc
darragh.obrien@sun.com and
Alex.Monaghan@aculab.com

Introduction

This chapter presents a novel and conceptually simple approach to pitch and time-scale modification of speech. Traditionally, pitch pulse onset times have played a crucial role in sinusoidal model-based speech transformation techniques. Examples of algorithms relying on onset estimation are those proposed by Quatieri and McAulay (1992) and George and Smith (1997). At each onset time all waves are assumed to be in phase, i.e. the phase of each is assumed to be some integer multiple of 2π. Onset time estimates thus provide a means of maintaining waveform shape and phase coherence in the modified speech. However, accurate onset time estimation is a difficult problem, and errors give rise to a garbled speech quality (Macon, 1996).

The harmonic-based approach described here does not rely on onset times to maintain phase coherence. Instead, post-modification waveform shape is preserved by exploiting the harmonic relationship existing between the sinusoids used to code each (voiced) frame, to cause them to be in phase at synthesis frame intervals. Furthermore, our modification algorithms are not based on PSOLA (Moulines and Charpentier, 1990) and therefore, in contrast to HNM (Stylianou *et al.*, 1995), analysis need not be pitch synchronous and the duplication/deletion of frames during scaling is avoided. Finally, time-scale expansion of voiceless regions is handled not through the use of a hybrid model but by increasing the variation in frequency of 'noisy' sinusoids, thus smoothing the spectrum and alleviating the

problem of tonal artefacts. Importantly, our approach allows for a straightforward implementation of joint pitch and time-scale modification.

Sinusoidal Modelling

Analysis

Pitch analysis is carried out on the speech signal using Entropic's pitch detection software[1] which is based on work by Talkin (1995). The resulting pitch contour, after smoothing, is used to assign an F0 estimate to each frame (zero if voiceless). Over voiced (and partially voiced) regions, the length of each frame is set at three times the local pitch period. Frames of length 20 ms are used over voiceless regions. A constant frame interval of 10 ms is used throughout analysis. A Hanning window is applied to each frame and its FFT calculated. Over voiced frames the amplitudes and phases of sinusoids at harmonic frequencies are coded. Peak picking is applied to voiceless frames. Other aspects of our approach are closely based on McAulay and Quatieri's (1986) original formulation of the sinusoidal model. For pitch modification, the estimated glottal excitation is analysed in the same way.

Time-Scale Modification

Because of the differences in the transformation techniques employed, time-scaling of voiced and voiceless speech are treated separately. Time-scale modification of voiced speech is presented first.

Voiced Speech

If their frequency is kept constant, the phases of the harmonics used to code each voiced frame repeat periodically every $2\pi/\omega_0$ s where ω_0 is the fundamental frequency expressed in rad s^{-1}. Each parameter set (i.e. the amplitudes, phases and frequencies at the centre of each analysis frame) can therefore be viewed as defining a periodic waveform. For any phase adjustment factor δ a new set of 'valid' (where valid means being in phase) phases can be calculated from

$$\psi_k^{'} = \psi_k + \omega_k \delta \tag{1}$$

Where ψ_k' is the new and ψ_k the original phase of the k^{th} sinusoid with frequency ω_k. After time-scale modification, harmonics should be in phase at each synthesis frame interval i.e. their new and original phases should be related by equation (1). Thus, the task during time-scaling is to estimate the factor δ for each frame, from which a new set of phases at each synthesis frame interval can be calculated. Equipped with phase information consistent with the new time-scale, synthesis is straightforward and is carried out as in McAulay and Quatieri (1986). A procedure for estimating δ is presented below.

After nearest neighbour matching (over voiced frames this simplifies to matching corresponding harmonics), has been carried out, the frequency track connecting the

[1] get_f0 Copyright Entropic Research Laboratory, Inc. 5/24/93.

fundamental of frame l with that of frame $l+1$ is computed as in McAulay and Quatieri (1986) and may be written as:

$$\theta_0(n) = \gamma + 2\alpha n + 3\beta n^2 \tag{2}$$

Time-scaling equation (2) is straightforward. For a given time-scaling factor, ρ, a new target phase, $\psi_0^{l+1'}$ must be determined. Let the new time-scaled frequency function be

$$\theta_0'(n) = \theta_0(n/\rho) \tag{3}$$

The new target phase, $\psi_0^{l+1'}$, is found by integrating equation (3) over the time interval ρS (where S is the analysis frame interval) and adding back the start phase ψ_0^l,

$$\int_0^{\rho S} \theta_0(n)dn + \psi_0^l = \rho S\left(\gamma + \alpha S + \beta S^2\right) + \psi_0^l \tag{4}$$

By evaluating equation (4) modulo $2\pi, \psi_0^{l+1'}$ is determined. The model (for F0) is completed by solving for α and β, again as outlined in McAulay and Quatieri (1986).

Applying the same procedure to each remaining matched pair of harmonics will, however, lead to a breakdown in phase coherence after several frames as waves gradually move out of phase. To overcome this, and to keep waves in phase, δ is calculated from (1) as:

$$\delta = \frac{\psi_0^{l+1} - \psi_0^{l+1'}}{\omega_0^{l+1}} \tag{5}$$

δ simply represents the linear phase shift from the fundamental's old to its new target phase value. Once δ has been determined, all new target phases, $\psi_k^{l+1'}$, are calculated from equation (1). Cubic phase interpolation functions may then be calculated for each sinusoid and resynthesis of time-scaled speech is carried out using equation (6):

$$s(n) = \sum_k A_k^l(n)\cos\left[\theta_k^l(n)\right] \tag{6}$$

It is necessary to keep track of previous phase adjustments when moving from one frame to the next. This is handled by Δ (see Figure 6.1) which must be applied, along with δ, to target phases thus compensating for phase adjustments in previous frames. The complete time-scaling algorithm is presented in Figure 6.1. It should be noted that this approach is different from that presented in O'Brien and Monaghan (1999a), where the difference between the time-scaled and original frequency tracks was minimised (see below for an explanation of why this approach was adopted). Here, in the interests of efficiency, the original frequency track is not computed.

Some example waveforms, taken from speech time-scaled using this method, are given in Figures 6.2, 6.3 and 6.4. As can be seen in the figures, the shape of the original is well preserved in the modified speech.

$\Delta = 0$
$\delta = 0$
For each Frame l
Begin
$\Delta = \Delta + \delta$
For ω_0
Begin
Adjust ψ_0^{l+1} by Δ
Compute frenquency track $\theta_0(n)$
Compute new frequency track $\theta_0(n)$
Solve for $\psi_0^{l+1'}$
Solve for δ
Compute phase function $\theta_0^l(n)$
End
For ω_k where $k \neq 0$
Begin
Adjust ψ_k^{l+1} by $\delta + \Delta$
Compute phase function $\theta_k^l(n)$
End
End

Figure 6.1 Pitch-scaling algorithm

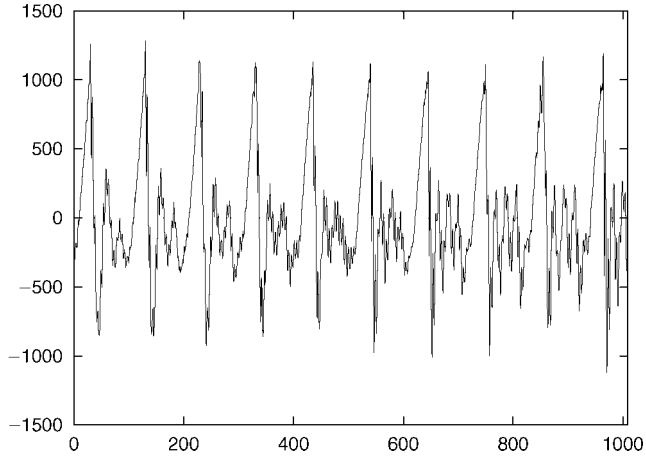


Figure 6.2 Original speech, $\rho = 1$

Voiceless Speech

In our previous work (O'Brien and Monaghan, 1999a) we attempted to minimise the difference between the original and time-scaled frequency tracks. Such an approach, it was thought, would help to preserve the random nature of frequency tracks in voiceless regions, thus avoiding the need for phase and frequency dithering or hybrid modelling and providing a unified treatment of voiced and

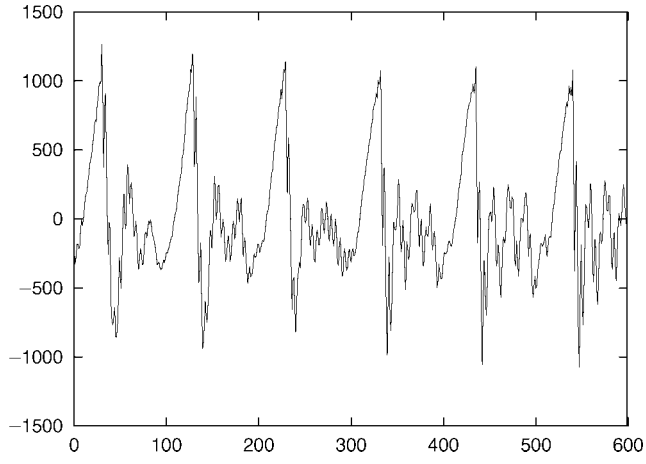


Figure 6.3 Time-scaled speech, $\rho = 0.6$

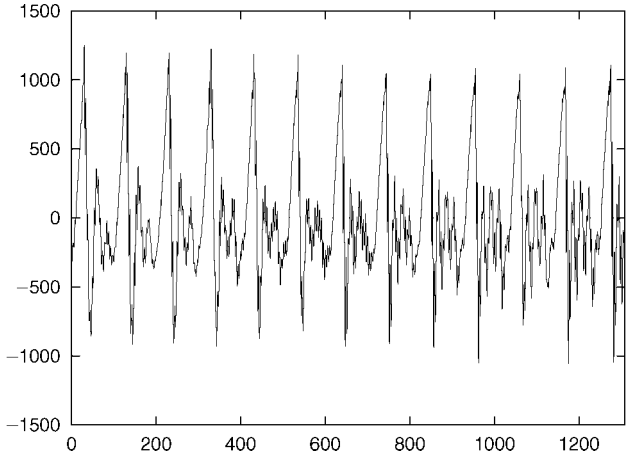


Figure 6.4 Time-scaled speech, $\rho = 1.3$

voiceless speech during time-scale modification. Using this approach, as opposed to computing the smoothest frequency track, meant slightly larger scaling factors could be accommodated before tonal artefacts were introduced. The improvement, however, was deemed insufficient to outweigh the extra computational cost incurred.

For this reason, frequency dithering techniques, to be applied over voiceless speech during time-scale expansion, were implemented. Initially, two simple methods of increasing randomness in voiceless regions were incorporated into the model:

- Upon birth or death of a sinusoid in a voiceless frame, a random start or target *phase* is assigned.
- Upon birth or death of a sinusoid in a voiceless frame, a random (but within a specified bandwidth) start or target *frequency* is assigned.

These simple procedures can be combined if necessary with shorter analysis frame intervals to handle most time-scale expansion requirements. However, for larger time-scale expansion factors, these measures may not be enough to prevent tonality. In such cases variation in frequency of 'noisy' sinusoids is increased, thereby smoothing the spectrum and helping to preserve perceptual randomness. This procedure is described in O'Brien and Monaghan (2001).

Pitch Modification

In order to perform pitch modification, it is necessary to separate vocal tract and excitation contributions to the speech production process. Here, an LPC-based inverse filtering technique, IAIF: Iterative Adaptive Inverse Filtering (Alku *et al.* 1991), is applied to the speech signal to yield a glottal excitation estimate which is sinusoidally coded. The frequency track connecting the fundamental of frame l with that of frame $l+1$ is then given by:

$$\theta_0(n) = \gamma + 2\alpha n + 3\beta n^2 \tag{7}$$

Pitch-scaling equation (7) is quite simple. Let λ^l and λ^{l+1} be the pitch modification factors associated with frames l and $l+1$ of the glottal excitation respectively. Interpolating linearly, the modification factor across the frame is given by:

$$\lambda(n) = \lambda^l + \frac{\lambda^{l+1} - \lambda^l}{S} n \tag{8}$$

where S is the analysis frame interval. The pitch-scaled fundamental can then be written as:

$$\theta_0'(n) = \theta_0(n)\lambda(n) \tag{9}$$

The new (unwrapped) target phase, $\psi_0^{l+1'}$ is found by integrating equation (9) over S and adding back the start phase, ψ_0^l:

$$\int_0^s \theta_0(n) + \psi_0^l = \frac{S}{12}\left[6\gamma\left(\lambda^l + \lambda^{l+1}\right) + 4\alpha S\left(\lambda^l + 2\lambda^{l+1}\right) + 3\beta S^2\left(\lambda^l + 3\lambda^{l+1}\right)\right] + \psi_0^l \tag{10}$$

Evaluating equation (10) modulo 2π gives $\psi_0^{l+1'}$ from which δ can be calculated and a new set of target phases derived.

Each start and target frequency is scaled by λ^l and λ^{l+1}, respectively. Composite amplitude values are calculated by multiplying excitation amplitude values by the LPC system magnitude response at each of the scaled frequencies. (Note that the excitation magnitude spectrum is not resampled but frequency-scaled.) Composite phase values are calculated by adding the new excitation phase values to the LPC system phase response measured at each scaled frequency. Re-synthesis of pitch-scaled speech may then be carried out by computing a phase interpolation function for each sinusoid and substituting into equation (11).

$$s^l(n) = \sum_k A_k^l(n)\cos\left[\theta_k^l(n)\right] \tag{11}$$

Except for the way $\psi_0^{l+1'}$ is calculated, pitch modification is quite similar to the time-scaling technique presented in Figure 6.1. The pitch-scaling algorithm is given in Figure 6.5. This approach is different from an earlier one presented by the authors (O'Brien and Monaghan, 1999b) where pitch-scaling was, in effect, converted to a time-scaling problem.

A number of speech samples were pitch modified using the method described above and the results were found to be of high quality. Some example waveforms, taken from pitch-scaled speech are given in Figures 6.6, 6.7 and 6.8. Again, it should be noted that the original waveform shape has been generally well preserved.

Joint Pitch and Time-Scale Modification

These algorithms for pitch and time-scale modification can be easily combined to perform joint modification. The frequency track linking the fundamental of frame l with that of frame $l+1$ can again be written as:

$$\theta_0(n) = \gamma + 2\alpha n + 3\beta n^2 \tag{12}$$

$\Delta = 0$
$\delta = 0$
For each Frame l
Begin
$\Delta = \Delta + \delta$
For ω_0
Begin
Adjust ψ_0^{l+1} by Δ
Compute frenquency track $\theta_0(n)$
Compute new frequency track $\theta_0(n)$
Solve for $\psi_0^{l+1'}$ and δ
Compute composite amplitude
Compute composite phase
Compute phase function $\theta_0^l(n)$
End
For ω_k where $k \neq 0$
Begin
Adjust ψ_k^{l+1} by $\delta + \Delta$
Compute composite amplitude
Compute composite phase
Compute phase function $\theta_k^l(n)$
End
End

Figure 6.5 Pitch-scaling algorithm

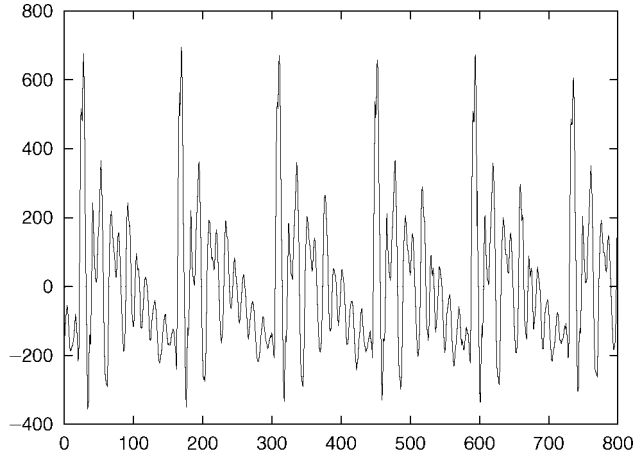


Figure 6.6 Original speech, $\lambda = 1$

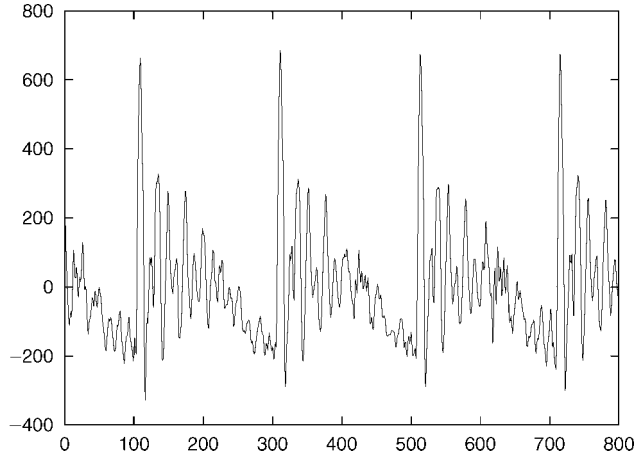


Figure 6.7 Pitch-scaled speech, $\lambda = 0.7$

The pitch and time-scaled track, where ρ is the time-scaling factor associated with frame l and λ^l and λ^{l+1} are the pitch modification factors associated with frames l and $l+1$ respectively, is given by:

$$\theta'_0(n) = \theta_0(n/\rho)\lambda(n/\rho) \tag{13}$$

where $\lambda(n)$ is the linearly interpolated pitch modification factor given in equation (8). Integrating equation (13) over the interval ρS and adding back the start phase, ψ_0^l, gives

$$\int_0^{\rho S} \theta'_0(n) + \psi_0^l = \frac{\rho S}{12}\left[6\gamma\left(\lambda^l + \lambda^{l+1}\right) + 4\alpha S\left(\lambda^l + 2\lambda^{l+1}\right) + 3\beta S^2\left(\lambda^l + 3\lambda^{l+1}\right)\right] + \psi_0^l \tag{14}$$

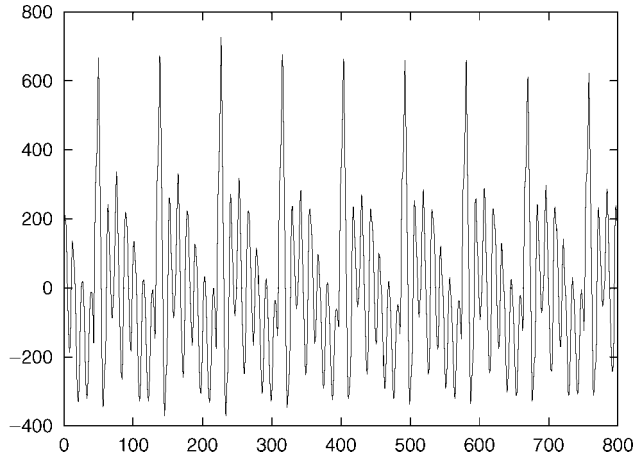


Figure 6.8 Pitch-scaled speech, $\lambda = 1.6$

Evaluating equation (14) modulo 2π gives $\psi_0^{l+1'}$ from which δ can be calculated and a new set of target phases derived. Using the scaled harmonic frequencies and new composite amplitudes and phases, synthesis is carried out to produce speech that is both pitch and time-scaled. Some example waveforms, showing speech (from Figure 6.6) which has been simultaneously pitch and time-scaled using this method, are given in Figures 6.9 and 6.10. In these examples, the same pitch and time-scaling factors have been assigned to each frame although, obviously, this need not be the case as both factors are mutually independent. As with the previous examples, waveform shape is well preserved.

Results

The time-scale and pitch modification algorithms described above were tested against other models in a prosodic transplantation task. The COST 258 coder evaluation server[2] provides a set of speech samples with neutral prosody and for each a set of associated target prosodic contours. Speech samples to be modified include vowels, fricatives (both voiced and voiceless) and continuous speech. Results from a formal evaluation (O'Brien and Monaghan, 2001) show our model's performance to compare very favourably with that of two other coders: HNM as implemented by Institut de la Communication Parlée, Grenoble, France (Bailly, Chapter 3, this volume) and a pitch-synchronous sinusoidal technique developed at the University of Vigo, Spain (Banga, Garcia-Mateo and Fernando-Salgado, Chapter 5, this volume).

Discussion

A high-quality yet conceptually simple approach to pitch and time-scale modification of speech has been presented. Taking advantage only of the harmonic

[2] *http://www.icp.grenet.fr/cost258/evaluation/server/cost 258_coders.html*

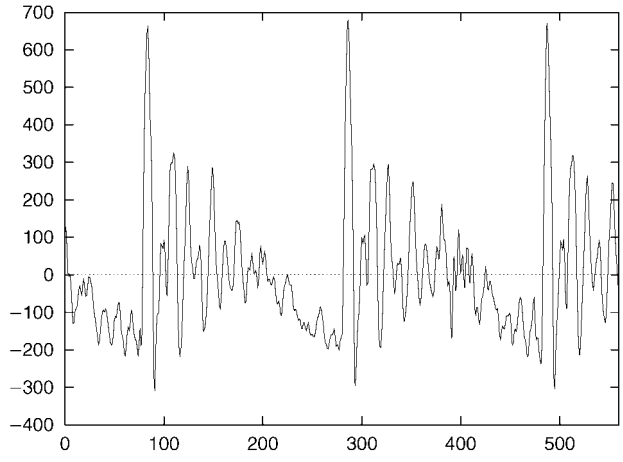


Figure 6.9 Pitch- and time-scaled speech, $\rho = 0.7, \lambda = 0.7$

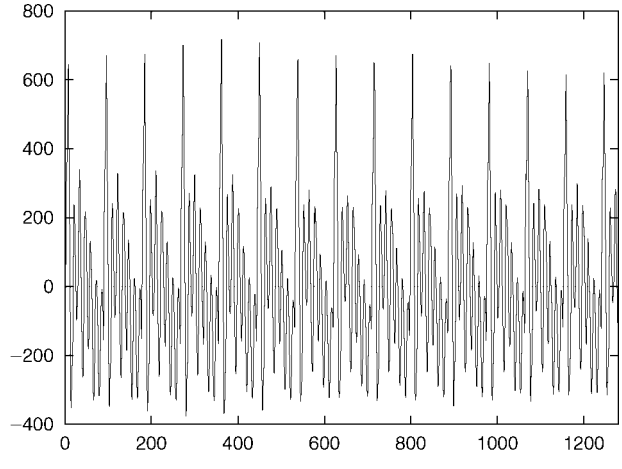


Figure 6.10 Pitch- and time-scaled speech, $\rho = 1.6, \lambda = 1.6$

structure of the sinusoids used to code each frame, phase coherence and waveform shape are well preserved after modification.

The simplicity of the approach stands in contrast to the shape invariant algorithms in Quatieri and McAulay (1992). Using their approach, pitch pulse onset times, used to preserve waveform shape, must be estimated in both the original and target speech. In the approach presented here, onset times play no role and need not be calculated. Quatieri and McAulay use onset times to impose a structure on phases and errors in their location lead to unnaturalness in the modified speech. In the approach described here, during modification phase relations inherent in the original speech are preserved. Phase coherence is thus guaranteed and waveform shape is retained. Obviously, our approach has a similar advantage over George and Smith's (1997) ABS/OLA modification techniques which also make use of pitch pulse onset times.

Unlike the PSOLA-inspired (Moulines and Charpentier, 1990) HNM approach to speech transformation (Stylianou *et al.*, 1995), using our technique no mapping need be generated from synthesis to analysis short-time signals. Furthermore, the duplication/deletion of information in the original speech (a characteristic of PSOLA techniques) is avoided: every frame is used once and only once during resynthesis.

The time-scaling technique presented here is somewhat similar to that used in George and Smith's ABS/OLA model in that the (quasi-)harmonic nature of the sinusoids used to code each frame is exploited by both models. However, the frequency (and associated phase) tracks linking one frame with the next and playing a crucial role in the sinusoidal model (McAulay and Quatieri, 1986), while absent from the ABS/OLA model, are retained here. Furthermore, our pitch modification algorithm is a direct extension of our time-scaling approach and is simpler than the 'phasor interpolation' mechanism used in the ABS/OLA model.

The incorporation of modification techniques specific to voiced and voiceless speech brings to light deficiencies in our analysis model. Voicing errors can seriously lower the quality of the re-synthesised speech. For example, where voiced speech is deemed voiceless, frequency dithering is wrongly applied, waveform dispersion occurs, and the speech is perceived as having an unnatural 'rough' quality. Correspondingly, where voiceless speech is analysed as voiced, its random nature is not preserved and the speech takes on a tonal character.

Apart from voicing errors, other problem areas also exist. Voiced fricatives, by definition, consist of a deterministic and a stochastic component and, because our model applies a binary $\pm$ voice distinction, cannot be accurately modelled. During testing, such sounds were modelled as a set of harmonics (i.e. as if purely voiced) and, while this approach coped with moderate time-scale expansion factors, a tonal artefact was introduced for larger degrees of modification.

The model could be improved and the problems outlined above alleviated by incorporating several of the elements used in HNM analysis (Stylianou *et al.*, 1995). First, leaving the rest of the model as it stands, a more refined pitch estimation procedure could be added to the analysis phase, i.e. as in HNM the pitch could be chosen to be that whose harmonics best fit the spectrum. Second, the incorporation of a voicing cut-off frequency would add the flexibility required to solve the problems mentioned in the previous paragraph. Above the cut-off point, frequency dithering techniques could be employed to ensure noise retained its random character. Below the cut-off point the speech would be modelled as a set of harmonics. The main computational burden incurred in implementing pitch and time-scale modification, using our approach, lies in keeping frequencies in phase. The use of a cut-off frequency, above which phases can be considered random, would significantly improve the efficiency of the approach as only frequencies below the cut-off point would require explicit phase monitoring. Obviously, the same idea can also be applied in purely voiceless regions to reduce processing. Finally, the inverse filtering technique currently being used (Alku *et al.*, 1991) is quite simple and is designed for efficiency rather than accuracy. A more refined algorithm should yield better quality results.

Acknowledgements

The authors gratefully acknowledge the support of the European co-operative action COST 258, without which this work would not have been possible.

References

Alku, P., Vilkman, E., and Laine, U.K. (1991). Analysis of glottal waveform in different phonation types using the new IAIF-method. Paper presented at the International Congress of Phonetic Sciences, Aix-en-Provence.

George, E.B. and Smith, M.J.T. (1997). Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add model. *IEEE Transactions on Speech and Audio Processing, 5*, 389–406.

Macon, M.W. (1996). Speech synthesis based on sinusoidal modeling. Unpublished doctoral dissertation, Georgia Institute of Technology.

McAulay, R.J. and Quatieri, T.F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *34*, 744–754.

Moulines, E. and Charpentier, F. (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.

O'Brien, D. and Monaghan, A.I.C. (1999a). Shape invariant time-scale modification of speech using a harmonic model. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (pp. 381–384). Phoenix, Arizona, USA.

O'Brien, D. and Monaghan, A.I.C. (1999b). Shape invariant pitch modification of speech using a harmonic model. *Proceedings of EUROSPEECH*, (pp. 381–384). Budapest, Hungary.

O'Brien, D. and Monaghan, A.I.C. (2001). Concatenative Synthesis based on a Harmonic Model. *IEEE Transactions on Speech and Audio Processing*, *9*, 11–20.

Quatieri, T.F. and McAulay, R.J. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, *40*, 497–510.

Stylianou, Y., Laroche, J., and Moulines, E. (1995). High quality speech modification based on a harmonic + noise model. *Proceedings of EUROSPEECH*, (pp. 451–454). Madrid, Spain.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W.B. Kleijn and K.K. Paliwal (eds), *Speech Coding and Synthesis*. Elsevier.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

7

Concatenative Speech Synthesis Using SRELP

Erhard Rank
Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology
Gusshausstrasse 25/E389
1040 Vienna, Austria
erank@nt.tuwien.ac.at

Introduction

The good quality of state-of-the-art speech synthesisers in terms of naturalness is mainly due to the use of concatenative synthesis: synthesis by concatenation of recorded speech segments usually yields more natural speech than model-based synthesis, such as articulatory synthesis or formant synthesis. Although model-based synthesis algorithms offer generally a better access to phonetic and prosodic parameters (see, for example, Ogden *et al.*, 2000), some aspects of human speech production cannot yet be fully covered, and concatenative synthesis is usually preferred by the users.

For concatenative speech synthesis, the recorded segments are commonly stored as mere time signals. In the synthesis stage too, time-domain processing with little computational effort is used for prosody manipulations, like TD-PSOLA (time-domain pitch synchronous overlap-and-add, see Moulines and Charpentier, 1990). Alternatively, no manipulations on the recorded speech are performed at all, and the selection of segments is optimised (Black and Campbell, 1995; Klabbers and Veldhuis, 1998; Beutnagel *et al.*, 1998). Both methods are reported to yield high ratings on intelligibility and naturalness when used in limited domains. TD-PSOLA can be successfully applied for general purpose synthesis with moderate prosodic manipulations and unit selection scores, if the database covers long parts of the synthesised utterances – particularly with a dedicated inventory for a certain task, like weather reports or train schedule information – but yields poor quality for example for proper names not included in the database.

Consequently, for speech synthesis applications not limited to a specific task and prosody manipulations beyond a certain threshold – not to mention attempts to change speaker characteristics (gender, age, attitude/emotion, etc.) – it is

advantageous not to be restricted by the inventory and to have flexible, and possibly phonologically interpretable synthesis and signal manipulation methods. This also makes it feasible to use inventories of reasonably small size for general purpose synthesis.

In this chapter, we describe a speech synthesis algorithm that uses a hybrid concatenative and linear predictive coding (LPC) approach with a simple method for manipulation of the prosodic parameters fundamental frequency (*f0*), segment duration, and amplitude, termed *simple residual excited linear predictive* (SRELP[1]) synthesis. This algorithm allows for large-scale modifications of fundamental frequency and duration at low computational cost in the synthesis stage. The basic concepts for SRELP synthesis are outlined, several variations of the algorithm are referenced, and the benefits and shortcomings are briefly summarised. We emphasise the benefits of using LPC in speech synthesis resulting from its relationship with the prevalent source-filter speech production model.

The SRELP synthesis algorithm is closely related to the multipulse excited LPC synthesis algorithm and to LP-PSOLA, also used for general purpose speech synthesis with prosodic manipulations, and to codebook excited linear prediction (CELP) re-synthesis without prosodic manipulations used for telephony applications.

The outline of this chapter is as follows: next we describe LPC analysis in general and the prerequisites for the SRELP synthesis algorithm, then, the synthesis procedure and the means for prosodic manipulation are outlined. The benefits of this synthesis concept compared to other methods are discussed, and also some of the problems encountered. The chapter ends with a summary and conclusion.

Preprocessing procedure

The idea of LPC analysis is to decompose a speech signal into a set of coefficients for a linear prediction filter (the 'inverse' filter) and a residual signal. The inverse filter shall compensate the influence of the vocal tract on the glottis pressure signal (Markel and Gray, 1976). This mimicking of the source-filter speech production model (Fant, 1970) allows for separate manipulations on the residual signal (related to the glottis pressure pulses) and on the LPC filter (vocal tract transfer function), and thus provides a way to independently alter the glottis-signal related parameters *f0*, duration, and amplitude via the residual and spectral envelope (e.g., formants) in the LPC filter. For SRELP synthesis, the recorded speech signal is pitch-synchronously LPC-analysed, and both the coefficients for the LPC filter and the residual signal are used in the synthesis stage. For best synthesis speed the LPC analysis is performed off-line, and the LPC filter coefficients and the residual signal are stored in an inventory employed for synthesis.

To perform SRELP synthesis, the analysis frame boundaries of voiced parts of the speech signal are placed in a way such that the estimated glottis closure instant is aligned in the center of a frame,[2] and LPC analysis is performed by processing

[1] We distinguish here between the terms RELP (residual excited linear predictive) synthesis for perfect reconstruction of a speech signal by excitation of the LPC filter with the residual, and SRELP for resynthesis of a speech signal with modified prosody.

the recorded speech by a finite-duration impulse response (FIR) filter with transfer function $A(z)$ (the inverse filter) to generate a residual signal S_{res} with a peak of energy of the residual also in the center of the frame. The filter transfer function $A(z)$ is obtained from LPC analysis based on the auto-correlation function or the covariance of the recorded speech signal, or by performing partial correlation analysis using a ladder filter structure (Makhoul, 1975; Markel and Gray, 1976).

Thus, for a correct choice of LPC analysis frames, the residual energy typically decays towards the frame borders for voiced frames, as in Figure 7.1a. For unvoiced frames the residual is noise-like with its energy evenly distributed over time and a fixed frame length is used, as indicated in Figure 7.1b.

For re-synthesis an all pole LPC filter with a transfer function $V(z) = 1/A(z)$ is used. This re-synthesis filter can be implemented in different ways: the straightforward implementation is a pure recursive infinite-duration impulse response (IIR) filter. Also, there are different kinds of lattice structures that implement the transfer function $V(z)$ (Markel and Gray, 1976). Note that due to the time-varying nature of speech the filter coefficients have to be re-adjusted regularly and thus switching transients will occur when the filter coefficients are changed. One thing to pay attention to is that the re-synthesis filter structure matches the analysis (inverse) filter structure or adequate adaptations of the filter state have to be performed when the coefficients are changed.

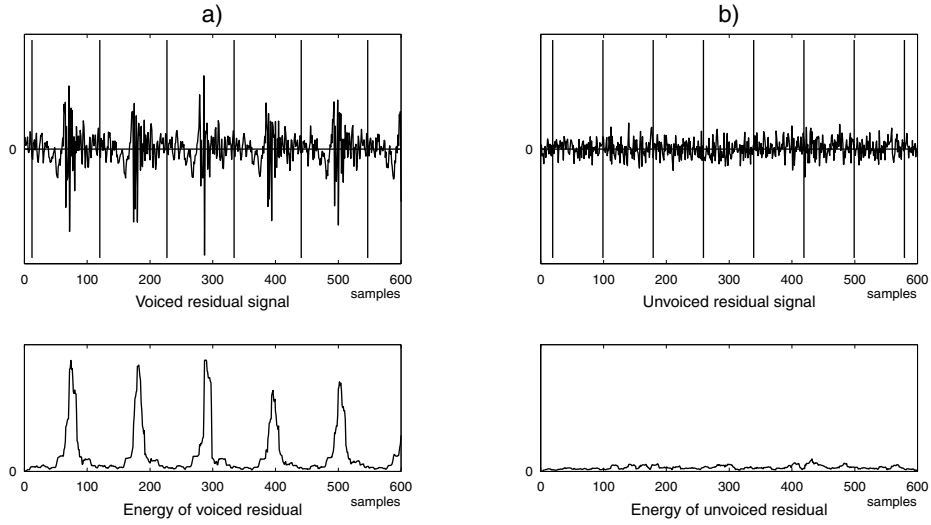


Figure 7.1 Residual signals and local energy estimate (estimated by twelve point moving average filtering of the sample power) (a) for a voiced phoneme (vowel /a/) and (b) for an unvoiced phoneme (/s/). The borders of the LPC analysis frames are indicated by vertical lines in the signal plots. In the unvoiced case frame borders are at fixed regular intervals of 80 samples. Note that in the voiced case the pitch-synchronous frames are placed such that the energy peaks of the residual corresponding to the glottis closure instant are centred within each frame.

[2] Estimation of the glottis closure instants (pitch extraction) is a complex task of its own (Hess, 1983), which is not further discussed here.

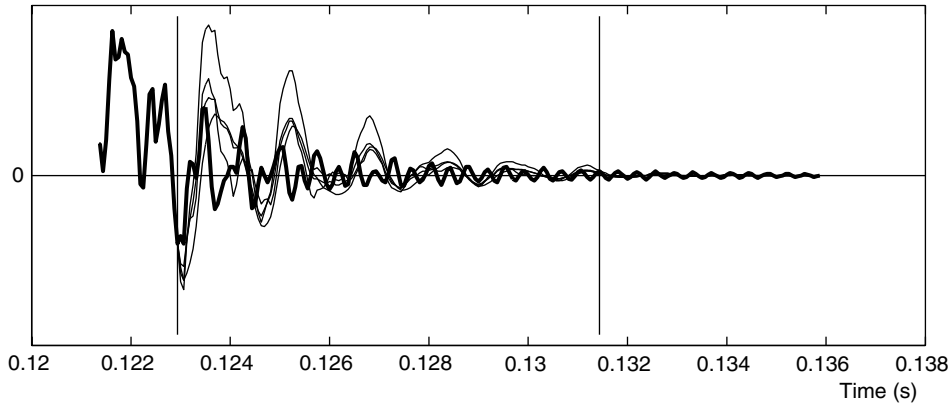


Figure 7.2 Transient behaviour caused by coefficient switching for different LPC filter realisations. The thick line shows the filter output with the residual zeroed at a frame border (vertical lines) and the filter coefficients kept constant. The signals plotted in thin lines expose transients evoked by switching the filter coefficients at frame borders for different filter structures. A statistic of the error due to the transients is given in Table 7.1.

On the other hand, the amplitude of the transients depends on the filter structure in general, as has been investigated in Rank (2000). To classify the error caused by switching filter coefficients the input for LPC synthesis filter (the residual signal) was set to zero at a frame border and the decay of the output speech signal was observed with and without switching the coefficients. An example of transients in the decaying output signal evoked by coefficient switching for different filter structures is shown in Figure 7.2. The signal plotted as thick line is without coefficient switching – and is the same for all different filter structures – whereas the signals in thin lines are evoked by switching coefficients of the direct form 2 IIR filter and several lattice filter types. A quantitative evaluation over the signals in the Cost 258 Signal Generation Test Array (see Bailly, Chapter 4, this volume) is depicted in Table 7.1. The maximum suppression of transients of −6.07 dB was achieved using the normalised lattice filter structure and correction of interaction between frames during LPC analysis (Ferencz *et al.*, 1999).

The implementation of the re-synthesis filter as a lattice filter can be interpreted as a discrete-time model for wave propagation in a one-dimensional waveguide

Table 7.1 Average error due to transients caused by filter coefficient switching for different LPC synthesis filter structures (2-multiplier, normalized, Kelly-Lochbaum (KL), and 1-multiplier lattice structure and direct form structure).

2-multiplier	Normalized	KL/1-multiplier	Direct form
−4.249 dB	−4.537 dB	−4.102 dB	−4.980 dB
−3.608 dB	−6.073 dB	−4.360 dB	−4.292 dB

Note: The values are computed as relative energy of the error signal in relation to the energy of the decaying signal without coefficient switching. The upper row is for simple LPC analysis over one frame, the lower row for LPC analysis over one frame with correction of the influence from the previous frame, where best suppression is achieved with the normalized lattice filter structure.

with varying wave impedance. The order of the LPC re-synthesis filter relates to the length of the human vocal tract equidistantly sampled with a spatial sampling distance corresponding to the sampling frequency of the recorded speech signal (Markel and Gray, 1976). The implementation of the LPC filter as a lattice filter is directly related to the lossless acoustic tube model of the vocal tract and has subtle advantages over the transversal filter structure, for example the prerequisites for easy and robust filter interpolation (see Rank, 1999 and p. 82).

Several possible improvements of the LPC analysis process should be mentioned here, such as analysis within the closed glottis interval only. When the glottis is closed, ideally the vocal tract is decoupled from the subglottal regions and no excitation is present. Thus, the speech signal in this interval will consist of freely decaying oscillations that are governed by the vocal tract transfer function only. An LPC filter obtained by closed-glottis analysis typically has larger bandwidths for the formant frequencies, compared to a filter obtained from LPC analysis over a contiguous interval (Wong *et al.*, 1979).

An inverse filtering algorithm especially designed for robust pitch modification in synthesis called low-sensitivity inverse filtering (LSIF) is described by Ansari, Kahn, and Macchi (1998). Here the bias of the LPC spectrum towards the pitch harmonics is overcome by a modification of the covariance matrix used for analysis by means of adding a symmetric Toeplitz matrix. This approach is also reported to be less sensitive to errors in pitch marking than pure SRELP synthesis.

Another interesting possibility is the LPC analysis with compensation for influences on the following frames (Ferencz *et al.*, 1999), as used in the analysis of transient behaviour described above. Here the damped oscillations generated during synthesis with the estimated LPC filter that may overlap with the next frames are subtracted from the original speech signal before analysis of these frames. This method may be especially useful for female voices, where the pitch period is shorter than for male voices, and the LPC filter has a longer impulse response in comparison to the pitch period.

Synthesis and prosodic manipulations

The SRELP synthesis model as described involves an LPC filter that directly models vocal tract properties and a residual signal resembling glottis pulses to some extent. The process for manipulating fundamental frequency and duration is now outlined in detail.

As described, the pitch-synchronous frame boundaries are set in such a way that the peak in the residual occurring at the glottis closure instant is centred within a frame. For each voiced frame, the residual vector $\mathbf{x}_{\text{res}}$ contains a high energy pulse in the center and typically decays towards the beginning and the end. To achieve a certain pitch, the residual vector $\mathbf{x}_{\text{res}}$ of a voiced frame is set to a length n_{res} according to the desired fundamental frequency *f0*. If this length is longer than the original frames residual length, the residual is zero padded at both ends. If it is shorter, the residual is truncated at both ends. The modified residual vectors are then concatenated to form the residual signal S_{res} which is used to excite the LPC synthesis filter with coefficients according to the residual frames. This is

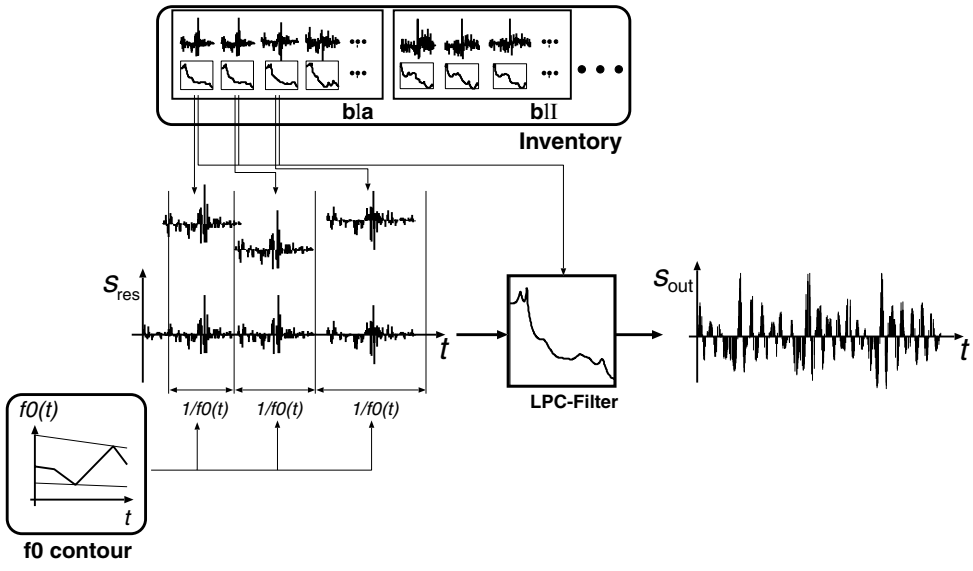


Figure 7.3 Schematic of SRELP re-synthesis. To achieve a given fundamental frequency contour $f0(t)$ at each point in time the pitch period is computed and used as length for the current frame. If the length of the original frame in the inventory is longer than the computed length, the residual of this frame is cut off at both ends to fit in the current frame (first frame of S_{res}). If the original frames length is shorter than the computed length the residual is zero padded at both ends (third frame of S_{res}). This results in a train of residual pulses S_{res} with the intended fundamental frequency. This residual signal is then fed through a LPC re-synthesis filter with the coefficients from the inventory corresponding to the residual frame to generate the speech output signal S_{out}.

illustrated in Figure 7.3 for a series of voiced frames. Thus, signal manipulations are restricted to the low energy part (the tails) of each frame residual. For unvoiced frames no manipulations on frame length are performed.

Duration modifications are achieved by repeating or dropping residual frames. Thus, segments of the synthesised speech can be uniformly stretched, or nonlinear time warping can be applied. A detailed description of the lengthening strategies used in a SRELP demisyllable synthesiser is given in Rank and Pirker (1998b). In our current synthesis implementation the original frames LPC filter coefficients are used during the stretching which is satisfactory when no large dilatation is performed.

The SRELP synthesis procedure as such is similar to the LP-PSOLA algorithm (Moulines and Charpentier, 1990) concerning the pitch synchronous LPC analysis, but no windowing and overlap-and-add process is performed.

Discussion

One obvious benefit of the SRELP algorithm is the simplicity of the prosody manipulations in the re-synthesis stage. This simplicity is of course tied to a higher complexity in the analysis stage – pitch prediction and LPC analysis – which is not

necessary for some other synthesis methods. But this simplicity results in fewer artifacts due to signal processing (like windowing). Better quality of synthetic speech than with other algorithms is achieved in particular for fundamental frequency changes of considerable size, especially for male voices transformed from the normal (125 Hz) to the low pitch (75 Hz) range.

Generally, the decomposition of the speech signal into vocal tract (LPC) filter and excitation signal (residual) allows for independent manipulations of parameters concerning residual (*f0*, duration, amplitude) and vocal tract properties (formants, spectral tilt, articulatory precision, etc.). This parametrisation promotes smoothing (parameter interpolation) independent for each parameter regime at concatenation points (Chappel and Hanson, 1998; Rank, 1999), but it can also be utilised for voice quality manipulations that can be useful for synthesis of emotional speech (Rank and Pirker, 1998c).

The capability of parameter smoothing at concatenation points is illustrated in Figure 7.4. The signals and spectograms each show part of a synthetic word concatenated from the first part of *dediete* and the second part of *tetiete* with the concatenation taking place in the vowel /i/. At the concatenation point, a mismatch of spectral envelope and fundamental frequency is encountered. This mismatch is clearly visible in the plots for hard concatenation of the time signals (case a). Concatenation artifacts can be even worse if the concatenation point is not related to the pitch cycles, as it is here. Hard concatenation in the LPC residual domain (case b) with no further processing already provides some smoothing by the LPC synthesis filter. With interpolation of the LPC filter (case c), the mismatch in spectral content can be smoothed and with interpolation of fundamental frequency using SRELP (case d), the mismatch of the spectral fine structure is removed also.

Interpolation of the LPC filter is performed in the log area ratio (LAR) domain, which corresponds to smoothing the transitions of the cross sections of an acoustic tube model for the vocal tract. Interpolation of LARs or direct interpolation of lattice filter coefficients also always provides stable filter behaviour. Fundamental frequency is interpolated on a logarithmic scale, i.e., in the tone domain.

SRELP synthesis has been compared to other synthesis techniques regarding prosody manipulation by Macchi *et al.* (1993). The possibility of using residual vectors and LPC filter coefficients from different frames has been investigated by Kezniki (1995). An approach using a phoneme-specific residual prototype library, including different pitch period lengths, is described by Fries (1994). The implementation of a demisyllable synthesiser for Austrian German using SRELP is described in Rank and Pirker (1998a, b), and can be tested over the worldwide web (http://www.ai.univie.ac.at/oefai/nlu/viectos). The application of the synthesis algorithm in the Festival speech synthesis system with American English and Mexican Spanish voices is described in Macon *et al.* (1997). Similar synthesis algorithms are described in Pearson *et al.* (1998) and in Ferencz *et al.* (1999). Also, this synthesis algorithm is one of several algorithms tested within the Cost 258 Signal Generation Test Array (Bailly, Chapter 4, this volume).

A problem mentioned already is the need for a good estimation of glottis closure instants. This often requires manual corrections which is a very time-consuming

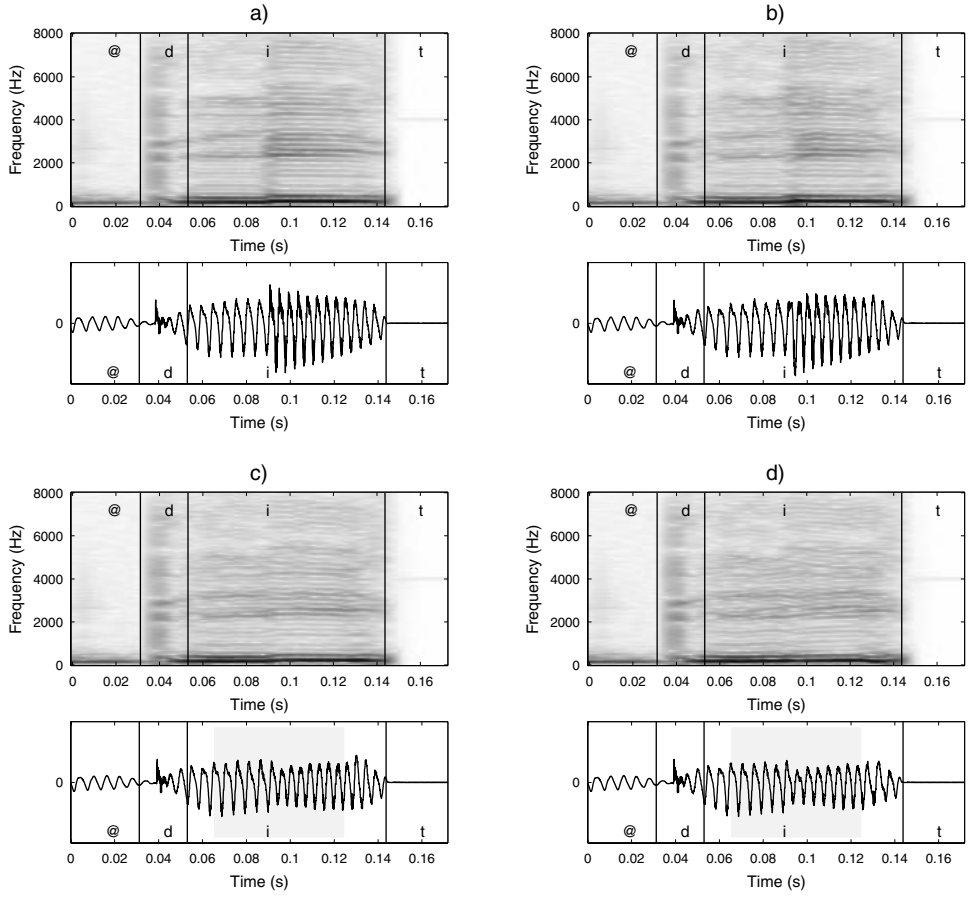


Figure 7.4 Smoothing of concatenation discontinuities. Concatenation of two half word segments with different segmental context inside the scope of the vowel /i/ using (a) concatenation in the speech signal domain at a pitch frame border; (b) concatenation of the LPC residual at a frame border without any other manipulations; (c) concatenation of the LPC residual at a frame border and interpolation of the LPC filter over the shaded region; and (d) concatenation of the LPC residual at a frame border and interpolation of the LPC filter and the fundamental frequency over the shaded region.

part of the analysis process. Another problem is the application of the synthesis algorithm on mixed excitation speech signals. For voiced fricatives the best solution seems to be pitch-synchronous frame segmentation, but no application of fundamental frequency modification. So for length modifications, the phase relations of the voiced signal part are preserved.

It is also notable that SRELP synthesis with modification of fundamental frequency yields more natural-sounding speech for low-pitch voices, compared to high pitch voices. Due to the shorter pitch period for a high-pitch voice, the impulse response of the vocal tract filter is longer in relation to the frame length, and there is a considerable influence on the following frame(s) (but see the remarks on p. 80).

Conclusion

SRELP speech synthesis provides the means for prosody manipulations at low computational cost in the synthesis stage. Due to the restriction of signal manipulations in the low energy part of the residual, signal processing artifacts are low, and good-quality synthetic speech is generated, in particular when performing large-scale modifications in fundamental frequency to the low pitch register. It also provides us with the means for parameter smoothing at concatenation points and for manipulations of the vocal tract filter characteristics. SRELP can be used for prosody manipulation in speech synthesisers with a fixed (e.g., diphone) inventory, or for prosody manipulation and smoothing in unit selection synthesis, when appropriate information (glottis closure instants, phoneme segmentation) is present in the database.

Acknowledgements

This work was carried out with the support of the European Cost 258 'The Naturalness of Synthetic Speech', including a fruitful short-term scientific mission to ICP, Grenoble. Many thanks go to Esther Klabbers, IPO, Eindhoven, for making available the signals for the concatenation task. Part of this work has been performed at the Austrian Research Institute for Artificial Intelligence (ÖFAI), Vienna, Austria, with financial support from the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (grant no. FWF P10822) and by the Austrian Federal Ministry of Science and Transport.

References

Ansari, R., Kahn, D., and Macchi, M.J. (1998). Pitch modification of speech using a low sensitivity inverse filter approach. *IEEE Signal Processing Letters, 5(3)*, 60–62.

Beutnagel, M., Conkie, A., and Syrdal, A.K. (1998). Diphone synthesis using unit selection. *Proc. of the Third ESCA/COCOSDA Workshop on Speech Synthesis* (pp. 185–190). Jenolan Caves, Blue Mountains, Australia.

Black, A.W. and Campbell, N. (1995). Optimising selection of units from speech databases for concatenative synthesis. *Proc. of Eurospeech '95*, Vol. 2 (pp. 581–584). Madrid, Spain.

Chappel, D.T. and Hanson, J.H.L. (1998). Spectral smoothing for concatenative synthesis. *Proc. of the 5th International Conference on Spoken Language Processing*, Vol. 5 (pp. 1935–1938). Sydney, Australia.

Fant, G. (1970). *Acoustic Theory of Speech Production*. Mouton.

Ferencz, A., Nagy, I., Kovács, T.-C., Ratiu, T., and Ferencz, M. (1999). On a hybrid time domain-LPC technique for prosody superimposing used for speech synthesis. *Proc. of Eurospeech '99*, Vol. 4. (pp. 1831–1834), Budapest, Hungary.

Fries, G. (1994). Hybrid time- and frequency-domain speech synthesis with extended glottal source generation. *Proc. of ICASSP '94*, Vol. 1 (pp. 581–584). Adelaide, Australia.

Hess, W. (1983). *Pitch Determination of Speech Signals: Algorithms and Devices*. Springer-Verlag.

Keznikl, T. (1995). Modifikation von Sprachsignalen für die Sprachsynthese (Modification of speech signals for speech synthesis, in German). *Fortschritte der Akustik, DAGA '95*, Vol. 2 (pp. 983–986). Saarbrücken, Germany.

Klabbers E. and Veldhuis, R. (1998). On the reduction of concatenation artefacts in diphone synthesis. *Proc. of the 5th International Conference on Spoken Language Processing*, Vol. 5 (pp. 1983–1986). Sydney, Australia.

Macchi, M., Altom, M.J., Kahn, D., Singhal, S., and Spiegel, M. (1993). Intelligibility as a function of speech coding method for template-based speech synthesis. *Proc. of Eurospeech '93* (pp. 893–896). Berlin, Germany.

Macon, M., Cronk, A., Wouters, J., and Klein, A. (1997). OGIresLPC: Diphone synthesizer using residual-excited linear prediction. *Tech. Rep. CSE-97-007*. Department of Computer Science, Oregon Graduate Institute of Science and Technology, Portland, OR.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proc. of the IEEE, 63(4)*, 561–580.

Markel, J.D. and Gray, A.H. Jr. (1976). *Linear Prediction of Speech*. Springer Verlag.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*, 453–467.

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovicová, J., and Heid, S. (2000). ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Computer Speech and Language, 14*, 177–210.

Pearson, S., Kibre, N., and Niedzielski, N. (1998). A synthesis method based on concatenation of demisyllables and a residual excited vocal tract model. *Proc. of the 5th International Conference on Spoken Language Processing*, Vol. 6 (pp. 2739–2742). Sydney, Australia.

Rank, E. (1999). Exploiting improved parameter smoothing within a hybrid concatenative/LPC speech synthesizer. *Proc. of Eurospeech '99*, Vol. 5 (pp. 2339–2342). Budapest, Hungary.

Rank, E. (2000). Über die Relevanz von alternativen LP-Methoden für die Sprachsynthese (On the relevance of alternative LP methods for speech synthesis, in German). *Fortschritte der Akustik, DAGA '2000*, Oldenburg, Germany.

Rank, E., and Pirker, H. (1998a). VieCtoS – speech synthesizer, technical overview. *Tech. Rep. TR–98–13*. Austrian Research Institute for Artificial Intelligence, Vienna, Austria.

Rank, E. and Pirker, H. (1998b). Realization of prosody in a speech synthesizer for German. *Computer Studies in Language and Speech*, *Vol. 1: Computers, Linguistics, and Phonetics between Language and Speech* (*Proc. of Konvens '98, Bonn, Germany.*), 169–178.

Rank, E. and Pirker, H. (1998c). Generating emotional speech with a concatenative synthesizer. *Proc. of the 5th International Conference on Spoken Language Processing*, Vol. 3 (pp. 671–674). Sydney, Australia.

Wong, D.Y., Markel, J.D., and Gray, A.H. Jr. (1979). Least squares glottal inverse filtering from the acoustic speech wave form. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27(4)*, 350–355.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Part II

Issues in Prosody

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

8

Prosody in Synthetic Speech
Problems, Solutions and Challenges

Alex Monaghan
Aculab Plc, Lakeside, Bramley Road
Mount Farm, Milton Keynes MK1 1PT, UK
Alex.Monaghan@aculab.com

Introduction

When the COST 258 research action began, prosody was identified as a crucial area for improving the naturalness of synthetic speech. Although the segmental quality of speech synthesis has been greatly improved by the recent development of concatenative techniques (see the section on signal generation in this volume, or Dutoit (1997) for an overview), these techniques will not work for prosody. First, there is no agreed set of prosodic elements for any language: the type and number of intonation contours, the set of possible rhythmic patterns, the permitted variation in duration and intensity for each segment, and the range of natural changes in voice quality and spectral shape, are all unknown. Second, even if we only consider the partial sets which have been proposed for some of these aspects of prosody, a database which included all possible combinations would be unmanageably large. Improvements in prosody in the foreseeable future are therefore likely to come from a more theoretical approach, or from empirical studies concentrating on a particular aspect of prosody.

In speech synthesis systems, prosody is usually understood to mean the specification of segmental durations, the generation of fundamental frequency (F_0), and perhaps the control of intensity. Here, we are using the term prosody to refer to all aspects of speech which are not predictable from the segmental transcription and the speaker characteristics: this includes short-term voice quality settings, phonetic reduction, pitch range, emotional and attitudinal effects. Longer-term voice quality settings and speech rate are discussed in the contributions to the section on speaking styles.

Problems and Solutions

Prosody is important for speech synthesis because it conveys aspects of meaning and structure which are not implicit in the segmental content of utterances. It

conveys the difference between new or important information and old or unimportant information. It indicates whether an utterance is a question or a statement, and how it is related to previous utterances. It expresses the speaker's beliefs about the content of the utterance. It even marks the boundaries and relations between several concepts in a single utterance. If a speech synthesiser assigns the wrong prosody, it can obscure the meaning of an utterance or even convey an entirely different meaning.

Prosody is difficult to predict in speech synthesis systems because the input to these systems contains little or no explicit information about meaning and structure, and such information is extremely hard to deduce automatically. Even when that information is available, in the form of punctuation or special mark-up tags, or through syntactic and semantic analysis, its realisation as appropriate prosody is still a major challenge: the complex interactions between different aspects of prosody (F_0, duration, reduction, etc.) are often poorly understood, and the translation of linguistic categories such as 'focus' or 'rhythmically strong' into precise acoustic parameters is influenced by a large number of perceptual and contextual factors.

Four aspects of prosody were identified for particular emphasis in COST 258:

- prosodic effects of focus and/or emphasis
- prosodic effects of speaking styles
- rhythm: what is rhythm, and how can it be synthesised?
- mark-up: what prosodic markers are needed at a linguistic (phonological) level?

These aspects are all very broad and complex, and will not be solved in the short term. Nevertheless, COST 258 has produced important new data and ideas which have advanced our understanding of prosody for speech synthesis. There has been considerable progress in the areas of speaking styles and mark-up during COST 258, and they have each produced a separate section of this volume. Rhythm is highly relevant to both styles of speech and general prosody, and several contributions address the problem of rhythmicality in synthetic speech.

The issue of focus or emphasis is of great interest to developers of speech synthesis systems, especially in emerging applications such as spoken information retrieval and dialogue systems (Breen, Chapter 37, this volume). Considerable attention was devoted to this issue during COST 258, but the resources needed to make significant progress in this pan-disciplinary area were not available. Some discussion of focus and emphasis is presented in the sections on mark-up and future challenges (Monaghan, Chapter 31, this volume; Caelen-Haumont, Chapter 36, this volume).

Contributions to this section range from acoustic studies providing basic data on prosodic phenomena, through applications of such data in the improvement of speech synthesisers, to new theories of the nature and organisation of prosodic phenomena with direct relevance to synthetic speech. This diversity reflects the current language-dependent state of prosodic processing in speech synthesis systems. For some languages (e.g. English, Dutch and Swedish) the control of several prosodic parameters has been refined over many years and recent improvements have come from the resolution of theoretical details. For most economically powerful European languages (e.g. French, German, Spanish and Italian) the necessary acoustic and phonetic data have only been available quite recently and their

implementation in speech synthesisers is relatively new. For the majority of European languages, and particularly those which have not been official languages of the European Union, basic phonetic research is still lacking: moreover, until the late 1990s researchers working on these languages generally did not consider the possibility of applying their results to speech synthesis. The work presented here goes some way towards evening out the level of prosodic knowledge across languages: considerable advances have been made in some less commonly synthesised languages (e.g. Czech, Portuguese and Slovene), often through the sharing of ideas and resources from more established synthesis teams, and there has also been a shift towards multilingual research whose results are applicable to a large number of languages.

The contributions by Teixeira and Freitas on Portuguese, Dobnikar on Slovene and Dohalská on Czech all advance the prosodic quality of synthetic speech in relatively neglected languages. The methodologies used by these researchers are all applicable to the majority of European languages, and it is to be hoped that they will encourage other neglected linguistic communities to engage in similar work. The results presented by Fackrell and his colleagues are explicitly multilingual, and although their work to date has concentrated on more commercially prominent languages, it would be equally applicable to, say, Turkish or Icelandic.

It is particularly pleasing to present contributions dealing with four aspects of the acoustic realisation of prosody (pitch, duration, intensity and vowel quality) rather than the more usual two. Very few previous publications have discussed variations of intensity and vowel quality in relation to synthetic speech, and the fact that this part includes three contributions on these aspects is an indication that synthesis technology is ready to use these extra dimensions of prosodic control. The initial results for intensity presented by Dohalská and Teixeira and Freitas, for Czech and Portuguese respectively, may well apply to several related languages and should stimulate research for other language families. The contribution by Widera, on perceived levels of vowel reduction, is based solely on German data but will obviously bear repetition for other Germanic and non-Germanic languages where vowel quality is an important correlate of prosodic prominence. The underlying approach of expressing prosodic structure as a sequence of prominence values is an interesting new development in synthesis research, and the consequent link between prosodic realisations and perceptual categories is an important one which is often neglected in current theory-driven and data-driven approaches alike (see 't Hart *et al*. (1990) for a full discussion).

As well as contributions dealing with F_0 and duration in isolation, this part presents two attempts to integrate these aspects of prosody in a unified approach. The model proposed by Mixdorff is based on the Fujisaki model of F_0 (Fujisaki and Hirose, 1984) in which pitch excursions have consequences for duration. The contribution by Zellner Keller and Keller concentrates on the rhythmic organisation of speech, which is seen as underlying the natural variations in F_0, duration and other aspects of prosody. This contribution is at a more theoretical level, as is Martin's analysis of F_0 in Romance languages, but both are aimed at improving the naturalness of current speech synthesis systems and provide excellent examples of best practice in the application of linguistic theory to speech technology.

Looking ahead

This part presents new methodologies for research into synthetic prosody, new aspects of prosody to be integrated into speech synthesisers, and new languages for synthesis applications. The implementation of these ideas and results for a large number of languages is an important step in the maturation of synthetic prosody, and should stimulate future research in this area.

Several difficult questions remain to be answered before synthetic prosody can rival its natural counterpart, including how to predict prosodic prominence (see Monaghan, 1993) and how to synthesise rhythm and other aspects of prosodic structure. Despite this, the goal of natural-sounding multilingual speech synthesis is becoming more realistic. It is also likely that better control of intensity, rhythm and vowel quality will lead to improvements in the segmental quality of synthetic speech.

References

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer.

Fujisaki, H. and Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)*, *5*, 233–241.

't Hart, J., Collier, R., and Cohen, A. (1990). *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press.

Monaghan, A.I.C. (1993). What determines accentuation? *Journal of Pragmatics*, *19*, 559–584.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

9

State-of-the-Art Summary of European Synthetic Prosody R&D

Alex Monaghan
Aculab Plc, Lakeside, Bramley Road
Mount Farm, Milton Keynes MK1 1PT, UK
Alex.Monaghan@aculab.com

Introduction

This chapter summarises contributions from approximately twenty different research groups across Europe. The motivations, methods and manpower of these groups vary greatly, and it is thus difficult to represent all their work satisfactorily in a concise summary. I have therefore concentrated on points of consensus, and I have also attempted to include the major exceptions to any consensus. I have not provided references to all the work mentioned in this chapter, as this would have doubled its length: a list of links to websites of individual research groups is provided on the Webpage, as well as an incomplete bibliography (sorted by country) for those requiring more information. Similar information is available online.[1] For a more historical perspective on synthetic prosody, see Monaghan (1991).

While every attempt has been made to represent the research and approaches of each group accurately, there may still be omissions and errors. It should therefore be pointed out that any such errors or omissions are the responsibility of this author, and that in general this chapter reflects the opinions of the author alone. I am indebted to all my European colleagues who provided summaries of their own work, and I have deliberately stuck very closely to the text of those summaries in many cases. Unattributed quotations indicate personal communication from the respective institutions. In the interests of brevity, I have referred to many institutions by their accepted abbreviations (e.g. IPO for Instituut voor Perceptie

[1] http://www.compapp.dcu.ie/~alex/cost258.html and http://www.unil.ch/imm/docs/LAIP/COST_258/cost258.htm

Onderzoek): a list of these abbreviations and the full names of the institutions are given at the end of this chapter, and may be cross-referenced with the material on the Webpage and on the COST 258 website.[2]

Overview

In contrast to US or Japanese work on synthetic prosody, European research has no standard approach or theory. In fact, there are generally more European schools of thought on modelling prosody than there are European languages whose prosody has been modelled. We have representatives of the linguistic, psycho-acoustic and stochastic approaches, and within each of these approaches we have phoneticians, phonologists, syntacticians, pragmaticists, mathematicians and engineers. Nevertheless, certain trends and commonalities emerge.

First, the modelling of fundamental frequency is still the goal of the majority of prosody research. Duration is gaining recognition as a major problem for synthetic speech, but intensity continues to attract very little attention in synthesis research. Most workers acknowledge the importance of interactions between these three aspects of prosody, but as yet very few have devoted significant effort to investigating such interactions.

Second, synthesis methodologies show a strong tendency towards stochastic approaches. Many countries which have not previously been at the forefront of international speech synthesis research have recently produced speech databases and are attempting to develop synthesis systems from these. Methodological details vary from neural nets trained on automatically aligned data to rule-based classifiers derived from hand-labelled corpora. In addition, these stochastic approaches tend to concentrate on the acoustic phonetic level of prosodic description, examining phenomena such as average duration and F0 by phoneme or syllable type, lengths of pause between different lexical classes, classes of pause between sentences of different lengths, and constancy of prosodic characteristics within and across speakers. These are all phenomena which can be measured without any labelling other than phonemic transcription and part-of-speech tagging.

Ironically, there is also widespread acknowledgement that structural and functional categories are the major determinants of prosody, and that therefore synthetic prosody requires detailed knowledge of syntax, semantics, pragmatics, and even emotional factors. None of these are easily labelled in spoken corpora, and therefore tend to be ignored in practice by stochastic research. Compared with US research, European work seems generally to avoid the more abstract levels of prosody, although there are of course exceptions, some of which are mentioned below.

The applications of European work on synthetic prosody range from R&D tools (classifiers, phoneme-to-speech systems, mark-up languages), through simple TTS systems and limited-domain concept-to-speech (CSS) applications, to fully-fledged unrestricted text input and multimedia output systems, information retrieval (IR) front ends, and talking document browsers. For some European languages, even

[2] http://www.unil.ch/imm/docs/LAIP/COST_258/cost258.htm

simple applications have not yet been fully developed: for others, the challenge is to improve or extend existing technology to include new modalities, more complex input, and more intelligent or natural-sounding output.

The major questions which must be answered before we can expect to make progress in most cases seem to me to be:

- What is the information that synthetic prosody should convey?
- What are the phonetic correlates that will convey it?

For the less ambitious applications, such as tools and restricted text input systems, it is important to ascertain which levels of analysis should be performed and what prosodic labels can reliably be generated. The objective is often to avoid assigning the wrong label, rather than to try and assign the right one: if in doubt, make sure the prosody is neutral and leave the user to decide on an interpretation. For the more advanced applications, such as 'intelligent' interfaces and rich-text processors, the problem is often to decide which aspects of the available information should be conveyed by prosodic means, and how the phonetic correlates chosen to convey those aspects are related to the characteristics of the document or discourse as a whole: for example, when faced with an input text which contains italics, bold, underlining, capitalisation, and various levels of sectioning, what are the hierarchic relations between these different formattings and can they all be encoded in the prosody of a spoken version?

Pitch, Timing and Intensity

As stated above, the majority of European work on prosody has concentrated on pitch, with timing a close second and intensity a poor third. Other aspects of prosody, such as voice quality and spectral tilt, have been almost completely ignored for synthesis purposes.

All the institutions currently involved in COST 258 who expressed an interest in prosody have an interest in the synthesis of pitch contours. Only two have concentrated entirely on pitch. All others report results or work in progress on pitch and timing. Only three institutions make significant reference to intensity.

Pitch

Research on pitch (fundamental frequency or abstract intonation contours) is mainly at a very concrete level. The 'J. Stefan' Institute in Slovenia (see Dobnikar, Chapter 14, this volume) is a typical case, concentrating on 'the microprosody parameters for synthesis purposes, especially ... modelling of the intra-word F_0 contour'. Several other institutions take a similar stochastic corpus-based approach. The next level of abstraction is to split the pitch contour into local and global components: here, the Fujisaki model is the commonest approach (see Mixdorff, Chapter 13, this volume), although there is a home-grown alternative (MOMEL and INTSINT: see Hirst, Chapter 32, this volume) developed at Aix-en-Provence.

Work at IKP is an interesting exception, having recently moved from the Fujisaki model to a 'Maximum Based Description' model. This model uses temporal alignment of pitch maxima and scaling of those maxima within a speaker-specific pitch range, together with sinusoidal modelling of accompanying rises and falls, to produce a smooth contour whose minima are not directly specified. The approach is similar to the Edinburgh model developed by Ladd, Monaghan and Taylor for the phonetic description of synthetic pitch contours.

Workers at KTH, Telenor, IPO, Edinburgh and Dublin have all developed phonological approaches to intonation synthesis which model the pitch contour as a sequence of pitch accents and boundaries. These approaches have been applied mainly to Germanic languages, and have had considerable success in both laboratory and commercial synthesis systems. The phonological frameworks adopted are based on the work of Bruce, 't Hart and colleagues, Ladd and Monaghan. A fourth approach, that of Pierrehumbert and colleagues (Pierrehumbert 1980; Hirschberg and Pierrehumbert 1986), has been employed by various European institutions. The assumptions underlying all these approaches are that the pitch contour realises a small number of phonological events, aligned with key elements at the segmental level, and that these phonological events are themselves the (partial) realisation of a linguistic structure which encodes syntactic and semantic relations between words and phrases at both the utterance level and the discourse level. Important outputs of this work include:

- classifications of pitch accents and boundaries (major, minor; declarative, interrogative; etc.);
- rules for assigning pitch accents and boundaries to text or other inputs;
- mappings from accents and boundaries to acoustic correlates, particularly fundamental frequency.

One problem with phonological work related to synthesis is that it has generally aimed at specifying a 'neutral' prosodic realisation of each utterance. The rules were mainly intended for implementation in TTS systems, and therefore had to handle a wide range of input with a small amount of linguistic information to go on: it was thus safer in most cases to produce a bland, rather monotonous prosody than to attempt to assign more expressive prosody and risk introducing major errors. This has led to a situation where most TTS systems can produce acceptable pitch contours for some sentence types (e.g. declaratives, yes/no questions) but not for others, and where the prosody for isolated utterances is much more acceptable than that for longer texts and dialogues. The paradox here is that most theoretical linguistic research on prosody has concentrated on the rarer, non-neutral cases or on the prosody of extended dialogues, but this research generally depends on pragmatic and semantic information which is simply not available to current TTS systems. In some cases, such as the LAIP system, this paradox has been solved by augmenting the prosody rules with performance factors such as rhythm and information chunking, allowing longer stretches of text to be processed simply.

The problem of specifying pitch contours linguistically in larger contexts than the sentence or utterance has been addressed by projects at KTH, IPO, Edinburgh, Dublin and elsewhere, but in most cases the results are still quite inconclusive.

Work at Edinburgh, for instance, is examining the long-standing problem of pitch register changes and declination between intonational phrases: to date, the results neither support a declination-based model nor totally agree with the competing distinction between initial, final and medial intonational phrases (Clark, 1999). The mappings from text to prosody in larger units are dependent on many unpredictable factors (speaking style, speaker's attitude, hearer's knowledge, and the relation between speaker and hearer, to name but a few). In dialogue systems, where the message to be uttered is generated automatically and much more linguistic information is consequently available, the level of linguistic complexity is currently very limited and does not give much scope for prosodic variation. This issue will be returned to in the discussion of applications below.

Timing

Work on this aspect of prosody includes the specification of segmental duration, duration of larger units, pause length, speech rate and rhythm. Approaches to segmental duration are exclusively stochastic. They include neural net models (University of Helsinki, Czech Academy of Sciences, ICP Grenoble), inductive learning (J. Stefan Institute), and statistical modelling (LAIP, Telenor, Aix). The Aix approach is interesting, in that it uses simple DTW techniques to align a natural signal with a sequence of units from a diphone database: the best alignment is aassumed to be the one where the diphone midpoints match the phone boundaries in the original. ÖFAI provide a lengthy justification of stochastic approaches to segmental duration, and recent work in Dublin suggests reasons for the difficulties in modelling segmental duration. Our own experience at Aculab suggests that while the statistical accuracy of stochastic models may be quite high, their naturalness and acceptability are still no better than simpler rule-based approaches.

Some researchers (LAIP, Prague Institute of Phonetics, Aix, ICP) incorporate rules at the syllable level, based particularly on Campbell's (1992) work. The University of Helsinki is unusual in referring to the word level rather than syllables or feet. The Prague Institute of Phonetics refers to three levels of rhythmic unit above the segment, and is the only group to mention such an extensive hierarchy, although workers in Helsinki are investigating phrase-level and utterance-level timing phenomena.

Several teams have investigated the length of pauses between units, and most others view this as a priority for future work. For Slovene, it is reported that 'pause duration is almost independent of the duration of the intonation unit before the pause', and seems to depend on speech rate and on whether the speaker breathes during the pause: there is no mention of what determines the speaker's choice of when to breathe. Similar negative findings for French are reported by LAIP. KTH have investigated pausing and other phrasing markers in Swedish, based on analyses of the linguistic and information structure of spontaneous dialogues: the findings included a set of phrasing markers corresponding to a range of phonetic realisations such as pausing and pre-boundary lengthening. Colleagues in Prague note that segmental duration in Czech seems to be related to boundary type in a similar way, and workers in Aix suggest a four-way classification of segmental

duration to allow for boundary and other effects: again, this is similar to suggestions by Campbell and colleagues.

Speech rate is mentioned by several groups as an important factor and an area of future research. Monaghan (1991) outlines a set of rules for synthesising three different speech rates, which is supported by an analysis of fast and slow speech (Monaghan, Chapter 20, this volume). The Prague Institute of Phonetics has recently developed rules for various different rates and styles of synthesis. A recent thesis at LAIP (Zellner, 1998) has examined the durational effects of speech rate in detail. The LAIP team is unusual in considering that the temporal structure can be studied independently of the pitch curve. Their prosodic model calculates temporal aspects before the melodic component. Following Fujisaki's principles, fully calculated temporal structures serve as the input to F_0 modelling. LAIP claims satisfactory results for timing in French using stochastic predictions for ten durational segment categories deduced from average segment durations. The resultant predictions are constrained by a rule-based system that minimises the undesirable effects of stochastic modelling.

Intensity

The importance of intensity, particularly its interactions with pitch and timing, is widely acknowledged. Little work has been devoted to it so far, with the exception of the two Czech institutions who have both incorporated control of intensity into their TTS rules (see Dohalska, Chapter 12, this volume). Many other researchers have expressed an intention to follow this lead in the near future.

Languages

Some of the different approaches and results above may be due to the languages studied. These include Czech, Dutch, English, Finnish, French, German, Norwegian, Slovene, Spanish and Swedish. In Finnish, for example, it is claimed that pitch does not play a significant linguistic role. In French and Spanish, the syllable is generally considered to be a much more important timing unit than in Germanic languages. In general, it is important to remember that different languages may use prosody in different ways, and that the same approach to synthesis will not necessarily work for all languages. One of the challenges for multilingual systems, such as those produced by LAIP or Aculab, is to determine where a common approach is applicable across languages and where it is not.

There are, however, several important methodological differences which are independent of the language under consideration. The next section looks at some of these methodologies and the assumptions on which they are based.

Methodologies

The two commonest methodologies in European prosody research are the purely stochastic corpus-based and the linguistic knowledge-based approaches. The former

is typified by work at ICP or Helsinki, and the latter by IPO or KTH. These methodologies differ essentially in whether the goal of the research is simply to model certain acoustic events which occur in speech (the stochastic approach) or to discover the contributions to prosody of various non-acoustic variables such as linguistic structure, information content and speaker characteristics (the knowledge-based approach). This is nothing new, nor is it unique to Europe. There are, however, some new and unique approaches both within and outside these established camps which deserve a mention here.

Research at ICP, for example, differs from the standard stochastic approach in that prosody is seen as 'a direct encoding of meaning via prototypical prosodic patterns'. This assumes that no linguistic representations mediate between the cognitive/semantic and acoustic levels. The ICP approach makes use of a corpus with annotation of P-Centres, and has been applied to short sentences with varying syntactic structures. Based on syntactic class (presumably a cognitive factor) and attitude (e.g. assertion, exclamation, suspicious irony), a neural net model is trained to produce prototypical durations and pitch contours for each syllable. In principle, prototypical contours from these and many other levels of analysis can be superimposed to create individual timing and pitch contours for units of any size.

Research at Joensuu was noted above as being unusually eclectic, and concentrates on assessing the performance of different theoretical frameworks in predicting prosody. ETH has similar goals, namely to determine a set of symbolic markers which are sufficient to control the prosody generator of a TTS system. These markers could accompany the input text (in which case their absence would result in some default prosody), or they could be part of a rich phonological description which specifies prominences, boundaries, contour types and other information such as focus domains or details of pitch range. Both the evaluation of competing prosodic theories and the compilation of a complete and coherent set of prosodic markers have important implications for the development of speech synthesis mark-up languages, which are discussed in the section on applications below.

LAIP and IKP both have a perceptual or psycho-acoustic flavour to their work. In the case of LAIP, this is because they have found that linguistic factors are not always sufficiently good predictors of prosodic control, but can be complemented by performance criteria. Processing speed and memory are important considerations for LAIPTTS, and complex linguistic analysis is therefore not always an option. For a neutral reading style, LAIP has found that perceptual and performance-related prosodic rules are often an adequate substitute for linguistic knowledge: evenly-spaced pauses, rhythmic alternations in stress and speech rate, and an assumption of uniform salience of information lead to an acceptable level of coherence and 'fluency'. However, these measures are inadequate for predicting prosodic realisations in 'the semantically punctuated reading of a greater variety of linguistic structures and dialogues', where the assumption of uniform salience does not hold true.

Recent research at IKP has concentrated on the notion of 'prominence', a psycholinguistic measure of the degree of perceived salience of a syllable and consequently of the word or larger unit in which that syllable is the most prominent. IKP proposes a model where each syllable is an ordered pair of segmental content

and prominence value. In the case of boundaries, the ordered pair is of boundary type (e.g. rise, fall) and prominence value. These prominence values are presumably assigned on the basis of linguistic and information structure, and encode hierarchic and salience relations, allowing listeners to reconstruct a prominence hierarchy and thus decode those relations.

The IKP theory assumes that listeners judge the prosody of speech not as a set of independent perceptions of pitch, timing, intensity and so forth, but as a single perception of prominence for each syllable: synthetic speech should therefore attempt to model prominence as an explicit synthesis parameter. 'When a synthetic utterance is judged according to the perceived prominence of its syllables, these judgements should reflect the prominence values [assigned by the system]. It is the task of the phonetic prosody control, namely duration, F0, intensity and reductions, to allow the appropriate perception of the system parameter.' Experiments have shown that phoneticians are able to assign prominence values on a 32–point scale with a high degree of consistency, but so far the assignment of these values automatically from text and the acoustic realisation of a value of, say, 22 in synthetic speech are still problematic.

Applications

By far the commonest application of European synthetic prosody research is in TTS systems, mainly laboratory systems but with one or two commercial systems. Work oriented towards TTS includes KTH, IPO, LAIP, IKP, ETH, Czech Academy of Sciences, Prague Institute of Phonetics, British Telecom, Aculab and Edinburgh. The FESTIVAL system produced at CSTR in Edinburgh is probably the most freely available of the non-commercial systems. Other applications include announcement systems (Dublin), dialogue systems (KTH, IPO, IKP, BT, Dublin), and document browsers (Dublin). Some institutions have concentrated on producing tools for prosody research (Joensuu, Aix, UCL) or on developing and testing theories of prosody using synthesis as an experimental or assessment methodology.

Current TTS applications typically handle unrestricted text in a robust but dull fashion. As mentioned above, they produce acceptable prosody for most isolated sentences and 'neutral' text, but other genres (email, stories, specialist texts, etc.) rapidly reveal the shallowness of the systems' processing. There are currently two approaches to this problem: the development of dialogue systems which exhibit a deeper understanding of such texts, and the treatment of rich-text input from which prosodic information is more easily extracted.

Dialogue systems predict appropriate prosody in their synthesised output by analysing the preceding discourse and deducing the contribution which each synthesised utterance should make to the dialogue: e.g. is it commenting on the current topic, introducing a new topic, contradicting or confirming some proposition, or closing the current dialogue? Lexical, syntactic and prosodic choices can be made accordingly. There are two levels of prosodic analysis involved in such systems: the extraction of the prosodically-relevant information from the context, and the mapping from that information to phonetic or phonological specifications.

Extracting all the relevant syntactic, semantic, pragmatic and other information from free text is not currently possible. Small-domain systems have been developed in Edinburgh, Dublin and elsewhere, but these systems generally only synthesise a very limited range of prosodic phenomena since that is all that is required by their input. The relation between a speaker's intended contribution to a dialogue, and the linguistic choices which the speaker makes to realise that contribution, is only poorly understood: the incorporation of more varied and expressive prosody into dialogue systems will require progress in the fields of NLP and HCI among others.

More work has been done on the relation between linguistic information and dialogue prosody. IPO has recently embarked on research into 'pitch range phenomena, and the interaction between the thematic structure of the discourse and turn-taking'. Research at Aculab is refining the mappings from discourse factors to accent placement which were first developed at Edinburgh in the BRIDGE spoken dialogue generation system. Work at KTH has produced 'a system whereby markers inserted in the text can generate prosodic patterns based on those we observe in our analyses of dialogues', but as yet these markers cannot be automatically deduced.

The practice of annotating the input to speech synthesis systems has led to the development of speech synthesis mark-up languages at Edinburgh and elsewhere. The type of mark-up ranges from control sequences which directly alter the phonetic characteristics of the output, through more generic markers such as or , to document formatting commands such as section headings. With such an unconstrained set of possible markers, there is a danger that mark-up will not be coherent or that only trained personnel will be able to use the markers effectively.

One option is to make use of a set of markers which is already used for document preparation. Researchers in Dublin have developed prosodic rules to translate common document formats (LaTeX, HTML, RTF, etc.) into spoken output for a document browser, with interfaces to a number of commercial synthesisers. Work at the University of East Anglia is pursuing a multi-modal approach developed at BT, whereby speech can be synthesised from a range of different inputs and combined with static or moving images: this seems relatively unproblematic, given appropriate input.

The SABLE initiative (Sproat *et al.*, 1998) is a collaboration between synthesis researchers in Edinburgh and various US laboratories which has proposed standards for text mark-up specifically for speech synthesis. The current proposals mix all levels of representation and it is therefore very difficult to predict how individual synthesisers will interpret the mark-up: future refinements should address this issue. SABLE's lead has been followed by several researchers in the USA, but so far not in Europe (see Monaghan, Chapter 31, this volume).

Prosody in COST 258

At its Spring 1998 meeting, COST 258 identified four priority areas for research on synthetic prosody: prosodic and acoustic effects of focus and/or emphasis, prosodic effects of speaking styles, rhythm, and mark-up. These were seen as the most promising areas for improvement in synthetic speech, and many of the contributions in

this volume address one or more of these areas. In addition, several participating institutions have continued to work on pre-existing research programmes, extending their prosodic rules to new aspects of prosody (e.g. timing and intensity) or to new classes of output (interrogatives, emotional speech, dialogue, and so forth). Examples include the contributions by Dobnikar, Dohalská and Mixdorff in this part.

The work on speaking styles and mark-up has provided two separate parts of this volume, without detracting from the broad range of prosody research presented in the present section. I have not attempted to include this research in this summary of European synthetic prosody R&D, as to do so would only serve to paraphrase much of the present volume. Both in quantity and quality, the research carried out within COST 258 has greatly advanced our understanding of prosody for speech synthesis, and thereby improved the naturalness of future applications. The multilingual aspect of this research cannot be overstated: the number of languages and dialects investigated in COST 258 greatly increases the likelihood of viable multilingual applications, and I hope it will encourage and inform development in those languages which have so far been neglected by speech synthesis.

Acknowledgements

This work was made possible by the financial and organisational support of COST 258, a co-operative action funded by the European Commission.

Abbreviated names of research institutions

Aculab – Aculab plc, Milton Keynes, UK.
Aix – Laboratoire Langue et Parole, Université de Provence, Aix-en-Provence, France.
BT – British Telecom Research Labs, Martlesham, UK.
Dublin – NCLT, Computer Applications, Dublin City University, Ireland.
Edinburgh – CSTR, Department of Linguistics, University of Edinburgh, Scotland, UK.
ETH – Speech Group, ETH, Zurich, Switzerland.
Helsinki – Acoustics Laboratory, Helsinki University of Technology, Finland.
ICP – Institut de la Communication Parlée, Grenoble, France.
IKP – Institut für Kommunikationsforschung und Phonetik, Bonn, Germany.
IPO – Instituut voor Perceptie Onerzoek, Technical University of Eindhoven, Netherlands.
J. Stefan Institute – 'J.Stefan' Institute, Ljubljana, Slovenia.
Joensuu – General Linguistics, University of Joensuu, Finland.
KTH – Department of Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden.
LAIP – LAIP, University of Lausanne, Switzerland.
ÖFAI – Österreichisches Forschungsinstitut für Artificial Intelligence, Vienna, Austria.
Prague – Institute of Phonetics, Charles Universit, Prague, Czech Republic.

Telenor – Speech Technology Group at Telenor, Kjeller, Norway
UCL – Phonetics and Linguistics, University College London, UK.

References

Campbell, W.N. (1992). *Multi-level Timing in Speech*. PhD thesis, University of Sussex.

Clark, R. (1999). Using prosodic structure to improve pitch range variation in text to speech synthesis. *Proceedings of ICPhS*, Vol. 1 (pp. 69–72). San Francisco.

Hirschberg, J. and Pierrehumbert, J.B. (1986). The intonational structuring of discourse. *Proceedings of the 24th ACL Meeting* (pp. 136–144). New York.

Monaghan, A.I.C. (1991). *Intonation in a Text-to-Speech Conversion System*. PhD thesis, University of Edinburgh.

Pierrehumbert, J.B. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology.

Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., and Lenzo, K. (1998). SABLE: A standard for TTS markup. *Proceedings of the 3rd International Workshop on Speech Synthesis* (pp. 27–30). Jenolan Caves, Australia.

Zellner, B. (1998). Caractérisation et prédiction du débit de parole en français, Une étude de cas. Unpublished doctoral thesis, University of Lausanne.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

10

Modelling F0 in Various Romance Languages

Implementation in Some TTS Systems

Philippe Martin
University of Toronto
Toronto, ON, Canada M55 1A1
philippe.martin@utoronto.ca

Introduction

The large variability observed in intonation data (specifically for the fundamental frequency curve) has for a long time constituted a puzzling challenge, precluding to some extent the use of systematic prosodic rules for speech synthesis applications. We will try to show that simple linguistic principles allow for the detection of enough coherence in prosodic data variations, to lead to a grammar of intonation specific for each language, and suitable for incorporation into TTS algorithms. We then give intonation rules for French, Italian, Spanish and Portuguese, together with their phonetic realisations. We then compare the actual realisations of three TTS systems to the theoretical predictions and suggest possible improvements by modifying the F0 and duration of the synthesised samples according to the theoretical model.

We will start to recap the essential features of the intonation model that for a given sentence essentially predicts the characteristics of pitch movements on stressed and final syllables as well as the rhythmic adjustments observed on large syntactic groups (Martin, 1987; 1999). This model accounts for the large variability inherent to prosodic data, and is clearly positioned outside the dominant phonological approaches currently used to describe sentence intonation (e.g. Beckman and Pierrehumbert, 1986). It contrasts as well with stochastic models frequently implemented in speech synthesis systems (see, for instance, Botinis *et al*., 1997). The dominant phonological approach has been discarded because it oversimplifies the data by using only high and low tones and by the absence of any convincing linguistic role given to intonation. Stochastic models, on the other hand, while delivering acceptable predictions of prosodic curves appear totally opaque as to the linguistic functions of intonation.

The approach chosen will explain some regularities observed in romance languages such as French, Italian, Spanish and Portuguese, particularly regarding pitch movements on stressed syllables. Applying the theoretical model to some commercially available TTS systems and modifying their output using a prosodic morphing program (WinPitch, 1996), we will comment upon observed data and improvements resulting from these modifications.

A Theory of Sentence Intonation

Many events in the speaker activity contribute to the fundamental frequency curve:

- the cycle of respiration, which determines respiratory pauses and the declination line of F0 observed inside an expiration phase;
- the fine variations in the vibrations of the vocal folds during phonation (producing micro-melodic effects);
- the influence of emotional state on the speaker, and its socialised counterpart, the speaker attitude;
- the declarative or interrogative modality of the sentence, and its variations: command, evidence, doubt and surprise;
- the hierarchical division of the sentence to facilitate the listener in the decoding of the organisation of what the speaker says.

We will focus on the latter aspect, which has local and global components:

- local (phonetic): pertains to the details of the F0 realisation conditioned by socio-geographic conditions;
- global (linguistic): pertains to the oral structuring of the sentence.

One of the linguistic aspects of intonation (which includes syllable F0, duration and intensity) deals with speech devices which signal cohesion and division among pronounced (syntactic) units. This aspect implies the existence of a prosodic structure PS, which defines a hierarchical organisation in the spoken sentence, *a priori* independent from the syntactic structure. The units organised by the PS are called prosodic words (or accentual units) containing only one stressed syllable (non-emphatic).

It can be shown (despite some recurrent scepticism among researchers in the field) that the PS is encoded by pitch movements located on stressed syllables (and occasionally on final syllables in romance languages other than French). These movements are not conditioned by the phonetic context, but rather by a set of rules to encode the PS and specific to each language. As other phonological entities such as vowels and consonants, the movements show phonological and phonetic characteristics such as neutralisation if locally redundant to indicate the prosodic structure, possible different phonetic realisation for each language and each dialect, etc.

The prosodic structure PS is not totally independent from syntax, and is governed by a set of constraints:

- size of the prosodic word (or accentual unit), which determines the maximum number of syllables depending of the rate of speech (typically 7) (Wioland, 1985);
- the stress clash condition, preventing the presence of two consecutive stressed syllables without being separated by a pause of some other phonetic spacing device (e.g. a consonant cluster of glottal stop) (Dell, 1984);
- the syntactic clash condition, preventing the grouping of accentual units not dominated by the same node in the syntactic structure (Martin, 1987);
- the eurhythmicity condition which express the tendency to prefer among all possible PS that can be associated with a given syntactic structure the one that balances the number of syllables in prosodic groups at the same level in the structure, or alternatively, the use of a faster speech rate for groups containing a large number of syllables, and a slower rate for groups with a small number of syllables (Martin, 1987).

A set of rules specific to each language then generates pitch movements from a given prosodic structure. The movements are described phonologically in terms of height (High–Low), slope (Rising–Falling), amplitude of melodic variation (Ample–Restrained), and so forth. Phonetically, they are realised as pitch variations taking place along the overall declination line of the sentence.

As they depend on other phonetic parameters such as speaker gender and emotion, rate of speech, etc., pitch contours do not have absolute values but maintain, according to their phonological properties, relations of differentiation with the other pitch markers appearing in the sentence. Therefore, they are defined by the differences they have to maintain with the other contours to function as markers of the prosodic structure, and not by some frozen value of pitch change and duration. In a sentence with a prosodic structure such as (A B) (C) for instance, where A, B and C are prosodic words (accentual units), and given a falling declarative contour on unit C, the B contour has to be different from A and C (in French it will be rising and long), and A must be differentiated with B and C. The differences are implemented according to rules, which in French, for instance, specify that a given pitch contour must have an opposite slope (i.e. rising vs. falling) to the pitch contour ending the prosodic group to which it belongs. So in (A B) (C), if C contour is falling, B will be rising and C falling. Furthermore, C will be differentiated from C by some other prosodic feature, in this case the height and amplitude of melodic variation (see details in Martin, 1987).

Given these principles, it is possible to discover the grammar of pitch contours for various languages. In French, for instance, for a PS ((A B) (C D)) (.....), where the first prosodic group corresponds to a subject noun phrase, we find:

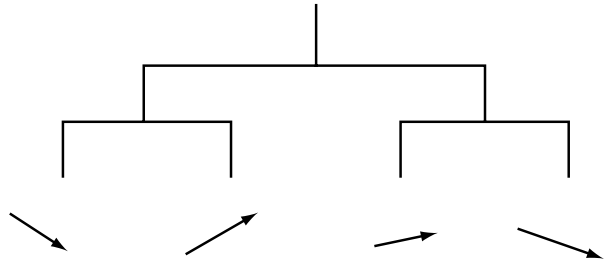


Figure 10.1

whereas for romance languages such as Italian, Spanish and Portuguese, we have:

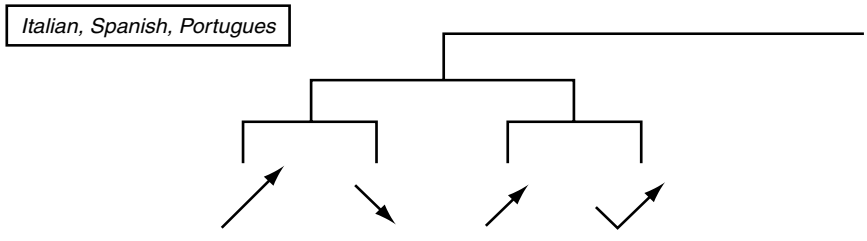


Figure 10.2

The phonetic realisations of these contours (i.e. the fine details of the melodic variations) will be of course different for each romance language of this group.

Figures 10.3 to 10.6 show F0 curves for French, Italian, Spanish and (European) Portuguese for an example using with very similar syntactic structure for each language. These curves were obtained by analysing sentences read by native speakers of the languages considered. Stressed syllables are shown by in circles (solid lines), group final syllables with circles in dotted lines.

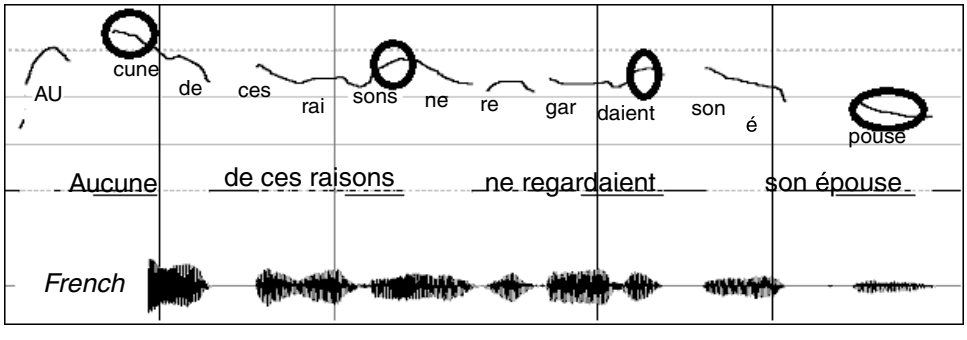


Figure 10.3 This example *Aucune de ces raisons ne regardaient son épouse* shows a declarative falling contour on *épouse*, to which is opposed the rising contour ending the group *Aucune de ces raisons*. At the second level of the hierarchical prosodic organisation on the sentence, the falling contour on *Aucune* is opposed to the final rise on the group *Aucune de ces raisons*, and the rise with moderate pitch variation on *ne regardaient* opposed to the final fall in *ne regardaient son épouse*

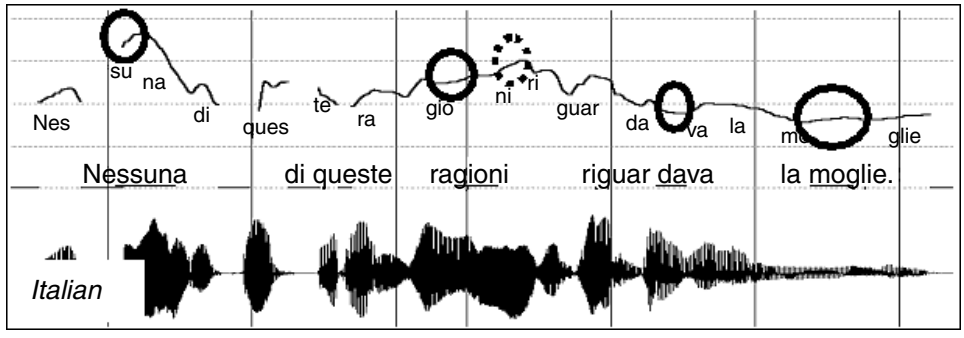


Figure 10.4 The first stressed syllable has a high and rising contour on *Nessuna*, opposed to a complex contour ending the group *Nessuna di queste ragioni*, where the rather flat F0 is located on the stressed syllable, and the final syllable has a rising pitch movement. The complex contour on such sentence initial prosodic groups has a variant where a rise is found on the (non-final) stressed syllable and any movement rise, flat or fall on the last syllable

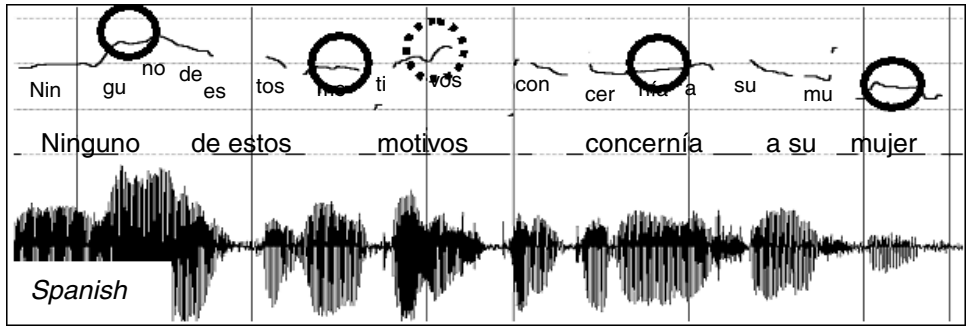


Figure 10.5 Spanish exhibits similar phonological pitch contours but with different phonetic realisations: rises are not so sharp and the initial pitch rise is not so high

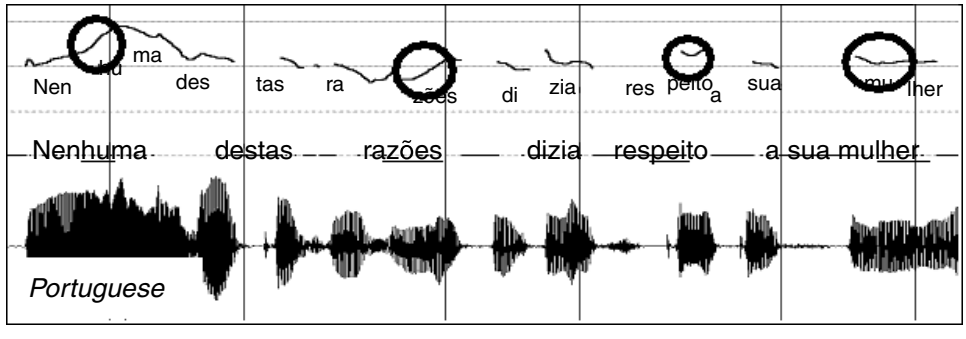


Figure 10.6 The same pitch variations appear on the Portuguese example, an initial rise and a rise of the final and stressed syllable ending the group *Nenhuma destas razões*

Comparison of Various TTS Systems for French

Among numerous existing TTS systems for French, Italian, Spanish and Portuguese, seven systems were evaluated for their realisations of F0 curves, and subjected to the theoretical predictions and the natural realisations on comparable examples: Bell Labs (2001), Elan (2000), LATL (2001), LAIPTTS (2001), L & H TTS (2000), Mons (2001) and SyntAix (2001). These systems were chosen for the availability of demo testing through the Internet at the time of writing.

Figures 10.8 to 10.15 successively show seven realisations of F0 curves: natural, Mons, Bell Labs, Elan, LATL, LAIPTTS, SyntAix, and L & H TTS. The examples commented here were taken from a much larger set representing a comprehensive group of various prosodic structures in the four romance languages studied.

For the first example, the theoretical sequence of prosodic contour is:

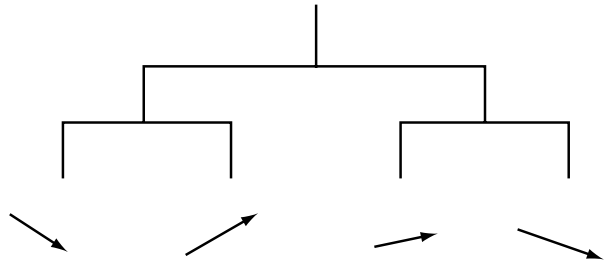


Figure 10.7

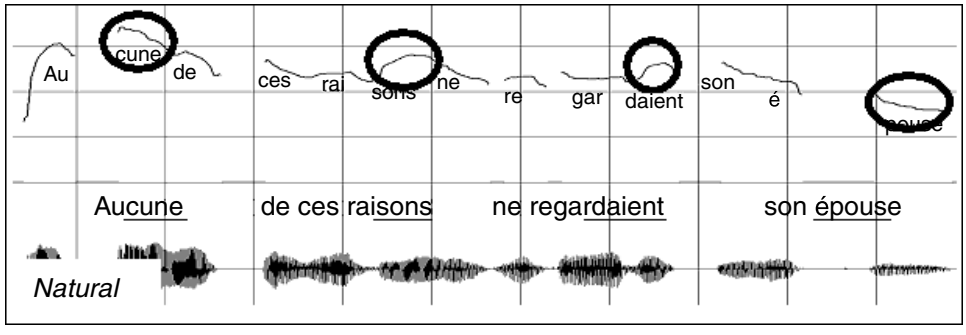


Figure 10.8 The natural speech shows the expected theoretical sequence of contours: falling high, rising, rising moderate and falling low

In the following figures, the straight lines traced over the F0 contour and (occasionally) on the speech wave represent the changes in F0 and segment duration made to modify the original F0 and segment duration. These changes were made using the using the WinPitch software (WinPitch, 1996). The original and modified speech sounds can be found on the Webpage in wave format.

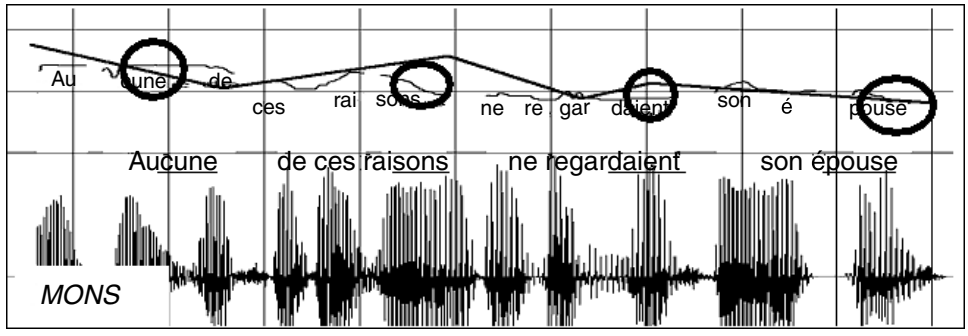


Figure 10.9 The Mons realisation of the same example exhibit contours in disagreement with theoretical and natural sequences. The effect of melodic variations changes through prosodic morphing can be judged from the re-synthesised wave sample

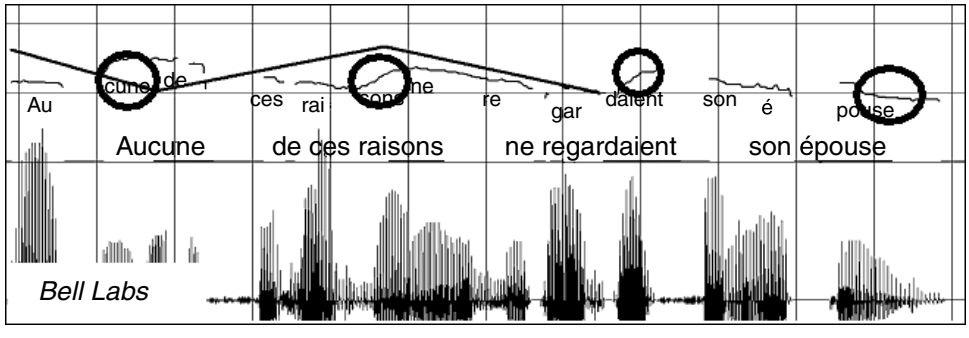


Figure 10.10 The Bell Labs realisation of the same example exhibit contours in good agreement with theoretical and natural sequences. Enhancing the amplitude induced a better perception of the major syntactic boundary

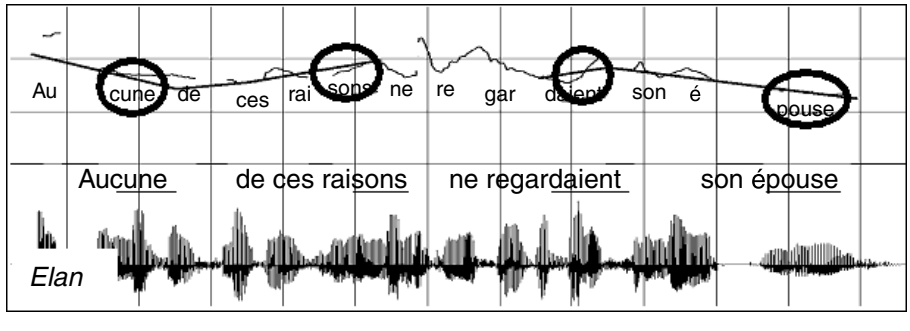


Figure 10.11 The ELAN realisation of the same example exhibit contours in agreement with theoretical and natural sequences, but augmenting the amplitude of melodic variations with prosodic morphing did enhance naturalness

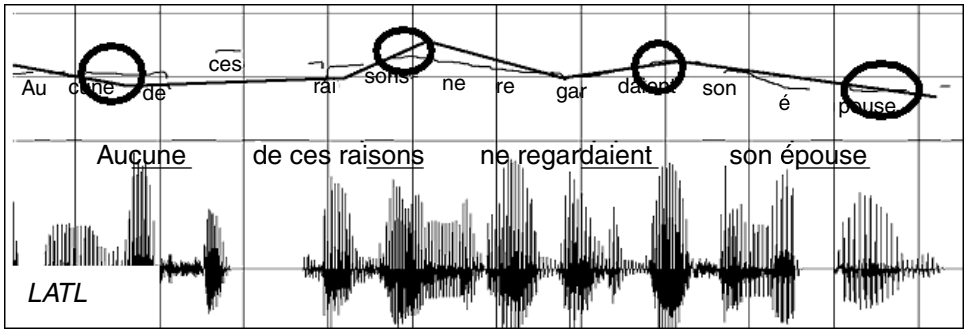


Figure 10.12 LATL. The pitch movements are somewhat in agreement with theoretical and natural sequences. Correcting a wrongly positioned pause on *ces* and enhancing the pitch variations improved the overall naturalness

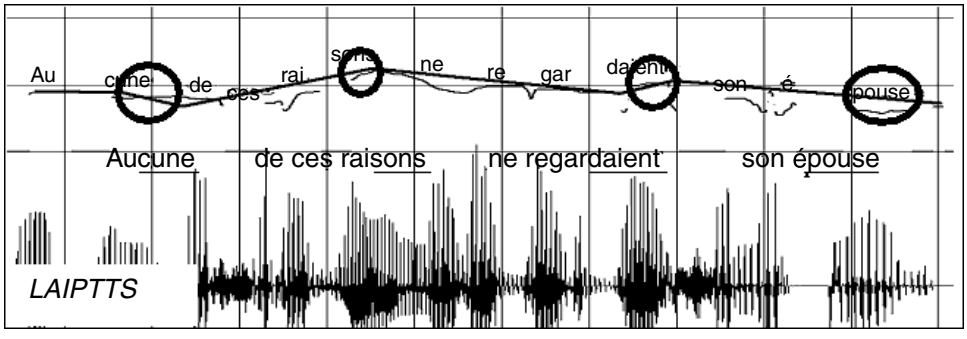


Figure 10.13 The LAIPTTS example manifests a very good match with natural and theoretical pitch movements. Re-synthesised speech using theoretical contrasts in fall and rise on stressed syllables brings no perceivable changes

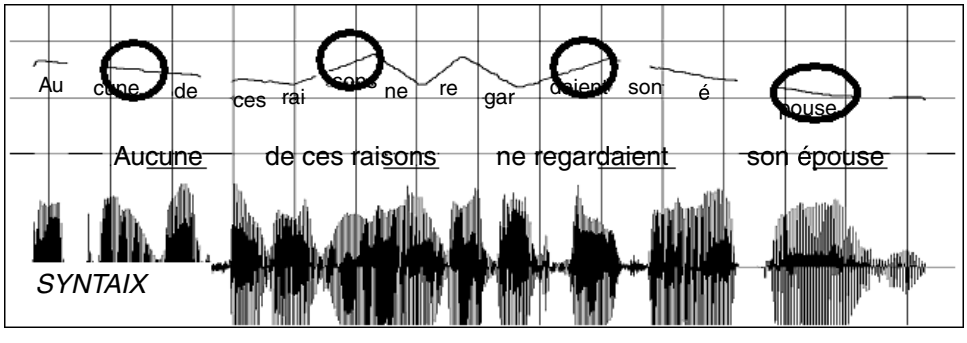


Figure 10.14 The SyntAix example manifests a good match with natural and theoretical pitch movements, and uses the rule of contrast of slope in melodic variation on *aucune* and *raison* (which seems somewhat in contradiction with the principles described in the author's paper, Di Cristo *et al*., 1997)

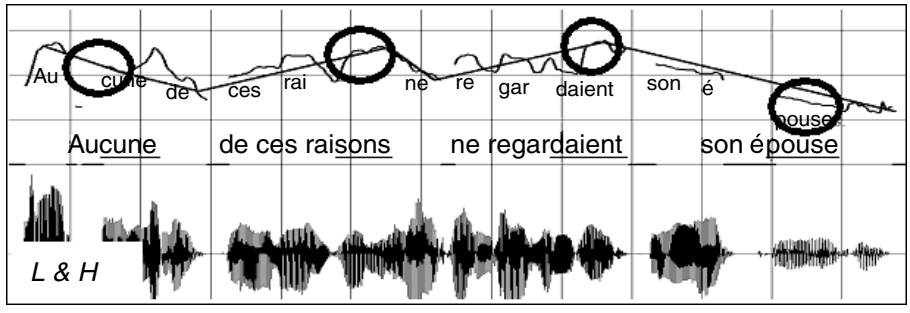


Figure 10.15 L & H: This example apparently uses unit selection for synthesis, and this case shows pitch contours on stressed syllables similar to natural and theoretical ones

The next example *Un groupe de chercheurs allemands a résolu l'énigme* has the following prosodic structure, indicated by a sequence of contours rising moderate, falling high, rising high, rising moderate and falling low.

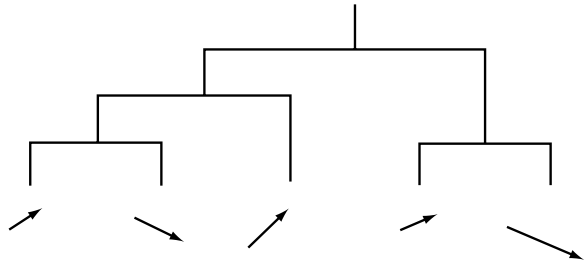


Figure 10.16

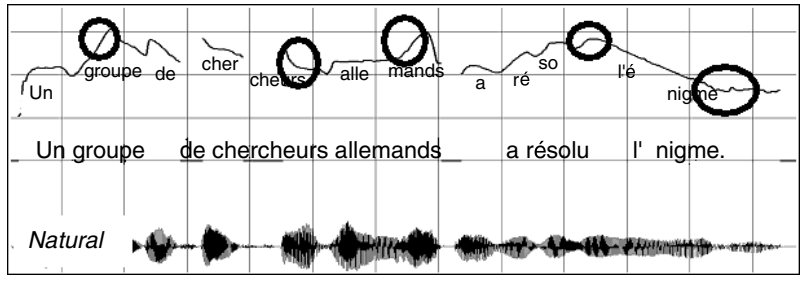


Figure 10.17 The natural F0 curve shows the expected variations and levels, with a neutralised realisation on the penultimate stressed syllable on *a résolu*

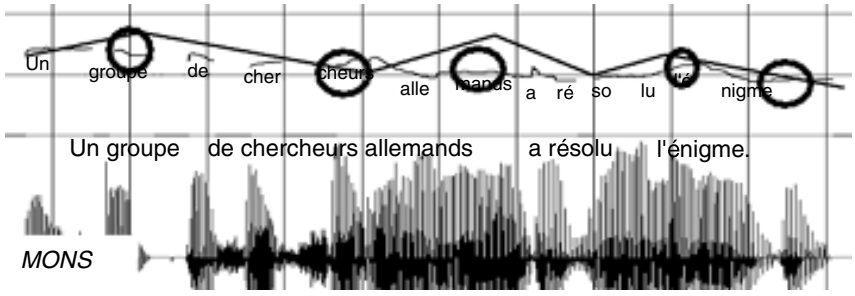


Figure 10.18 Mons. This realisation diverges considerably from the predicted and natural contours, with a flat melodic variation on the main division of the prosodic structure (final syllable of *allemand*). The re-synthesised sample uses the theoretical pitch movements to improve naturalness

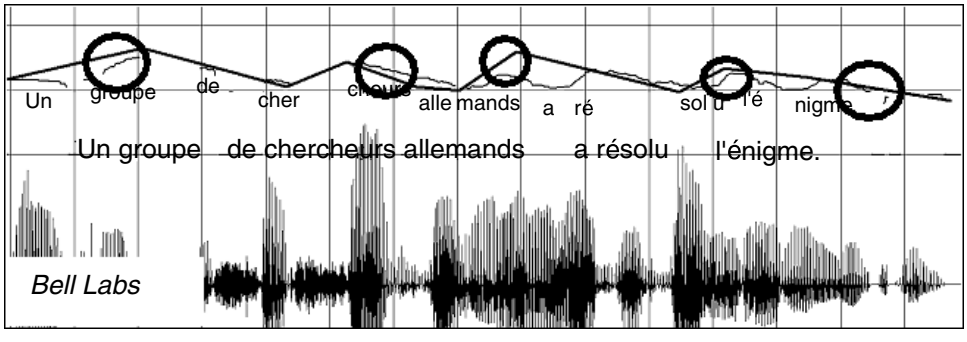


Figure 10.19 Bell Labs. This realisation is somewhat closer to the predicted and natural contours, except for the insufficient rise on the final syllable of the group *Un groupe de chercheurs allemand*. Re-synthesis was done by augmenting the rise on the stressed syllable of *allemand*

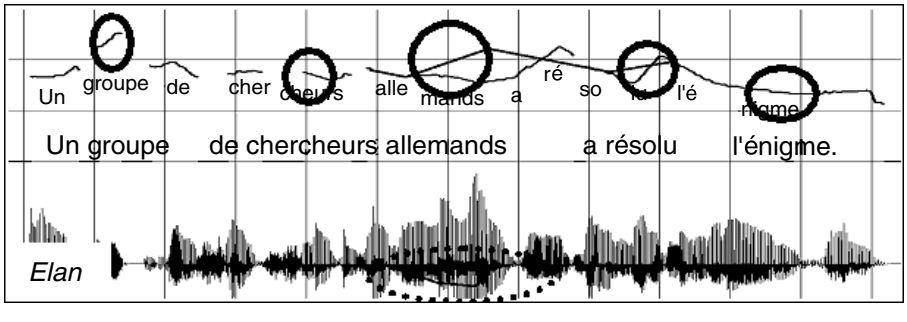


Figure 10.20 Elan. This realisation is close to the predicted and natural contours, except for the rise on the final syllable of the group *Un groupe de chercheurs allemand*. Augmenting the rise on the stressed syllable of *allemand* and selecting a slight fall on the first syllable of *a résolu* did improve naturalness considerably

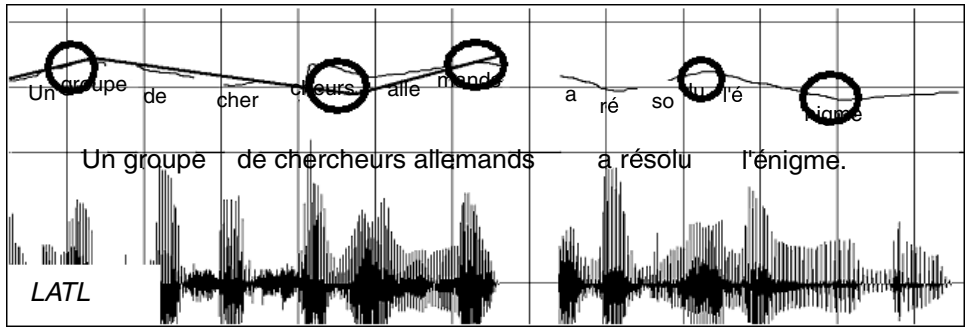


Figure 10.21 LATL. This realisation is close to the predicted and natural contours

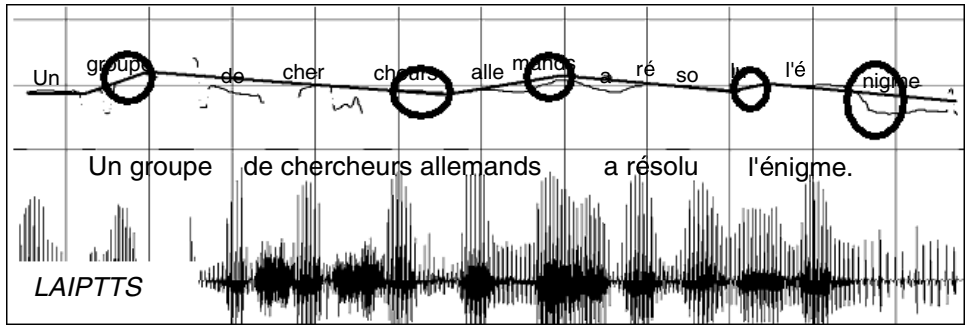


Figure 10.22 LAIPTTS. Each of the pitch movements on the stressed syllable are close to natural observations and theoretical predictions. Modifying the pitch variation according to the sequence seen above brings almost no change in naturalness

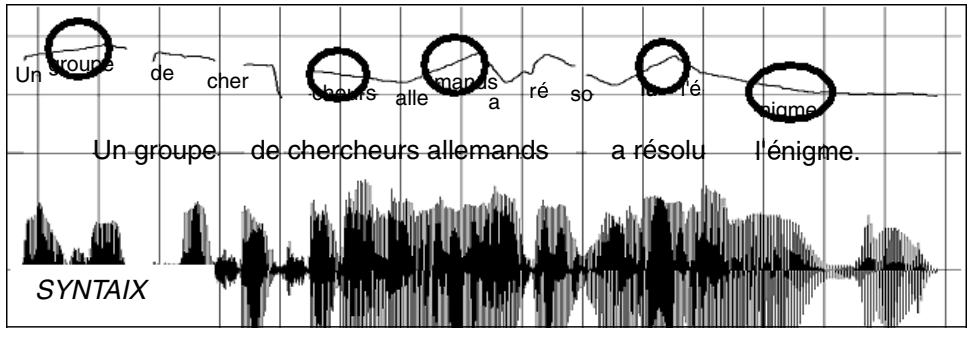


Figure 10.23 SyntAix. Here again there is a good match with natural and theoretical pitch movements, using slope contrast in melodic variation

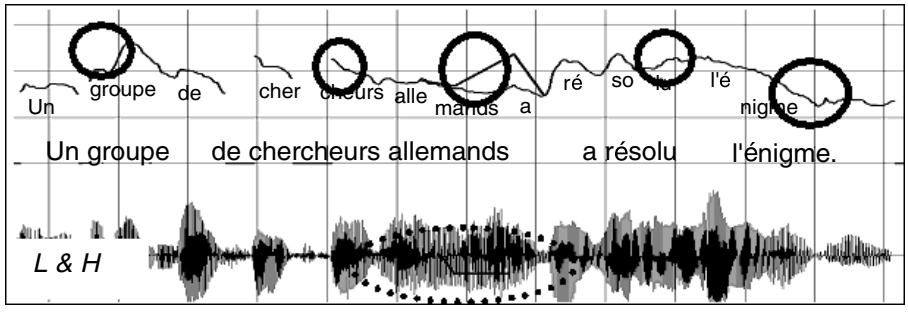


Figure 10.24 L & H. The main difference with the theoretical sequence pertains to the lack of rise on *allemands*, which is not perceived as stressed. Giving it a pitch rise and syllable lengthening will produce a more natural sounding sentence

The next set of examples deal with examples in Italian. The sentence *Alcuni edifici si sono rivelati pericolosi* is associated with a prosodic structure indicated by stressed syllables with high-rise, complex rise, moderate rise and fall low. The complex rising contour has variants, which depend on the complexity of the structure and the final or non-final position of the group last stress (see more details in Martin, 1999).

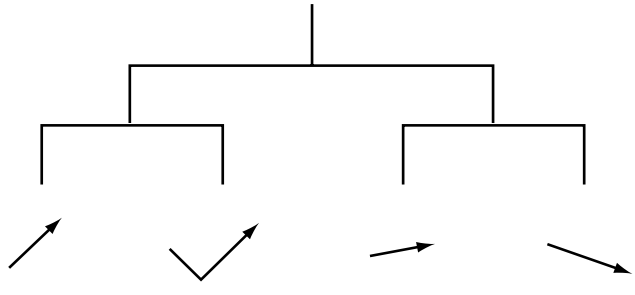


Figure 10.25

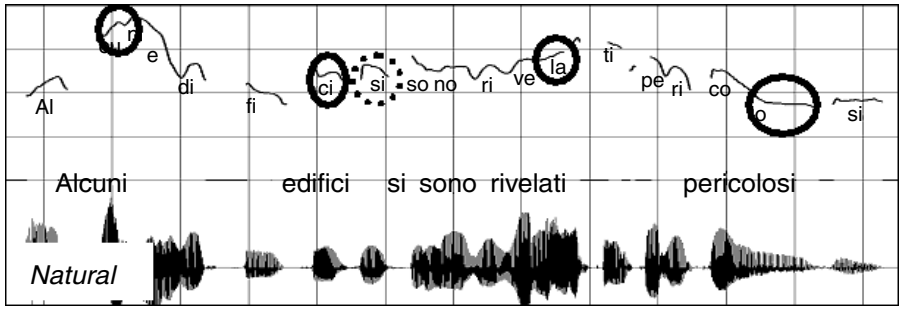


Figure 10.26 Natural. The F0 evolution for the natural realisation shows the predicted movements on the sentence stressed syllables

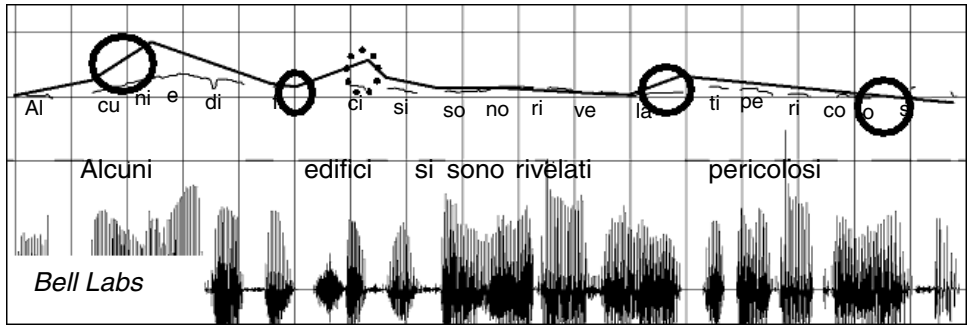


Figure 10.27 The Bell Labs sample shows the initial rise on *grupo*, but no complex contour on *edifici* (low flat on the stressed syllable, and rise of the final syllable). This complex contour is implemented in the re-synthesised version

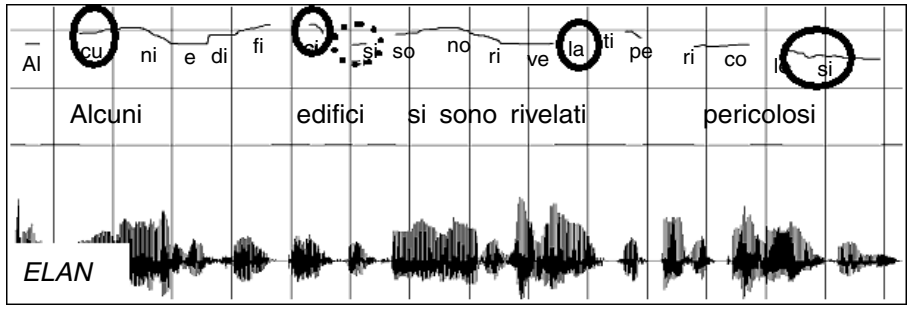


Figure 10.28 Elan. The pitch contours on stressed syllables are somewhat closer to theoretical and natural movements

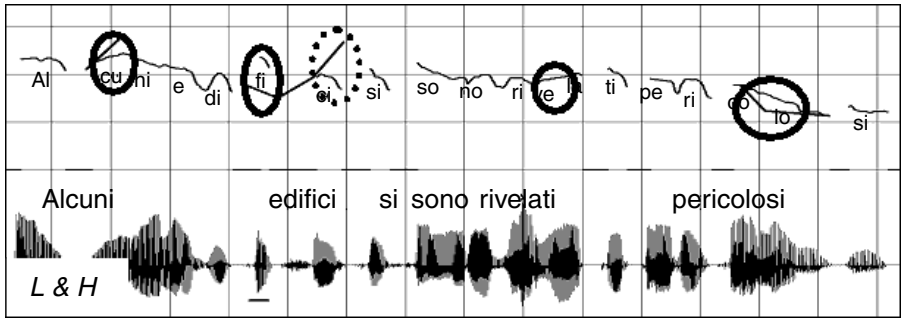


Figure 10.29 L & H pitch curve is somewhat close to the theoretical predictions, but enhancing the complex contour pitch changes on *edifici* with a longer stressed syllable did improve the overall auditory impression

A Spanish example is *Un grupo de investigadores alemanes ha resuelto l'enigma*. The corresponding prosodic structure and prosodic markers are:

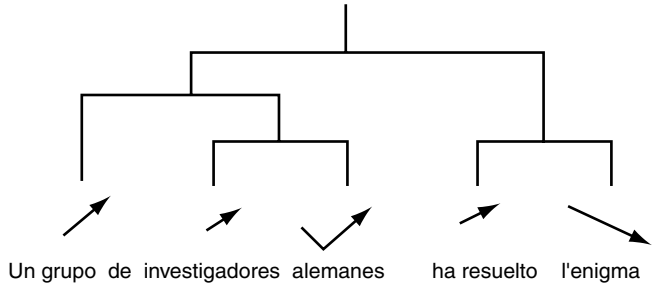


Figure 10.30

In this prosodic hierarchy, we have an initial high rise, a moderate rise, a complex rise, a moderate rise and a fall low.

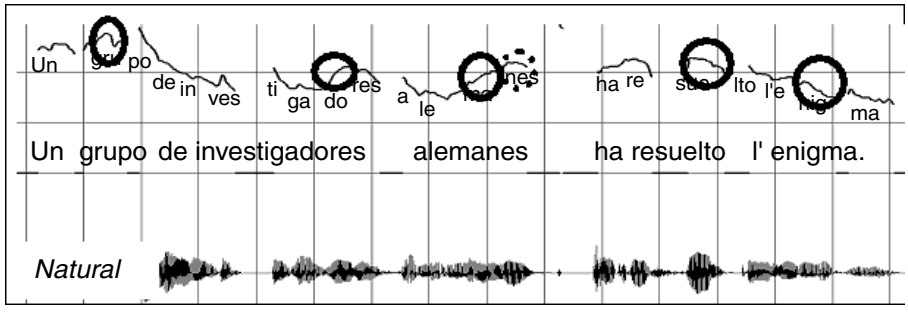


Figure 10.31 Natural. The natural example shows a stress rise and falling final variant of the complex rising contour ending the group *Un grupo de investigadores alemanes*

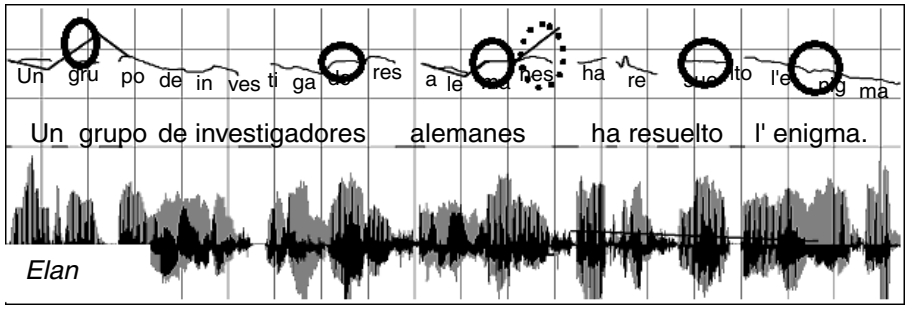


Figure 10.32 The ELAN example lacks the initial rise on *grupo*. Augmenting the F0 rise on the final syllable of *alemanes* did improve the perception of the prosodic organisation of the sentence

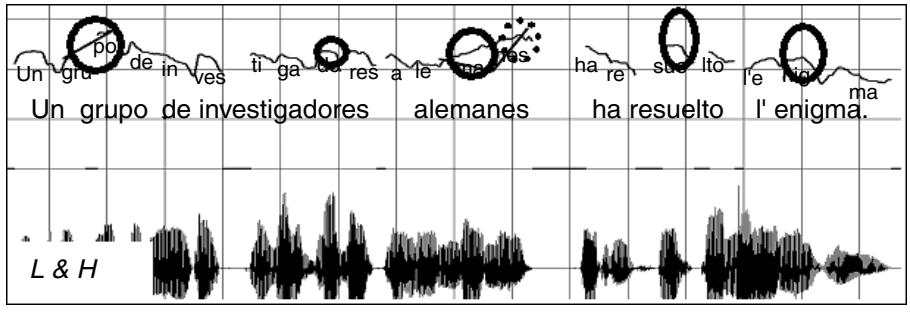


Figure 10.33 L & H. In this realisation, the initial rise and the complex rising contour were modified to improve the synthesis of sentence prosody

Conclusion

F0 curves depend on many parameters such as sentence modality, presence of focus and emphasis, syntactic structure, etc. Despite considerable variations observed in the data, a model pertaining to the encoding of a prosodic structure by pitch contours located on stressed syllables reveals the existence of a prosodic grammar specific to each language. We subjected the theoretical predictions of this model for French, Italian, Spanish and Portuguese to actual realisations of F0 curves produced by various TTS systems as well as natural speech. This comparison is of course quite limited as it involves mostly melodic variations in isolated sentences and ignores important timing aspects. Nevertheless, in many implementations for French, we can observe that pitch curves obtained either by rule or from unit selection approach are close to natural and theoretical predictions (this was far less the case a few years ago). In languages such as Italian and Spanish, however, the differences are more apparent and their TTS implementation could benefit from a more systematic use of linguistic description on sentence intonation.

Acknowledgements

This research was carried out in the framework of COST 258.

References

Beckman, M.E. and Pierrehumbert, J.B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, *3*, 255–309.

Bell Labs (2001) http://www.bell-labs.com/project/tts/french.html

Botinis, A., Kouroupetroglou, and Carayiannis, G. (eds) (1997). *Intonation: Theory, Models and Applications. Proceedings ESCA Workshop on Intonation*. Athens, Greece.

Dell, F. (1984). L'accentuation dans les phrases en français. In F. Dell, D. Hirst, and J.R. Vergnaud (eds), *Forme sonore du langage* (pp. 65–122). Hermann.

Di Cristo, A., Di Cristo, P., and Véronis, J. (1997). A metrical model of rhythm and intonation for French text-to-speech synthesis. In A. Botinis, Kouroupetroglou, and

G. Carayiannis, (eds), *Intonation: Theory, Models and Applications, Proceedings ESCA Workshop on Intonation* (pp. 83–86). Athens, Greece.
Elan (2000): http://www.lhsl.com/realspeak/demo.cfm
LAIPTTS (2001): http://www.unil.ch/imm/docs/LAIP/LAIPTTS.html
LATL (2001): http://www.latl.ch/french/index.htm
L & H TTS (2000): http://www.elan.fr/speech/french/index.htm
Martin, P. (1987). Prosodic and rhythmic structures in French. *Linguistics*, *25–5*, 925–949.
Martin, P. (1999). Prosodie des langues romanes: Analyse phonétique et phonologie. *Recherches sur le français parlé. Publications de l'Université de Provence*, *15*, 233–253.
Mons (2001) http://babel.fpms.ac.be/French/
SyntAix (2001) http://www.lpl.univ-aix.fr/∼roy/cgi-bin/metlpl.cgi
WinPitch (1996) http://www.winpitch.com/
Wioland, F. (1985). *Les Structures rythmiques du français*. Slatkine-Champion.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

11

Acoustic Characterisation of the Tonic Syllable in Portuguese

João Paulo Ramos Teixeira and Diamantino R.S. Freitas
E.S.T.I.G.-I.P. Bragança and C.E.F.A.T. (F.E.U.Porto), Portugal
joaopt@ipb.pt,dfreitas@fe.up.pt

Introduction

In developing prosodic models to improve the naturalness of synthetic speech it is assumed by some authors (Andrade and Viana, 1988; Mateus *et al.*, 1990; Zellner, 1998) that accurate modelling of tonic syllables is crucially important. This requires the modification of the acoustic parameters duration, intensity and F0, but there are no previously published works that quantify the variation of these parameters for Portuguese.

F0, duration or intensity variation in the tonic syllable may depend on their function in the context, the word length, the position of the tonic syllable in the word, or the position of this word in the sentence (initial, medial or final). Contextual function will not be considered, since it is not generally predictable by a TTS system, and the main objective is to develop a quantified statistical model to implement the necessary F0, intensity and duration variations on the tonic syllable for TTS synthesis.

Method

Corpus

A short corpus was recorded with phrases of varying lengths in which a selected tonic syllable that always contained the phoneme [ϵ] was analysed, in various positions in the phrases and in isolated words, bearing in mind that this study should be extended, in a second stage, to a larger corpus with other phonemes and with refinements in the method resulting from the first stage.

Two words were considered for each of the three positions of the tonic syllable (final, penultimate and antepenultimate stress). Three sentences were created with

each word, and one sentence with the word isolated was also considered, giving a total of 24 sentences. The characteristics of the tonic syllable were then extracted and analysed in comparison to a neighbouring reference syllable (unstressed) in the same word (e.g. **fe***rro*, A**mé***lia*, *ca***fé**: bold = tonic syllable, italic = reference syllable).

Recording Conditions

The 24 sentences were read by three speakers (H, J and E), two males and one female. Each speaker read the material three times. Recording was performed directly to a PC hard disk using a 50 cm unidirectional microphone and a sound card (16 bits, 11 kHz). The room used was only moderately soundproofed.

Signal Analysis

The MATLAB package was used for analysis, and appropriate measuring tools were created. All frames were first classified into voiced, unvoiced, mixed and silence. Intensity in dB was calculated as in Rowden (1992), and in voiced sections the F0 contour was extracted using a cepstral analysis technique (Rabiner and Schafer, 1978). These three aspects of the signal were verified by eye and by ear. The following values were recorded for tonic syllables (T) and reference syllables (R): syllable duration (D_T – tonic and D_R – reference), maximum intensity (I_T and I_R), and initial (F_A and F_C) and final (F_B and F_D) F0 values, as well as the shape of the contour.

Results

Duration

The relative duration for each tonic syllable was calculated by the relation (D_T / D_R) $\times$ 100 (%). For each speaker the average relative duration of the tonic syllable was determined and tendencies were observed for the position of the tonic syllable in the word and the position of this word in the phrase. The low values for the standard deviation in Figure 11.1 show that the patterns and ranges of variation are quite similar across the three speakers, leading us to conclude that variation in relative duration of the tonic syllable is speaker independent.

Figure 11.2 shows the average duration $\pm 2 \cdot \sigma$ (σ-standard deviation) of the tonic relative to the reference syllable for all speakers at 95% confidence. A general increase can be seen in the duration of the tonic syllable from the beginning to the end of the word. Rules for tonic syllable duration can be derived from Figure 11.2, based on position in the word and the position of the word in the phrase. Table 11.1 summarises these rules.

Note that when the relative duration is less than 100%, the duration of the tonic syllable will be reduced. For instance, in the phrase '**Ho**je **é** dia do An**tó**nio to**mar** ca**fé**', the tonic syllable duration will be determined according to Table 11.2.

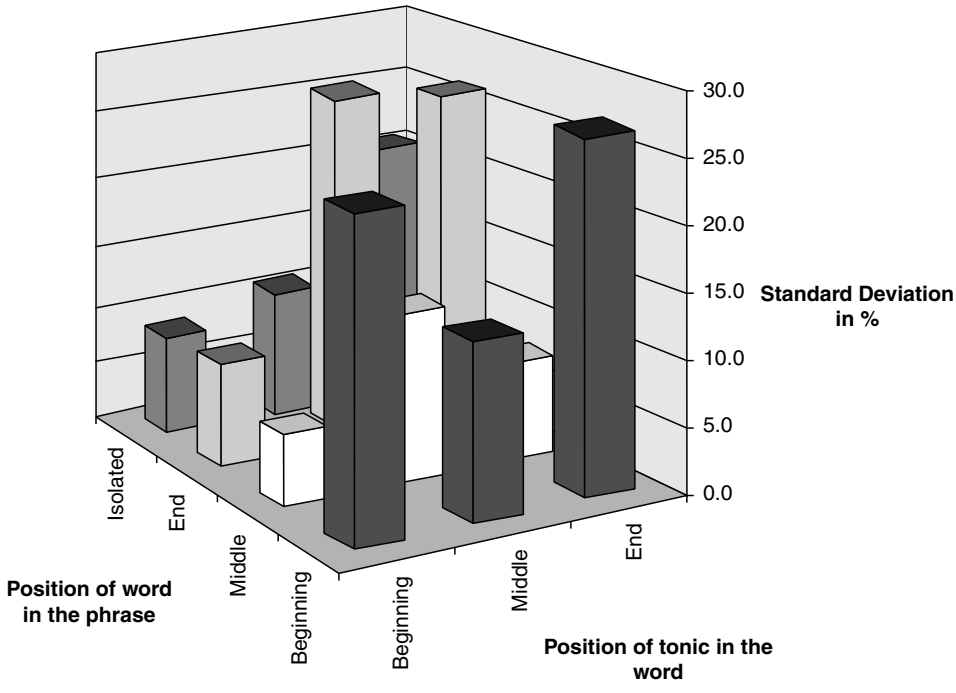


Figure 11.1 Standard deviation of average duration for the three speakers

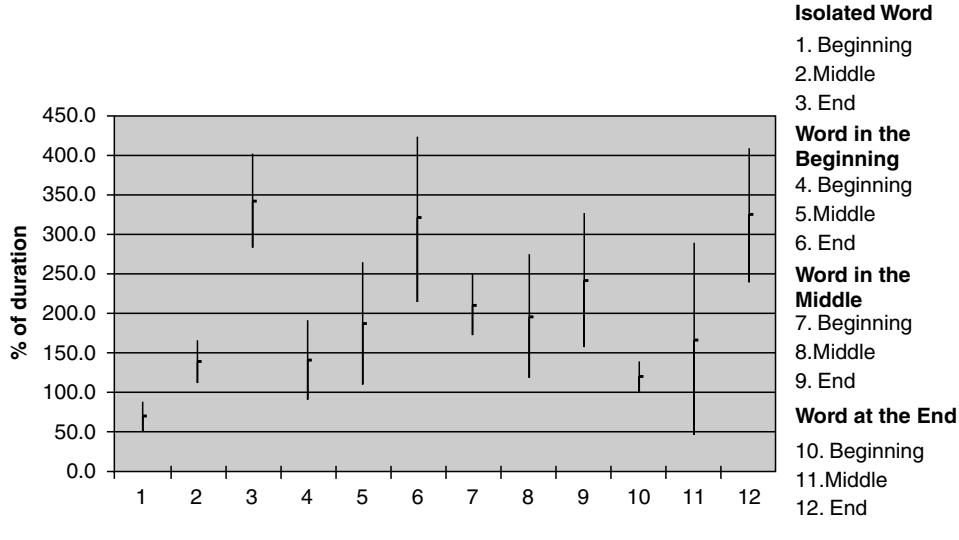


Figure 11.2 Average relative duration of tonic syllable for all speakers (95% confidence)

There are still some questions about these results. First, the reference syllable differs segmentally from the tonic syllable. Second, the results were obtained for a specific set of syllables and may not apply to other syllables. Third, in synthesising

Table 11.1 Duration rules for tonic syllables, values in %

Tonic syllable position	Isol. word	Phrase initial	Phrase medial	Phrase final
Beginning of word	69	140	210	120
Middle of word	139	187	195	167
End of word	341	319	242	324

Table 11.2 Example of application of duration rules

Tonic syllable	Position in word	Position of word in phrase	Relative duration (%)*
Ho	beginning	beginning	140
é	beginning	middle	210
tó	middle	middle	195
mar	end	middle	242
fé	end	end	324

Note: *Relative to the reference syllable.

a longer syllable, which constituents are longer? Only the vowel, or also the consonants? Does the type of consonant (stop, fricative, nasal, lateral) matter? A future study with a much larger corpus and a larger number of speakers will address these issues.

Depending on the type of synthesiser, these rules must be adapted to the characteristics of the basic units and to the particular technique. In concatenative diphone synthesis, for example, stressed vowel units are generally longer that the corresponding unstressed vowel and thus a smaller adjustment of duration will usually be necessary for the tonic vowel. However, the same cannot be said for the consonants in the tonic syllable.

Intensity

For each speaker the average intensity variation between tonic and reference syllables ($I_{T(dB)}$–$I_{R(dB)}$) was determined, in dB, according to the position of the tonic syllable in the word and the position of this word in the phrase. There are cross-speaker patterns of decreasing relative intensity in the tonic syllable from the beginning to the end of the word. Figure 11.3 shows the average intensity of the tonic syllable, plus and minus two standard deviations (95% confidence).

The standard deviation between speakers is shown in Figure 11.4. The pattern of variation for this parameter is consistent across speakers.

In contrast to the duration parameter, a general decreasing trend can be seen in tonic syllable intensity as its position changes from the beginning to the end of the word. Again, a set of rules can be derived from Figure 11.3, giving the change in intensity of the tonic syllable according to its position in the word

and in the phrase. Table 11.3 shows these rules. It can be seen that in cases 1, 2, 10 and 11 the inter-speaker variability is high and the rules are therefore unreliable.

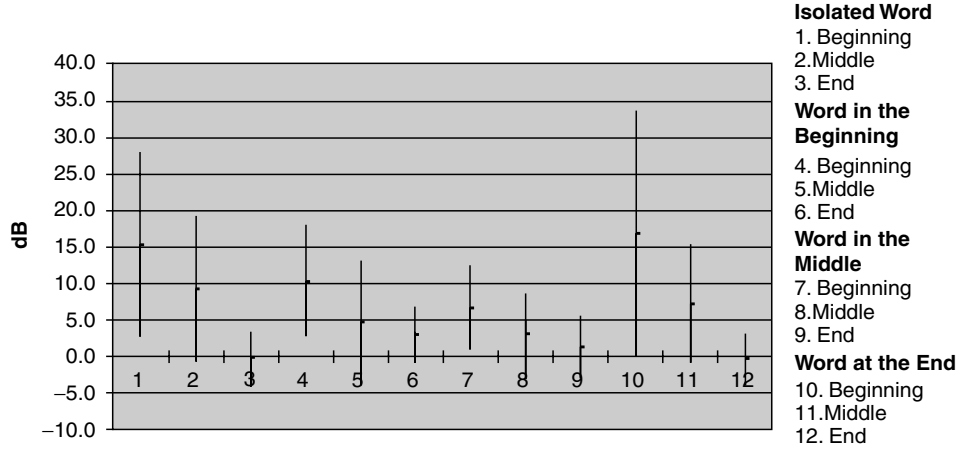


Figure 11.3 Average intensity of tonic syllable for all speakers (95% confidence)

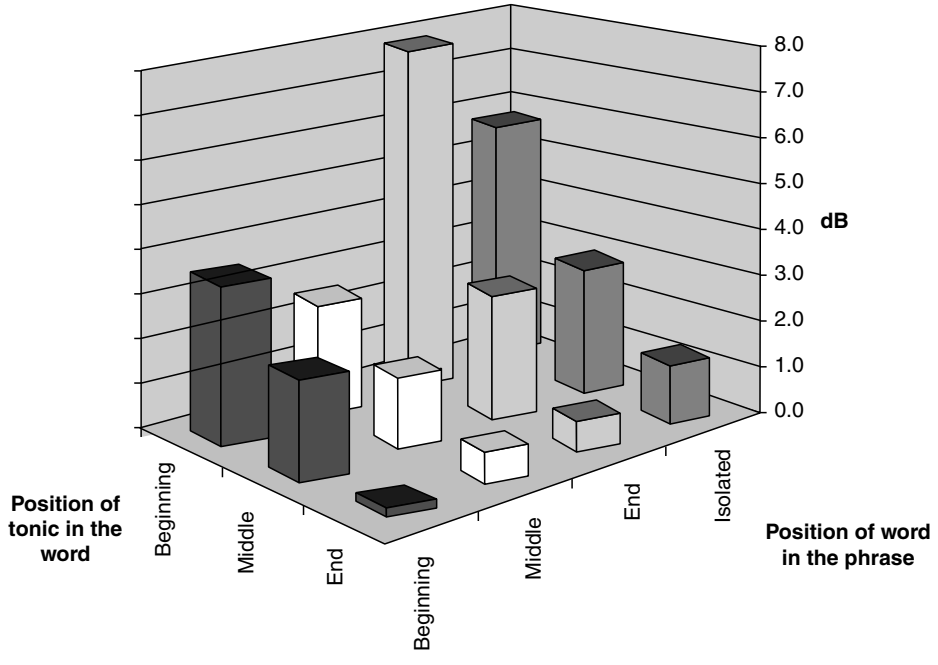


Figure 11.4 Standard deviation of intensity variation for the three speakers

Table 11.3 Change of intensity in the tonic syllable, values in dB

Tonic syllable position in the word	Isol. word	Phrase initial	Phrase medial	Phrase final
Beginning	15.2	10.3	6.6	16.8
Middle	9.2	4.6	3.0	7.2
End	−0.4	2.8	1.3	−0.4

Table 11.4 Example of the application of intensity rules

Tonic syllable	Position in the word	Position of word in phrase	Intensity (dB)*
Ho	beginning	beginning	10.3
é	beginning	middle	6.6
tó	middle	middle	3.0
mar	end	middle	1.3
fé	end	end	−0.4

Note: *Variation relative to the reference syllable.

Fundamental Frequency

The difference in F0 variation between tonic and reference syllables relative to the initial value of F0 in the tonic syllable $(((F_A - F_B) - (F_D - F_C))/F_A \times 100\ (\%))$ was determined for all sentences. As these syllables are in neighbouring positions, the common variation of F0 is the result of sentence intonation. The difference in

Table 11.5 F0 variation in the tonic syllable, values in %

Tonic syllable position in the word	Isol. word	Phrase initial word	Phrase medial word	Phrase final word
Beginning		5		
Middle				
end	−21	12.5		−12

Table 11.6 Example of the application of F0 rules

Tonic syllable	Position in word	Position of word in phrase	% of F0 variation*
o	beginning	beginning	5
é	beginning	middle	
tó	middle	middle	
mar	end	middle	
fé	end	end	−12

Note: *Relative to the F0 value at the beginning of the tonic syllable.

F0 variation in these two syllables is due to the tonic position. There are some cross-speaker tendencies, and some minor variations that seem irrelevant. Figure 11.5 shows average relative variation of F0, plus or minus two standard deviation, of the tonic syllable for all speakers.

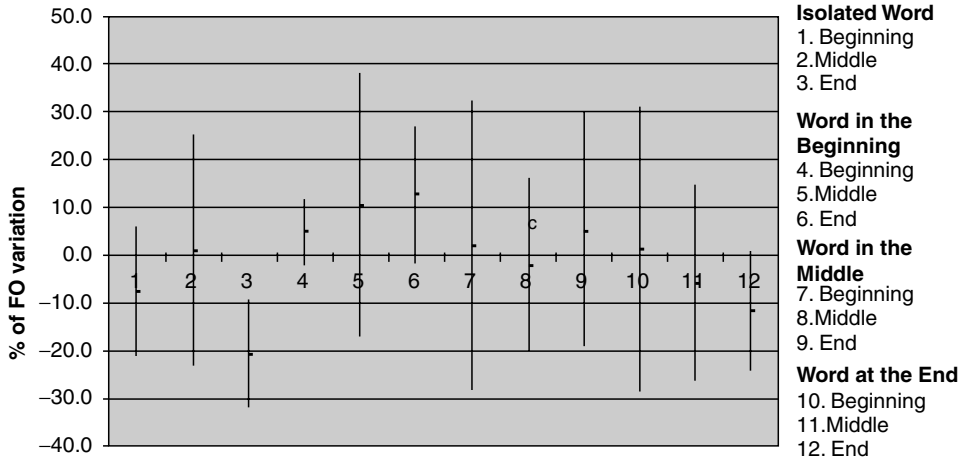


Figure 11.5 Average relative variation of F0 in tonic syllable for all speakers (95% confidence)

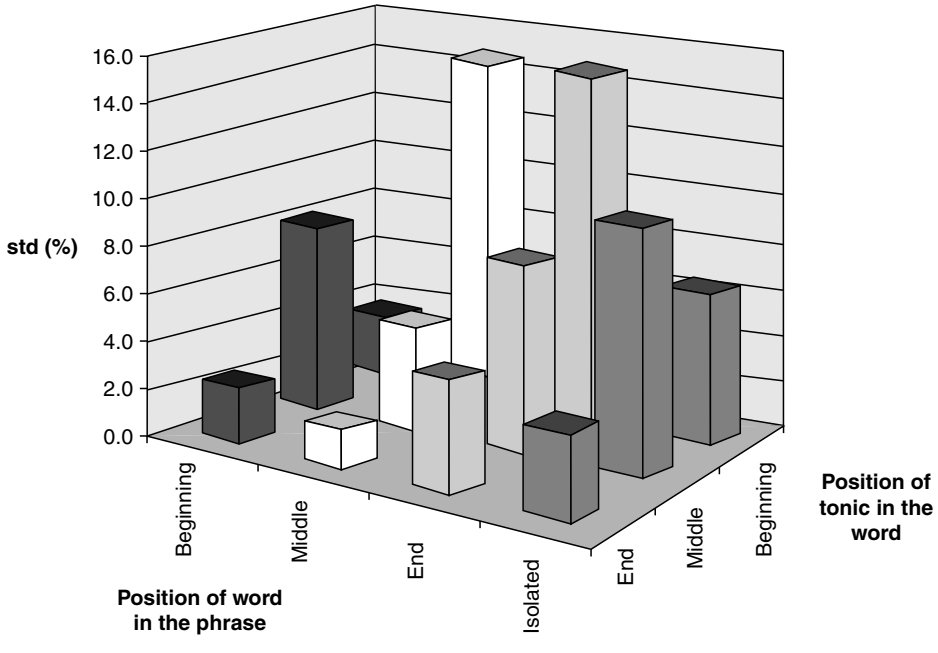


Figure 11.6 Standard deviation of F0 variation for the three speakers

Figure 11.6 shows the standard deviation for the three speakers. In some cases (low standard deviation) the F0 variation in tonic syllable is similar for the three speakers, but in other cases (high standard deviation) the F0 variation is very different. Reliable rules can therefore only be derived in a few cases. Table 11.5 shows the cases that can be taken as a rule. Table 11.6 gives an example of the application of these rules to the phrase '**Ho**je **é** dia do An**tó**nio to**mar** ca**fé**'.

Although only the values for F0 variation are reported here, the shape of the variation is also important. The patterns were observed and recorded. In most cases they can be approximated by exponential curves.

Conclusion

Some interesting variations of F0, duration and intensity in the tonic syllable have been shown as a function of their position in the word, for words in initial, medial and final position in the phrase and for isolated words. The analysis of the data is quite complex due to its multi-dimensional nature. The variations by position in the word are shown in Figures 11.2, 11.3 and 11.5, comparing the sets [1,2,3], [4,5,6], [7,8,9] and [10,11,12]. The average values of these sets show the effect of the position of the word in the phrase.

First, the variation of average relative duration and intensity of the tonic syllable are opposite in phrase-initial, phrase-final and isolated words. Comparing the variation in average relative duration in Figure 11.2 and average relative variation of F0 in Figure 11.5, the effect of syllable position in the word is similar in the cases of phrase-initial and phrase-medial words, but opposite in phrase-final words. Third, for intensity and relative F0 variation shown in Figures 11.3 and 11.5 respectively, opposite trends can be observed for phrase-initial words but similar trends for phrase-final words. In phrase-medial and isolated words the results are too irregular for valid conclusions. These qualitative comparisons are summarised in Table 11.7.

Finally, there are some general tendencies across all syllable and word positions. There is a regular increase in the relative duration of the tonic syllable, up to 200%. Less regular variation in intensity can be observed, moderately decreasing (2–3 dBs) as the word position varies from the beginning to the middle of the phrase, but increasing (2–4 dBs) phrase-finally and in isolated words. For F0 relative variation,

Table 11.7 Summary of qualitative trends for all word positions in the phrase

	Word position			
Character. quantity	Isolated	Beginning	Middle	End
Relative duration	↑	↑	↗	↑
Intensity	↓	↓	↓	↓
Relative F0 variation	↘*	↑	→	↘

Note: *Irregular variation.

the most significant tendency is a regular decrease from the beginning to the end of the phrase, but in isolated words the behaviour is irregular with an increase at the beginning of the word.

In informal listening tests of each individual characteristic in synthetic speech, the most important perceptual parameter is F0 and the least important is intensity. Duration and F0 are thus the most important parameters for a synthesiser.

Future Developments

This preliminary study clarified some important issues. In future studies the reference syllable should be similar to the tonic syllable for comparisons of duration and intensity values, and should be contiguous to the tonic in a neutral context. Consonant duration should also be controlled. These conditions are quite hard to fulfil in general, leading to the use of nonsense words containing the same syllable twice.

For duration and F0 variations a larger corpus of text is needed in order to increase the confidence levels. The default duration of each syllable should be determined and compared to the duration in tonic position. The F0 variation in the tonic syllable is assumed to be independent of segmental characteristics. The number and variety of speakers should also increase so that the results are more generally applicable.

Acknowledgements

The authors express their acknowledgement to COST 258 for the unique opportunities of exchange of experiences and knowledge in the field of speech synthesis.

References

Andrade, E. and Viana, M. (1988). Ainda sobre o ritmo e o Acento em Português. *Actas do 4° Encontro da Associação Portuguesa de Linguística*. Lisboa, 3–5.

Mateus, M., Andrade, A., Viana, M., and Villalva, A. (1990). *Fonética, Fonologia e Morfologia do Português*. Lisbon: Universidade Aberta.

Rabiner, L. and Schafer, R. (1978). *Digital Processing of Speech Signals*. Prentice-Hall.

Rowden, C. (1992). *Speech Processing*. McGraw-Hill.

Zellner, B. (1998). Caractérisation et prédiction du débit de parole en français. Unpublished doctoral thesis, University of Lausanne.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

12

Prosodic Parameters of Synthetic Czech

Developing Rules for Duration and Intensity

Marie Dohalská, Jana Mejvaldová and Tomáš Duběda
Institute of Phonetics, Charles University
Nám. J. Palacha 2, Praha 1, 116 38, Czech Republic
dohalska@ff.cuni.cz

Introduction

In our long-term research into the prosody of natural utterances at different speech rates (with special attention to the fast speech rate) we have observed some fundamental tendencies in the behaviour of duration (D) and intensity (I). A logical consequence of this was the incorporation of duration and intensity variations into our prosodic module for Czech synthesis, in which these two parameters had been largely ignored. The idea was to enrich the variations of fundamental frequency (F0), which had borne in essence the whole burden of prosodic changes, by adding D and I (Dohalská-Zichová and Duběda, 1996). Although we agree that fundamental frequency is the most important prosodic feature determining the acceptability of prosody (Bolinger, 1978), we claim that D and I also play a key role in the naturalness of synthetic Czech. A high-quality TTS system cannot be based on F0 changes alone. It has often been pointed out that the timing component cannot be of great importance in a language with a phonological length distinction like Czech (e.g. *dal* 'he gave' vs. *dál* 'further': the first vowel is short, the second long). However, we have found that apparently universal principles of duration (Maddieson, 1997) still apply to Czech (Palková, 1994).

We asked ourselves not only if the quality of synthetic speech is acceptable in terms of intelligibility, but we have also paid close 'phonetic' attention to its acceptability and aesthetic effect. Monotonous and unnatural synthesis with low prosodic variability might lead, on prolonged listening, to attention decrease in the listeners and to general fatigue.

Another problem is the fact that speech synthesis for handicapped people or in industrial systems has to meet special demands from the users. Thus, the speech rate may have to be very high (blind people use a rate up to 300% of normal) or

very low for extra intelligibility, which results in both segmental and prosodic distortions. At present, segments cannot be modified (except by shortening or lengthening), but prosody has to be studied for this specific goal. It is precisely in this situation, which involves many hours of listening, that monotonous prosody can have an adverse effect on the listener.

Methodology

The step-by-step procedure used to develop models of D and I was as follows:

1. Analysis of natural speech.
2. Application of the values obtained to synthetic sentences.
3. Manual adjustment.
4. Iterative testing of the acceptability of individual variants.
5. Follow-up correction according to the test results.
6. Selection of a general prosodic pattern for the given sentence type.

The modelling of synthetic speech was done with our ModProz software, which permits manual editing of prosodic parameters. In this system, the individual sounds are normalised in the domains of frequency (100 Hz), duration (average duration within a large corpus) and intensity (average). Modification involves adding or subtracting a percentage value.

The choice of evaluation material was not random. Initially, we concentrated on short sentences (5–6 syllables) of an informative character. All the sentences were studied in sets of three: statement, yes-no question, and wh-question (Dohalská *et al.*, 1998).

The selected sentences were modified by hand, based on measured data (natural sentences with the same wording pronounced by six speakers) and with an immediate feedback for the auditory effect of the different modifications, in order to obtain the most natural variant. We paid special attention to the interdependence of D and I, which turned out to be very complex. We focused on the behaviour of D and I at the beginnings and at the ends of stress groups with a varying number of syllables. The final fade-out at the end of an intonation group turned out to be of great importance.

Our analysis showed opposite tendencies of the two parameters at the end of rhythmic units. On the final syllable of a 2-syllable final unit, a rapid decrease in I was observed (down to 61% of the default value on average, but in many cases even 30–25%), while the D value rises to 138% of the default value for a short vowel, and to 370% for a long vowel on average. The distinction between short and long vowels is phonological in Czech. We used automatically generated F0 patterns which were kept constant throughout the experiment. Thus, the influence of D and I could be directly observed.

We are also aware of the role of timbre (sound quality), the most important segmental prosodic feature. However, our present synthesis system does not permit any variations of timbre, because the spectral characteristics of the diphones are fixed.

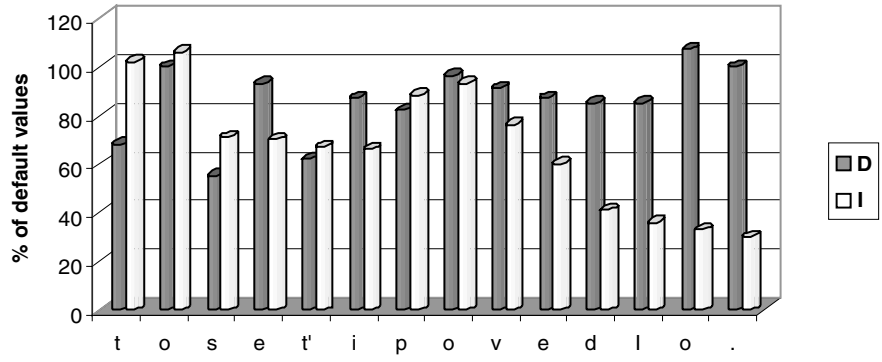


Figure 12.1 Manually adjusted D and I values in the sentence *To se ti povedlo*. (You pulled that off.) with high acceptability

An example of manually adjusted values is given in Figure 12.1. The sentence *To se ti povedlo* (You pulled that off) consists of two stress units with initial stress (*'To se ti / 'povedlo*).

Phonostylistic Variants

As well as investigating the just audible difference of D, I and F0 (Dohalská *et al.*, 1999) in various positions and contexts, we also tested the 'maximum acceptable' values of these two parameters per individual phonemes, especially at the end of the sentence (20 students and 5 teachers, comparison of two sentences in terms of acceptability). We tried to model different phonostylistic variants (Léon, 1992; Dohalská-Zichová and Mejvaldová, 1997) and to study the limit values of F0, D and I, as well as their interdependencies, without decreasing the acceptability too much. We found that F0 – considered often to be the dominant, if not the only phonostylistic factor – has to be accompanied by suitable variations of D and I. Some phonostylistic variants turned out to be dependent on timbre and they could not be modelled by F0, D and I.

We focused on a set of short everyday sentences, e.g. *What's the time?* or *You pulled that off*. Results for I are presented in Figure 12.2, as percentages of the default values for D and I. The maximum acceptable value for intensity (176%) was found on the initial syllable of a stress unit. This is not surprising, as Czech has regular stress on the first syllable of a stress unit. Figure 12.3 gives the results for D: a short vowel can reach 164% of default duration in the final position, but beyond this limit the acceptability falls. In all cases, the output sentences were judged to be different phonostylistic variants of the basic sentence. The phonostylistic colouring is due mainly to carefully adjusted variations of D and I, since we kept F0 as constant as possible.

We proceeded gradually from manual modelling to the formalisation of optimal values, in order to produce a set of typical values for D, I and F0 which were valid for a larger set of sentences. The parameters should thus represent a sort of compromise between the automatic prosody system and the prosody adjusted by hand.

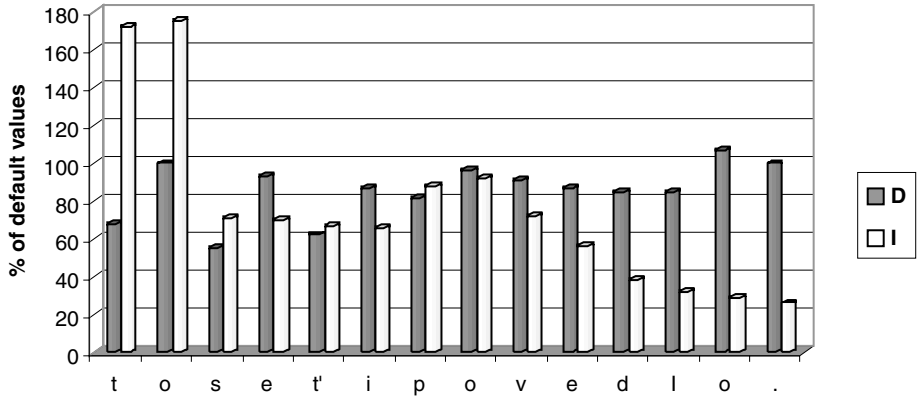


Figure 12.2

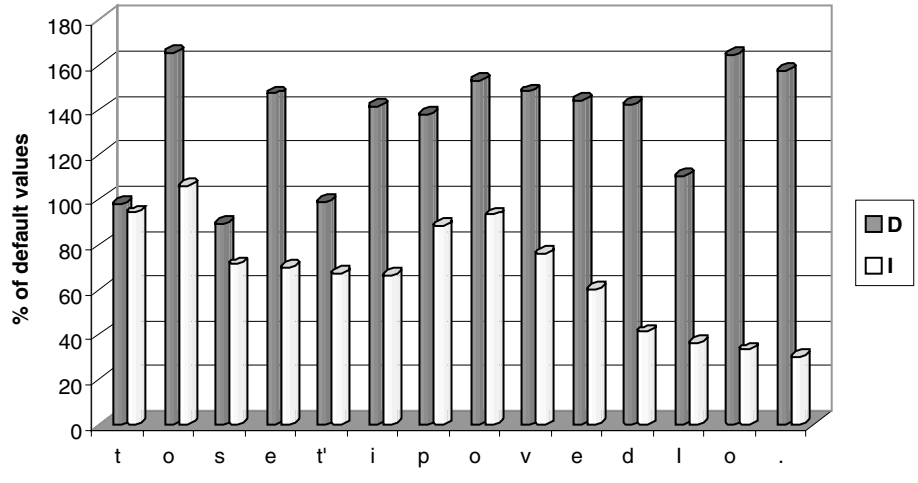


Figure 12.3

Implementation

To incorporate optimal values into the automatic synthesis program, we transplanted the modelled D and I curves onto other sentences with comparable rhythmic structure (with almost no changes to the default F0 values). We used not only declarative sentences, but also wh-questions and yes/no-questions. Naturally, the F0 curve had to be modified for interrogative sentences. The variations of I are independent of the type of sentence (declarative/interrogative), and seem to be general rhythmic characteristics of Czech, allowing us to use the same values for all sentence types.

The tendencies found in our previous tests with extreme values of D and I are valid also for neutral sentences (with neutral phonostylistic information). Highest intensity occurs on the initial, stress-bearing syllable of a stress unit, lowest intensity at the end of the unit. The same tendency is observed across a whole sentence, with the largest intensity drop in the final syllable. It should be noted that the decrease is greater (down to 25%) in an isolated sentence, while in continuous speech, the same decrease would sound unnatural or even comical.

We are currently formalising our observations with the help of new software christened *Epos* (Hanika and Horák, 1998). It was created to enable the user to construct sets of prosodic rules, and thus to formalise regularities in the data. The main advantage of this program is a user-friendly interface which permits rule editing via a formal language, without modifying the source code. While creating the rules, the user can choose from a large set of categories: position of the unit within a larger unit, nature of the unit, length of the unit, type of sentence, etc.

Acknowledgements

This research was supported by the COST 258 programme.

References

Bolinger, D. (1978). Intonation across languages. *Universals of Human Language* (pp. 471–524). Stanford.

Dohalská, M., Duběda, T., Mejvaldová, J. (1998). Preception limits between assertive and interrogative sentences in Czech. *8th Czech German Workshop, Speech processing* (pp. 28–31). Praha.

Dohalská, M., Duběda, T., and Mejvaldová, J. (1999). Perception of synthetic sentences with indistinct intonation in Czech. *Proceedings of the International Congress of Phonetic Sciences* (pp. 2335–2338). San Francisco.

Dohalská, M. and Mejvaldová, J. (1998). Les critères prosodiques des trois principaux types de phrases (testés sur le tchèque synthétique). XXIIèmes *Journées d'Etude sur la Parole* (pp. 103–106). Martigny.

Dohalská-Zichová, M. and Duběda, T. (1996). Rôle des changements de la durée et de l'intensité dans la synthèse du tchèque. XXIèmes *Journées d'Etude sur la Parole* (pp. 375–378). Avignon.

Dohalská-Zichová, M. and Mejvaldová, J. (1997). Où sont les limites phonostylistiques du tchèque synthétique? *Actes du XVIe Congrès International des Linguistes*. Paris.

Hanika, J. and Horák, P. (1998). Epos – a new approach to the speech synthesis. *Proceedings of the First Workshop on Text, Speech and Dialogue* (pp. 51–54). Brno.

Léon, P. (1992). *Précis de phonostylistique: Parole et expressivité*. Nathan.

Maddieson, I. (1997). Phonetic universals. In W.J. Hardcastle and J. Laver, *The Handbook of Phonetic Sciences* (pp. 619–639). Blackwell Publishers.

Palková, Z. (1994). *Fonetika a fonologie češtiny*. Karolinum.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

13

MFGI, a Linguistically Motivated Quantitative Model of German Prosody

Hansjörg Mixdorff
Dresden University of Technology, 01062 Dresden, Germany
mixdorff@tfh-berlin.de

Introduction

The intellegibility and perceived naturalness of synthetic speech strongly depend on the prosodic quality of a TTS system. Although some recent systems avoid this problem by concatenating larger chunks of speech from a database (see, for instance, Stöber *et al*., 1999), an approach which preserves the natural prosodic structure at least throughout the chunks chosen, the question of optimal unit selection still calls for the development of improved prosodic models. Furthermore, the lack of prosodic naturalness of conventional TTS systems indicates that the production process of prosody and the interrelation between the prosodic features of speech is still far from being fully understood.

Earlier work by the author was dedicated to a model of German intonation which uses the well-known quantitative Fujisaki model of the production process of F0 (Fujisaki and Hirose, 1984) for parameterising F0 contours, the **M**ixdorff-**F**ujisaki Model of **G**erman **I**ntonation (short MFGI). In the framework of MFGI, a given F0 contour is described as a sequence of linguistically motivated tone switches, major rises and falls, which are modelled by onsets and offsets of accent commands connected to accented syllables, or by so-called *boundary tones*. Prosodic phrases correspond to the portion of the F0 contour between consecutive phrase commands (Mixdorff, 1998). MFGI was integrated into the TU Dresden TTS system DRESS (Hirschfeld, 1996) and produced high naturalness compared with other approaches (Mixdorff and Mehnert, 1999).

Perception experiments, however, indicated flaws in the duration component of the synthesis system and gave rise to the question of how intonation and duration models should interact in order to achieve the highest prosodic naturalness possible. Most conventional systems like DRESS employ separate modules for gener-

ating F0 and segment durations. These modules are often developed independently and use features derived from different data sources and environments. This ignores the fact that the natural speech signal is coherent in the sense that intonation and speech rhythm are co-occurrent and hence strongly correlated. As part of his post-doctoral thesis, the author of this chapter decided to develop a prosodic module which takes into account the relation between melodic and rhythmic properties of speech. The model is henceforth to be called an 'integrated prosodic model'. For its F0 part this integrated prosodic model still relies on the Fujisaki model which is combined with a duration component. Since the Fujisaki model proper is language independent, constraints must be defined for its application to German. These constraints, which differ from the implementation by Möbius *et al.* (1993), for instance, are based on earlier works on German intonation discussed in the following section.

Linguistic Background of MFGI

The early work by Isačenko (Isačenko and Schädlich, 1964) is based on perception experiments using synthesised stimuli with extremely simplified F0 contours. These were designed to verify the hypothesis that the syntactic functions of German intonation can be modelled using tone switches between two constant F0 values connected to accented, so-called *ictic* syllables and *pitch interrupters* at syntactic boundaries.

The stimuli were created by 'monotonising' natural utterances at two constant frequencies and splicing the corresponding tapes at the locations of the tone switches (see Figure 13.1 for an example). The experiments showed a high consistency in the perception of intended syntactic functions in a large number of subjects.

The tutorial on German sentence intonation by Stock and Zacharias (1982) further develops the concept of tone switches introduced by Isačenko. They propose phonologically distinctive elements of intonation called *intonemes* which are characterised by the occurrence of a tone switch at an accented syllable. Depending on their communicative function, the following classes of intonemes are distinguished:

- *Information intoneme I* ↓ Declarative-final accents, falling tone switch. Conveying a message.
- *Contact intoneme C* ↑ Question-final accents, rising tone switch. Establishing contact.
- *Non-terminal intoneme N* ↑ Non-final accents, rising tone switch. Signalling non-finality.

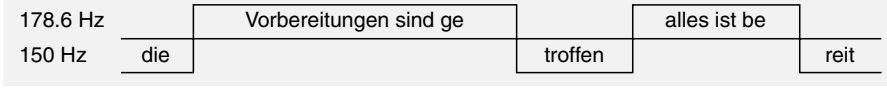


Figure 13.1 Illustration of the splicing technique used by Isačenko. Every stimulus is composed of chunks of speech monotonized either at 150 or 178.6 Hz

Any intonation model for TTS requires information about the appropriate accentuation and segmentation of an input text. In this respect, Stock and Zacharias' work is extremely informative as it provides default accentuation rules (word accent, phrase and sentence accents), and rules for the prosodic segmentation of sentences into *accent groups*.

The Fujisaki Model

The mathematical formulation used in MFGI for parameterising F0 contours is the well-known Fujisaki model. Figure 13.2 displays a block diagram of the model which has been shown to be capable of producing close approximations to a given contour from two kinds of input commands: phrase commands (impulses) and accent commands (stepwise functions). These are described by the following model parameters (henceforth referred to as *Fujisaki parameters*): *Ap*: phrase command magnitude; *T0*: phrase command onset time; α: time constant of phrase command; *Aa*: accent command amplitude; *T1*: accent command onset time; *T2*: accent command offset time; β: time constant of accent command; *Fb*, the 'base frequency', denoting the speaker-dependent asymptotic value of F0 in the absence of accent commands.

The phrase component produced by the phrase commands accounts for the global shape of the F0 contour and corresponds to the declination line. The accent commands determine the local shape of the F0 contour, and are connected to accents. The main attraction of the Fujisaki model is the physiological interpretation which it offers for connecting F0 movements with the dynamics of the larynx (Fujisaki, 1988), a viewpoint not inherent in other current intonation models which mainly aim at breaking down a given F0 contour into a sequence of 'shapes' (e.g. Taylor, 1995; Portele *et al.*, 1995).

MFGI's Components

Following Isačenko and Stock, an F0 contour in German can be adequately described as a sequence of tone switches. These tone switches can be regarded as basic

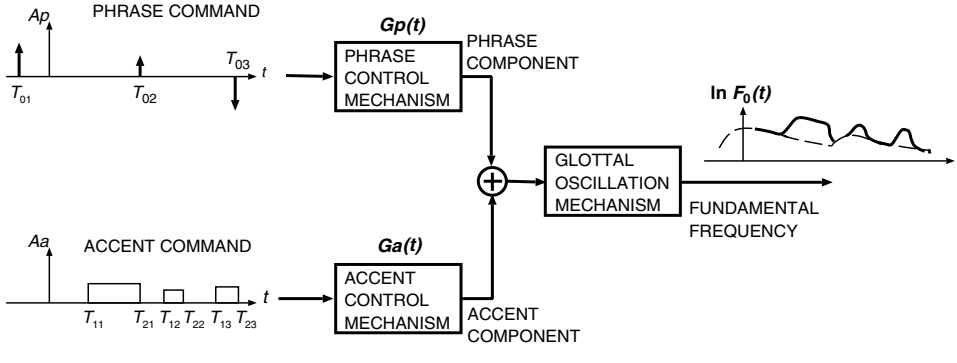


Figure 13.2 Block diagram of the Fujisaki model (Fujisaki and Hirose, 1984)

intonational elements. The term intoneme proposed by Stock shall be adopted to classify those elements that feature tone switches on accented syllables. Analogously with the term *phoneme* on the segmental level, the term intoneme describes intonational units that are quasi-discrete and denote phonological contrasts in a language. Although the domain of an intoneme may cover a large portion of the F0 contour, its characteristic feature, the tone switch, can be seen as a discrete event. By means of the Fujisaki model, intonemes can be described not only qualitatively but quantitatively, namely by the timing and amplitude of the accent commands to which they are connected. Analysis of natural F0 contours (Mixdorff, 1998) indicated that further elements – not necessarily connected to accented syllables – are needed. These occur at prosodic boundaries, and will be called *boundary tones* (marked by $B\uparrow$) using a term proposed by Pierrehumbert (1980).

Further discussion is needed as to how the portions of the F0 contour pertaining to a particular intoneme can be delimited. In an acoustic approach, for instance, an intoneme could be defined as starting with its characteristic tone switch and extending until the characteristic tone switch of the following accented syllable. In the present approach, however, a division of the F0 contour into portions belonging to meaningful units (words or groups of words) is favoured, as the location of accented syllables is highly dependent on constituency, i.e. the choice of words in an utterance and the location of their respective word accent syllables. Unlike other languages, German has a vast variety of possible word accent locations for words with the same number of syllables. Hence the delimitation of intonemes is strongly influenced by the lexical and syntactic properties of a particular utterance. We therefore follow the notion of *accent group* as defined by Stock, namely the grouping of clitics around an accented word as in the following example: 'Ich s'ah ihn // mit dem F'ahrrad // über die Br'ücke fahren' ('I saw him ride his bike across the bridge') where 'denotes accented syllables and // denotes accent group boundaries.

Analysis of natural F0 contours showed that every utterance starts with a phrase command, and major prosodic boundaries in utterance-medial positions are usually linked with further commands. Hence, the term *prosodic phrase* denotes the part of an utterance between two consecutive phrase commands. It should be noted that since the phrase component possesses a finite time constant, a phrase command usually occurs shortly before the segmental onset of a prosodic phrase, typically a few hundred ms. The phrase component of the Fujisaki model is interpreted as a *declination component* from which rising tone switches depart and to which falling tone switches return.

Speech Material and Method of Analysis

In its first implementation, for generating Fujisaki parameters from text, MFGI relied on a set of rules (Mixdorff, 1998, p. 238 ff.). These were developed based on the analysis of a corpus which was not sufficiently large for employing statistical methods, such as neural networks or CART trees for predicting model parameters. For this reason, most recently a larger speech database was analysed in order to determine the statistically relevant predictor variables for the integrated

prosodic model. The corpus is part of a German corpus compiled by the Institute of Natural Language Processing, University of Stuttgart and consists of 48 minutes of news stories read by a male speaker (Rapp, 1998). The decision to use this database was taken for several reasons: The data is real-life material and covers unrestricted informative texts produced by a professional speaker in a neutral manner. This speech material appears to be a good basis for deriving prosodic features for a TTS system which in many applications serves as a reading machine.

The corpus contains boundary labels on the phone, syllable and word levels and linguistic annotations such as part-of-speech. Furthermore, prosodic labels following the Stuttgart G-ToBI system (Mayer, 1995) are provided. The Fujisaki parameters were extracted using a novel automatic multi-stage approach (Mixdorff, 2000). This method follows the philosophy that not all parts of the F0 contour are equally salient, but are 'highlighted' to a varying degree by the underlying segmental context. Hence F0 modelling in those parts pertaining to accented syllable nuclei (the locations of tone switches) needs to be more accurate than along low-energy voiced consonants in unstressed syllables, for instance.

Results

Figure 13.3 displays an example of analysis, showing from top to bottom: the speech waveform, the extracted and model-generated F0 contours, the ToBI tier, the text of the utterance, and the underlying phrase and accent commands.

Accentuation

The corpus contains a total number of 13 151 syllables. For these a total number of 2931 accent commands were computed. Of these 2400 are aligned with syllables labelled as accented. Some 177 unaccented syllables preceding prosodic boundaries exhibit an accent command corresponding to a boundary tone $B\uparrow$. A rather small number of 90 accent commands are aligned with accented syllables on their rising as well as on their falling slopes, hence forming *hat patterns*.

Alignment

The *information intoneme* $I\downarrow$, and the *non-terminal intoneme* $N\uparrow$ can be reliably identified by the alignment of the accent command with respect to the accented syllable, expressed as $T1_{dist} = T1 - t_{on}$; and $T2_{dist} = T2 - t_{off}$ where t_{on} denotes the syllable onset time and t_{off} the syllable offset time. Mean values of $T1_{dist}$ and $T2_{dist}$ for I-intonemes are -47.5 ms and -47.1 ms compared with 56.0 ms and 78.4 ms for N-intonemes. N-intonemes preceding a prosodic boundary exhibit additional offset delay (mean $T2_{dist} = 125.5$ ms). This indicates that in these cases, the accent command offset is shifted towards the prosodic boundary.

A considerable number of accented syllables (N = 444) was detected which had not been assigned any accent labels by the human labeller. Figure 13.3 shows such an instance where in the utterance 'Die fran'zösische Re'gierung hat in einem

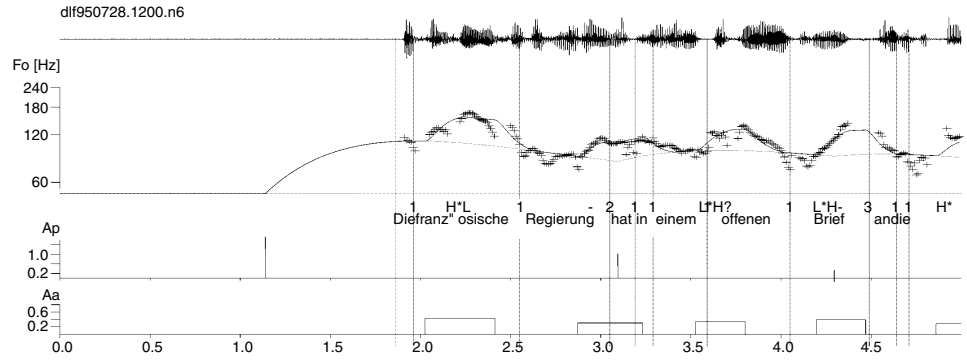


Figure 13.3 Initial part of an utterance from the database. The figure displays from top to bottom: (1) the speech waveform, (2) the extracted (+−signs) and estimated (solid line) F0 contours, (3) the ToBI labels and text of utterance, (4) the underlying phrase commands (impulses) and accent commands (steps)

'offenen 'Brief…' ('In an 'open 'letter, the 'French 'government…'), an accent command was assigned to the word 'Re'gierung', but not a tone label. Other cases of unlabelled accents were lexically stressed syllables in function words, which are usually unaccentable.

Prominence

Table 13.1 shows the relative frequency of accentuation depending on the part-of-speech of the word. As expected, nouns and proper names are accented more frequently than verbs, which occupy a middle position in the hierarchy, whereas function words such as articles and prepositions are very seldom accented. For the categories that are frequently accented, the right-most column lists a mean *Aa* reflecting some degree of relative prominence depending on the part of speech. As

Table 13.1 Occurrence, frequency of accentuation and mean *Aa* for selected parts of speech

Part of speech	Occurrence	Accented %	Mean *Aa*
Nouns	1262	75.8	0.28
Proper names	311	78.4	0.32
Adjectives conjugated	333	71.6	0.25
Adjectives non-conjugated	97	85.7	0.28
Past participle of full verbs	172	77.3	0.29
Finite full verbs	227	42.7	0.30
Adverbs	279	41.9	0.29
Conjunctions	115	2.6	–
Finite auxiliary verb	219	3.0	–
Articles	804	1.0	–
Prepositions	621	2.0	–

can be seen, differences found in these mean values are small. As shown in Wolters & Mixdorff (2000), word prominence is more strongly influenced by the syntactic relationship between words than simply by parts-of-speech.

A very strong factor influencing the *Aa* assigned to a certain word is whether it precedes a deep prosodic boundaries. Pre-boundary accents and boundary tones exhibit a mean *Aa* of 0.34 against 0.25 for phrase-initial and -medial accents.

Phrasing

All inter-sentence boundaries were found to be aligned with the onset of a phrase command. Some 68% of all intra-sentence boundaries exhibit a phrase command, with the figure rising to 71% for 'comma boundaries'. The mean phrase command magnitude *Ap* for intra-sentence boundaries, inter-sentence boundaries and paragraph onsets is 0.8, 1.68, and 2.28 respectively, which shows that *Ap* is a useful indicator of boundary strength. In Figure 13.4 the phrase component extracted for a complete news paragraph is displayed: sentence onsets are marked with arrows. As can be seen, the magnitudes of the underlying phrase commands nicely reflect the phrasal structure of the paragraph.

About 80% of prosodic phrases in this data contain 13 syllables or less. Hence phrases in the news utterances examined are considerably longer than the corresponding figure of eight syllables found in Mixdorff (1998) for simple readings. This effect may be explained by the higher complexity of the underlying texts, but also by the better performance of the professional announcer.

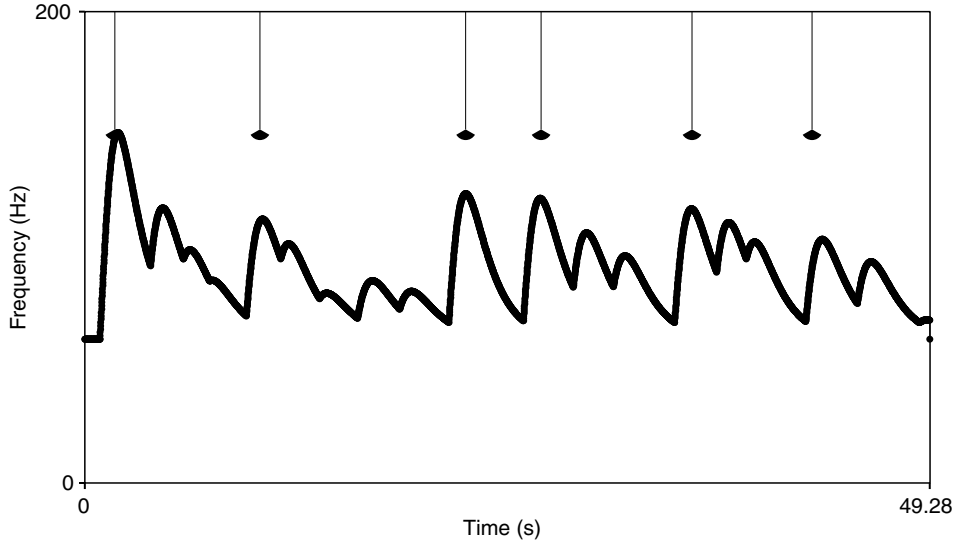


Figure 13.4 Profile of the phrase component underlying a complete news paragraph. Sentence onsets are marked with vertical arrows

A Model of Syllable Duration

In order to align an F0 contour with the underlying segmental string, F0 model parameters need to be related to the timing grid of an utterance. As was shown for the timing of intonemes in the preceding section, the syllable appears to be an appropriate temporal unit for 'hooking up' F0 movements pertaining to accents.

The timing of tone switches can thus be expressed by relating *T1* and *T2* to syllable onset and offset times respectively. In a similar fashion, the phrase command onset time *T0* can be related to the onset time of the first syllable in the corresponding phrase, namely by the distance between *T0* and the segmental onset of the phrase.

A regression model of the syllable duration was hence developed which separates the duration contour into an *intrinsic* part related to the (phonetic) syllable structure and a second, *extrinsic* part related to linguistic factors such as accentuation and boundary influences. The largest extrinsic factors were found to be (1) the degree of accentuation (with the categories 0: 'unstressed', 1: 'stressed, but unaccented', 2: 'accented', where 'accented' denotes a syllable that bears a tone switch); and (2) the strength of the prosodic boundary to the right of a syllable, together accounting for 35% of the variation in syllable duration. Pre-boundary lengthening is therefore reflected by local maxima of the extrinsic contour. The number of phones – as could be expected – proves to be the most important intrinsic factor, followed by the type of the nuclear vowel (the reduction-prone schwa or non-schwa). These two features alone account for 36% of the variation explained. Figure 13.5 shows an example of a

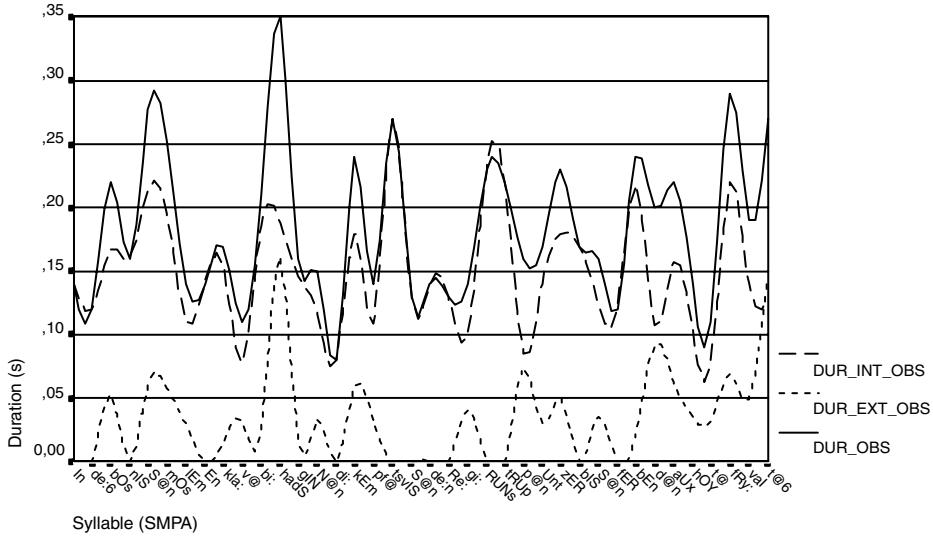


Figure 13.5 Example of smoothed syllable duration contours for the utterance 'In der bosnischen Moslem-Enklave Bihac gingen die Kämpfe zwischen den Regierungstruppen und serbischen Verbänden auch heute früh weiter' ('*In the Bosnian Muslim-enclave of Bihac, fights between the government troops and Serbian formations still continued this morning*'). The solid line indicates measured syllable duration, the dashed line intrinsic syllable duration and the dotted line extrinsic syllable duration. At the bottom, the syllabic SMPA-transcription is displayed.

smoothed syllable duration contour (solid line) decomposed into intrinsic (dotted line) and extrinsic (dashed line) components.

Compared with other duration models, the model presented here still incurs a considerable prediction error as it yields a correlation of only 0.79 between observed and predicted syllable durations (compare 0.85 in Zellner Keller (1998) for instance). Possible reasons for this shortcoming include the following:

- the duration model is not hierarchical, as factors from several temporal domains (i.e. phonemic, syllabic and phrasal) are superimposed on the syllabic level, and the detailed phone structure is (not yet) taken into account;
- syllabification and transcription information in the database are often erroneous, especially for foreign names and infrequent compound words which were not transcribed using a phonetic dictionary, but by applying default grapheme-to-phoneme rules.

Conclusion

This chapter discussed the linguistically motivated prosody model MFGI which was recently applied to a large prosodically labelled database. It was shown that model parameters can be readily related to the linguistic information underlying an utterance. Accent commands are typically aligned with accented syllables or syllables bearing boundary tones. Higher level boundaries are marked by the onset of phrase commands whereas the detection of lower-level boundaries obviously requires the evaluation of durational factors. For this purpose a syllable duration model was introduced. As well as the improvement of the syllable duration model, work is in progress to combine intonation and duration models into an integrated prosodic model.

References

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In O. Fujimura (ed.), *Vocal Physiology: Voice Production, Mechanisms and Functions* (pp. 347–355). Raven Press Ltd.

Fujisaki, H. and Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E), 5(4)*, 233–241.

Hirschfeld, D. (1996). The Dresden text-to-speech system. *Proceedings of the 6th Czech-German Workshop on Speech Processing* (pp. 22–24). Prague, Czech Republic.

Isačenko, A. and Schädlich, H. (1964). *Untersuchungen über die deutsche Satzintonation*. Akademie-Verlag.

Mayer, J. (1995). *Transcription of German Intonation: The Stuttgart System. Technischer Bericht*, Institut für Maschinelle Sprachverarbeitung. Stuttgart-University.

Mixdorff, H. (1998). *Intonation Patterns of German – Model-Based Quantitative Analysis and Synthesis of F0 Contours*. PhD thesis TU Dresden (http://www.tfh-berlin.de/~mixdorff/thesis.htm).

Mixdorff, H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. *Proceedings ICASSP 2000*, Vol. 3 (pp. 1281–1284). Istanbul, Turkey.

Mixdorff, H. & Mehnert, D. (1999). Exploring the naturalness of several German high-quality-text-to-speech systems. *Proceedings of Eurospeech '99*, Vol. 4 (pp 1859–1862). Budapest, Hungary.

Möbius, B., Pätzold, M., and Hess, W. (1993). Analysis and synthesis of German F0 contours by means of Fujisaki's model. *Speech Communication*, *13*, 53–61.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT.

Portele, T., Krämer, J., and Heuft, B. (1995). Parametrisierung von Grundfrequenzkonturen. *Fortschritte der Akustik – DAGA '95* (pp. 991–994). Saarbrücken.

Rapp, S. (1998). *Automatisierte Erstellung von Korpora für die Prosodieforschung*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Stuttgart University.

Stöber, K., Portele, T., Wagner, P., and Hess, W. (1999). Synthesis by word concatenation. *Proceedings of EUROSPEECH '99*., Vol. 2 (pp. 619–622). Budapest.

Stock, E. and Zacharias, C. (1982). *Deutsche Satzintonation*. VEB Verlag Enzyklopädie.

Taylor, P. (1995). The rise/fall/connection model of intonation. *Speech Communication*, *15(1)*, 169–186.

Wolters, M. and Mixdorff, H. (2000). Evaluating radio news intonation: Autosegmental vs. superpositional modeling. *Proceedings of ICSLP 2000*. Vol. 1 (pp. 584–585) Beijing, China.

Zellner Keller, B. (1998). Prediction of temporal structure for various speech rates. In N. Campbell (ed.), *Volume on Speech Synthesis*. Springer-Verlag.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

14

Improvements in Modelling the F0 Contour for Different Types of Intonation Units in Slovene

Ales Dobnikar
Institute J. Stefan, SI-1000 Ljubljana, Slovenia
ales.dobnikar@ijs.si

Introduction

This chapter presents a scheme for modelling the F0 contour for different types of intonation units for the Slovene language. It is based on results of analysing F0 contours, using a quantitative model on a large speech corpus. The lack of previous research into Slovene prosody for the purpose of text-to-speech synthesis meant that an approach had to be chosen and rules had to be developed from scratch. The F0 contour generated for a given utterance is defined as the sum of a global component, related to the whole intonation unit, and local components related to accented syllables.

Speech Corpus and F0 Analyses

Data from ten speakers were collected, resulting in a large corpus. All speakers were professional Slovene speakers on national radio, five males (labelled M1–M5) and five females (labelled F1–F5). The largest part of the speech material consists of declarative sentences, in short stories, monologues, news, weather reports and commercial announcements, containing sentences of various types and complexities (speakers M1–M4 and F1–F4). This speech database contains largely neutral prosodic emphasis and aims to be maximally intelligible and informative. Other parts of the corpora are interrogative sentences with yes/no and wh-questions and imperative sentences (speakers M5 and F5).

In the model presented here, an intonation unit is defined as any speech between two pauses greater than 30 ms. Shorter pauses were not taken as intonation unit

Table 14.1 No. of intonation units and total duration for each speaker in the corpus

Label	No. of intonation units	Length
F1	71	172.3
F2	34	102.3
F3	39	98
F4	64	146.6
F5	51	97.5
M1	33	91.5
M2	38	101.1
M3	45	75.9
M4	64	151.9
M5	51	93.3

boundaries, because this length is the minimum value for the duration of Slovene phonemes. Table 14.1 shows the speakers, the number of intonation units and the total duration of intonation units.

The scheme for modelling F0 contours is based on the results of analysing F0 contours using the INTSINT system (Hirst *et al.*, 1993; Hirst and Espesser, 1994; Hirst, 1994; Hirst and Di Cristo, 1995), which incorporates some ideas from TOBI transcription (Silverman *et al.*, 1992; Llisterri, 1994). The analysis algorithm uses a spline fitting approach that reduces F0 to a number of target points. The F0 contour is built up by interpolation between these points. The target points can then be automatically coded into INTSINT symbols, but the orthographic transcription of the intonation units or boundaries must be manually introduced and aligned with the target points.

Duration of Pauses

Pauses have a very important role in the intelligibility of speech. In normal conversations, typically half of the time consists of pauses; in the analysed readings they represent 18% of the total duration. The results show that pause duration is independent of the duration of the intonation unit before the pause. Pause duration depends only on whether the speaker breathes in during the pause.

Pauses, the standard boundary markers between successive intonation units, are classified into five groups with respect to types and durations:

- at new topics and new paragraphs, not marked in the orthography; these always represent the longest pauses, and always include breathing in;
- at the end of sentences, marked with a period, exclamation mark, question mark or dots;
- at prosodic phrase boundaries within the sentences, marked by comma, semicolon, colon, dash, parentheses or quotation marks;

- at rhythmic boundaries within the clause, often before the conjunctions *in*, *ter* (and), *pa* (but), *ali* (or), etc.
- at places of increased attention to a word or group of words.

Taking into account the fact that pause durations vary greatly across different speaking styles, the median was taken as a typical value because the mean is affected by extreme values which occur for different reasons (physical and emotional states of the speaker, style, attitude, etc.). The durations proposed for pauses are therefore in the range between the first and the third quartile, located around the median, and are presented in Table 14.2. This stochastic variation in pause durations avoids the unnatural, predictable nature of pauses in synthetic speech.

Modelling the F0 Contour

The generation of intonation curves for various types of intonation in the speech synthesis process consists of two main phases:

- segmentation of the text into intonation units;
- definition of the F0 contour for specific intonation units.

For automatic generation of fundamental frequency patterns in synthetic speech, a number of different techniques have been developed in recent years. They may be classified into two broad categories. One is the so-called 'superpositional approach', which regards an F0 contour as consisting of two or more superimposed components (Fujisaki, 1993; Fujisaki and Ohno, 1993), and the other is termed the 'linear approach' because it regards an F0 contour as a linear succession of tones, each corresponding to a local specification of F0 (Pierrehumbert, 1980; Ladd, 1987; Monaghan, 1991). For speech synthesis the first approach is more common, where the generated F0 contour of an utterance is the sum of a global component, related to the whole intonation unit, and local components related to accented syllables (see Figure 14.1).

Table 14.2 Pause durations proposed for Slovene

Type of pauses	Orthographic delimiters	Durations [ms]
At prefaces, between paragraphs, new topics of readings, ...		1430–1830
At the end of clauses	'.' '...' '?' '!'	780–1090
At places of prosodic phrases inside clauses	',' ';' ':' '-' '(...)' '"..."'	100–180; $t_m < 2.3$ s 400–440; $t_m \geq 2.3$ s
At places of rhythmical division of some clauses	before the Slovene conj. words *in*, *ter* (and), *pa* (but), *ali* (or)...	100–130; $t_m < 2.9$ s 360–390; $t_m \geq 2.9$ s
At places of increased attention to some word or part of the text	no classical orthographic delimiters	60–70

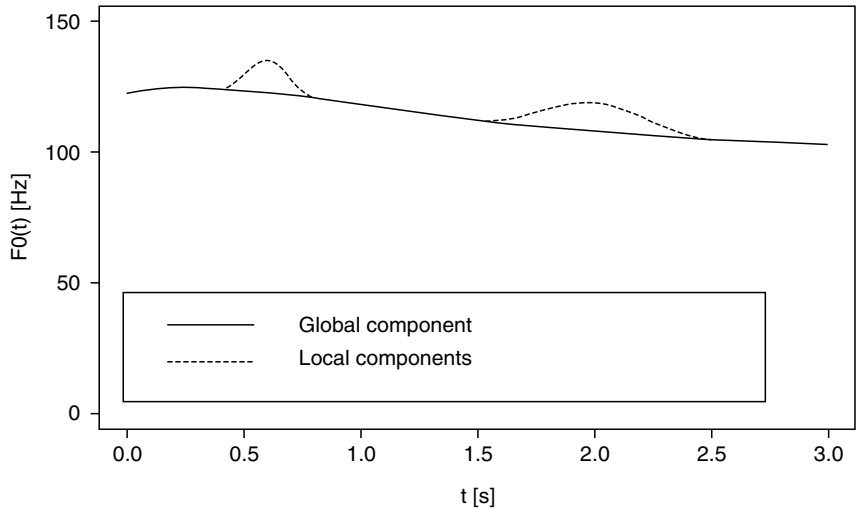


Figure 14.1 Definition of the F0 contour as the sum of global and local components

The global component gives the baseline F0 contour for the whole intonation unit, and often rises at the beginning of the intonation unit and slightly decreases towards the end. It depends on:

- the type of intonation unit (declarative, imperative, yes/no or wh-question);
- the position of the intonation unit (initial, medial, final) in a complex sentence with two or more intonation units;
- the duration of the whole intonation unit.

The local components model movements of F0 on accented syllables:

- the rise and fall of F0 on accented syllables in the middle of the intonation unit;
- the rise of F0 at the end of the intonation unit, if the last syllable is accented;
- the fall of F0 at the beginning of the intonation unit, if the first syllable is accented.

The F0 contour is defined by a function, composed of global $G(t)$ and local $L_i(t)$ components (Dobnikar, 1996; 1997):

$$F0 = G(t) + \sum_i L_i(t) \tag{1}$$

For the approximation of the global component an exponential function was adopted:

$$G(t) = F_k e^{A_z \alpha (t+0.5) e^{-\alpha (t+0.5)}} \tag{2}$$

and a cosine function for local components:

$$L_i(t) = G(T_{pi})A_{pi}(1 + \cos(\pi \frac{T_{pi} - t}{d_i})) \quad (3)$$

where the expression $(T_{pi} - t)$ must be in the range $(-d_i, d_i)$, otherwise $L_i(t) = 0$.

The symbols in these equations denote:

F_k = asymptotic final value of F0 in the intonation unit
A_z = parameter for the onset F0 value in the intonation unit
α = parameter for F0 shape control
T_{pi} = time of the i-th accent
A_{pi} = magnitude of the i-th accent
d_i = duration of the i-th accent contour

The parameters are modified during the synthesis process depending on syntactico-semantic analysis, speaking rate and microprosodic parameters. The values of global component parameters in the generation process (F_k, A_z, α) therefore depend on the relative height of the synthesised speech register, the type and position of intonation units in complex clauses, and the duration of the intonation unit.

F_k is modified according to the following heuristics (see Figure 14.2):

- If the clause is an independent intonation unit, then F_k could be the average final value of synthesised speech or the average final values obtained in analysed speech corpus ($F_k = 149$ Hz for female and $F_k = 83$ for male speech).
- If the clause is constructed with two or more intonation units, then:
  - the F_k value of the first intonation unit is the average final intonation unit multiplied by 1.075;
  - the F_k value of the last intonation unit is the average final intonation unit multiplied by 0.89;
  - the middle intonation unit(s), if any exist, have for F_k value defined average final values F_k.

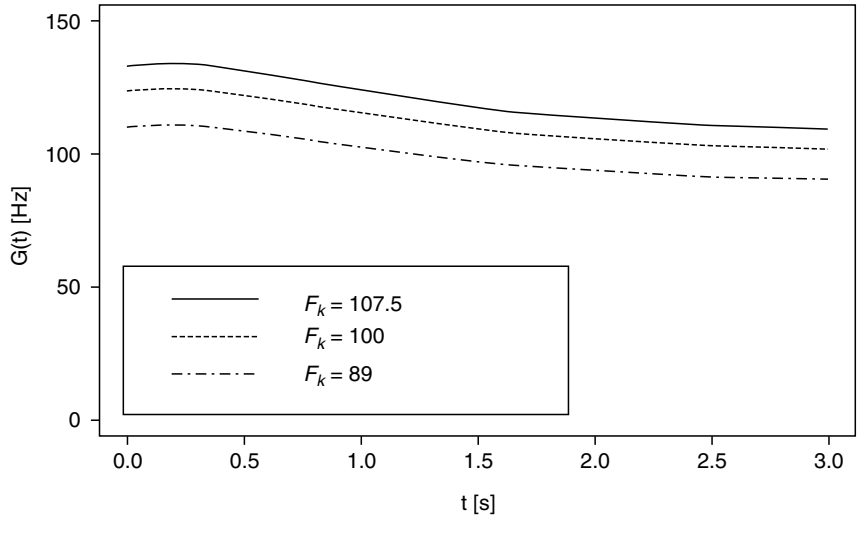


Figure 14.2 Influence of F_k values on the global component G(t)

The value of A_z (onset F0) depends on the type and position of the intonation unit in a complex sentence with two or more intonation units in the same clause. Figure 14.3 illustrates the influence of A_z on the global component.

Analysis revealed that in all types of intonation unit in Slovene readings, a falling baseline with positive values of A_z is the norm (Table 14.3).

The parameter α, dependent on the overall duration of the intonation unit T, specifies the global F0 contour and slope (Figure 14.4) and is defined as:

$$\alpha = 1 + \frac{4}{\sqrt{(T+1)^3}} \tag{4}$$

Parameter values for local components depend on the position (T_{pi}), height (A_{pi}, see Figure 14.5) and duration (d_i, see Figure 14.6) of the i-th accent in the intonation unit. Most of the primary accents in the analysed speech corpus occur at the

Table 14.3 Values for A_z for different types of intonation unit

Type of intonation and position of intonation unit		A_z
Declarative	independent intonation unit or starting intonation unit in a complex clause	0.47
Declarative	last intonation unit in a complex clause	0.77
Wh-question		1
YES/NO question		0.23
Imperative		0.7

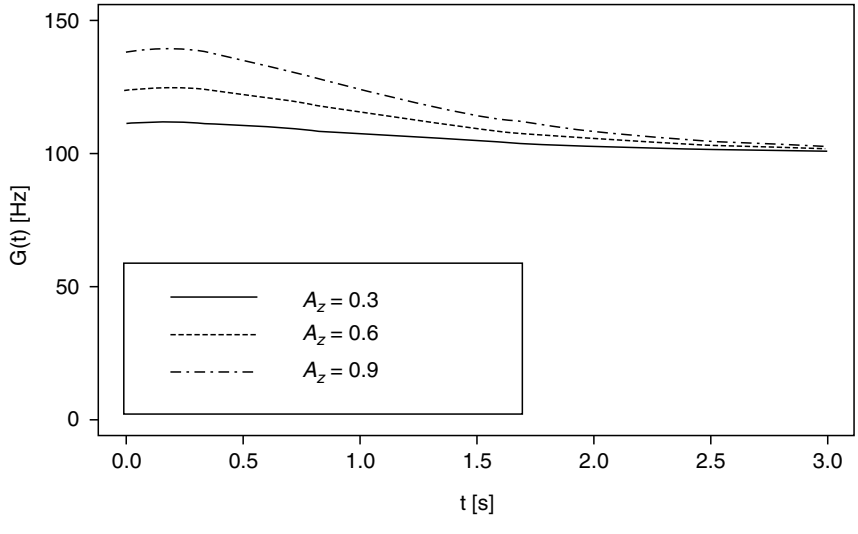


Figure 14.3 Influence of A_z values on the global component G(t)

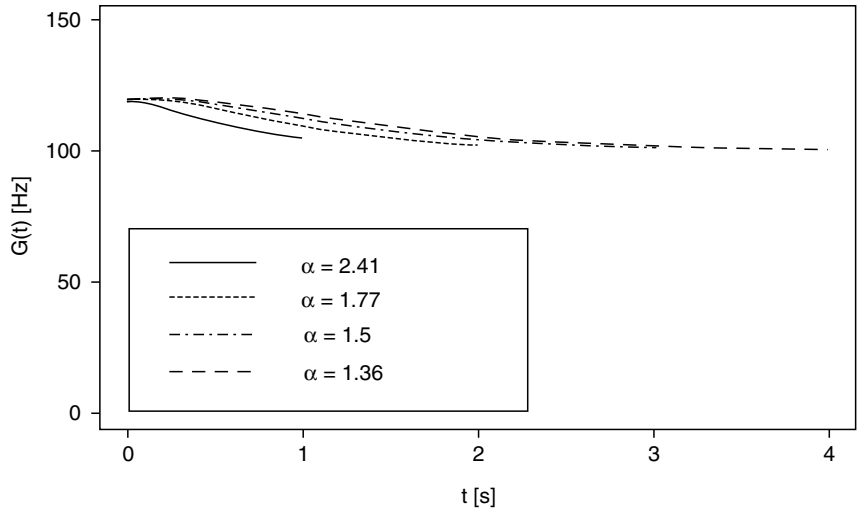


Figure 14.4 Influence of parameter α on the global component G(t)

beginning of intonation units (63%); others occur in the middle (16%) and at the end (21%). Comparison of the average values of F0 peaks at accents shows that these values are independent of the values of the global component and are dependent solely on the level of accentuation (primary or secondary accent). Exact values for local components are defined in the high-level modules of the synthesis system according to syntactic-semantic analysis, speaking rate and microprosodic parameters.

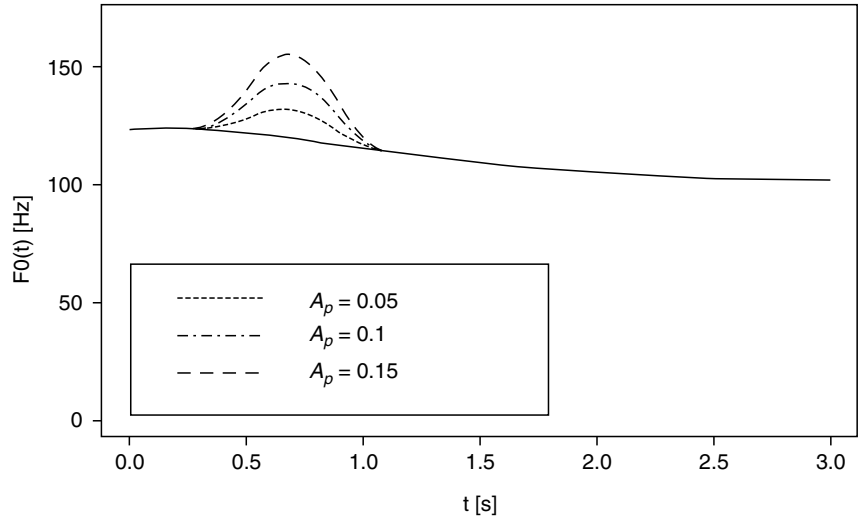


Figure 14.5 Influence of A_p on the local component L(t)

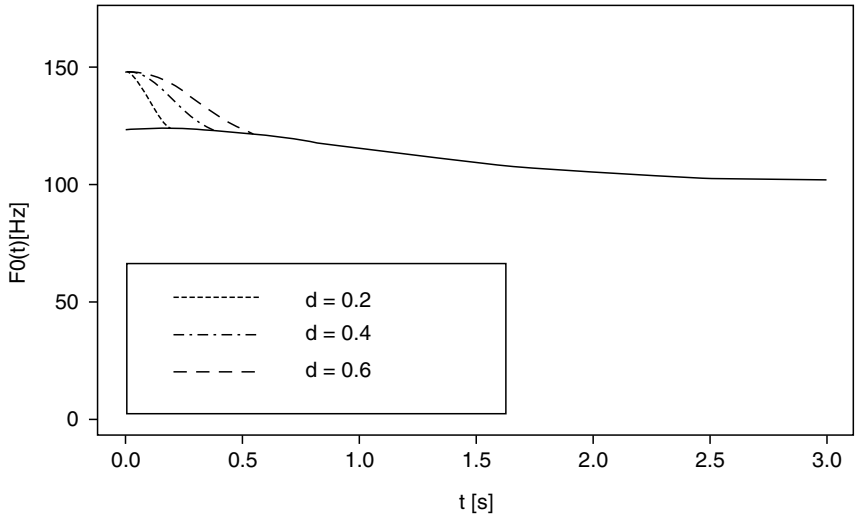


Figure 14.6 Influence of parameter d on the local component L(t)

Results

Figures 14.7, 14.8 and 14.9 show results obtained for declarative, interrogative and imperative sentences. The original F0 contour, modelled by the INTSINT system, is indicated by squares. The proposed F0 contour, generated with the presented

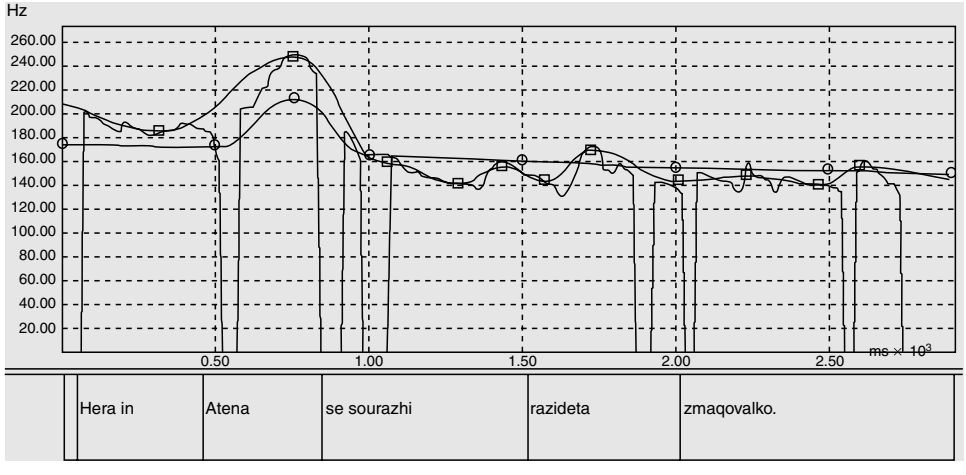


Figure 14.7 Synthetic F0 contour for a declarative sentence, uttered by a female: 'Hera in Atena se sovražni razideta z zmagovalko.' English: 'Hera and Athena hatefully separate from the winner.'
Parameter values:
G(t): $(T = 3s, F_k = 149\,Hz, A_z = 0.47, \alpha = 1.5)$
L(t): $(A_p = 0.13, T_p = 0, d = 0.5s)$

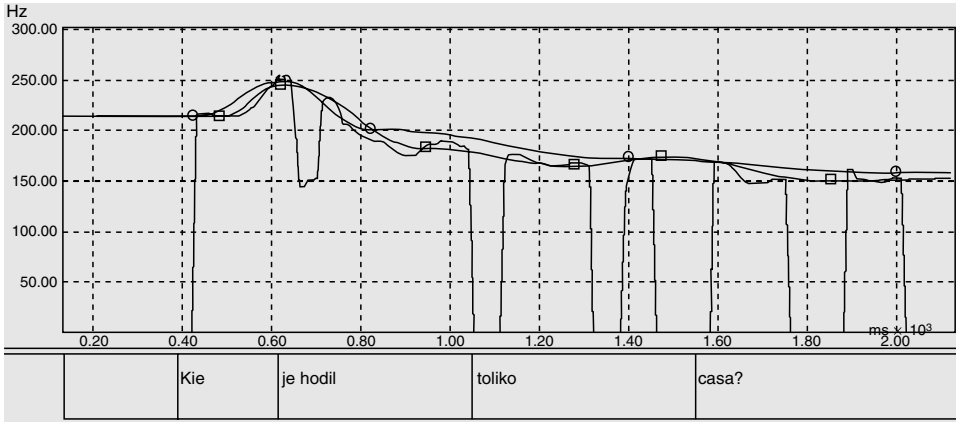


Figure 14.8 Synthetic F0 contour for a Slovene wh-question, uttered by a female 'Kje je hodil toliko časa?' English: 'Where did he walk for so long?'
Parameter values:
G(t): $(T = 1.6s, F_k = 149\,Hz, A_z = 1, \alpha = 1.95)$
L(t) : $(A_p = 0.13, T_p = 0.2\,s, d = 0.2s)$

equations is indicated by circles. Parameter values for the synthetic F0 are given below the figures: T is the duration of the intonation unit.

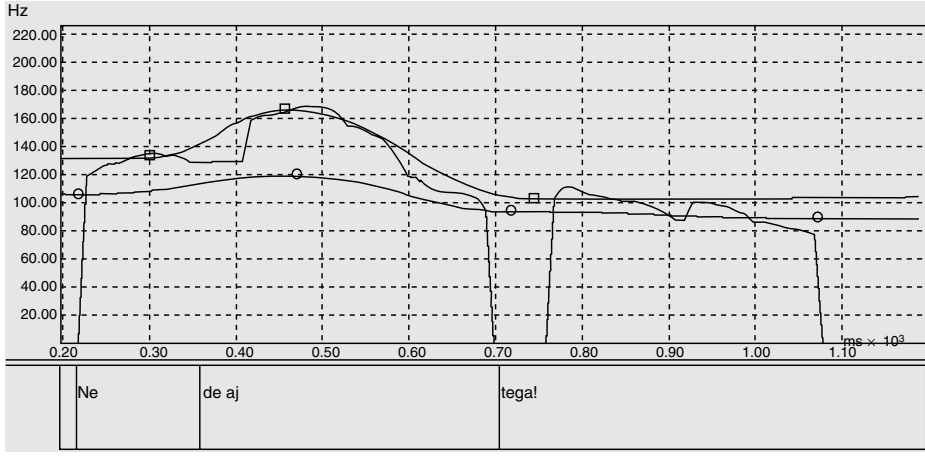


Figure 14.9 Synthetic F0 contour for a Slovene imperative sentence, uttered by a male 'Ne delaj tega!' English: 'Don't do that!'
Parameter values:
G(t): $(T = 0.86s, F_k = 83\,Hz, A_z = 0.7, \alpha = 2.7)$
L(t) : $(A_p = 0.22, T_p = 0.25\,s, d = 0.25s)$

Conclusion

The synthetic F0 contours, based on average parameter values, confirm that the model presented here can simulate natural F0 contours acceptably. In general, for generation of an acceptable F0 contour we need to know the relationship between linguistic units and the structure of the utterance, which includes syntactic-semantic analysis, duration of the intonation unit (related to a chosen speaking rate) and microprosodic parameters. The similarity of natural and synthetic F0 contours is considerably improved if additional information (especially levels and durations of accents) is available.

References

Dobnikar, A. (1996). Modeling segment intonation for Slovene TTS system. *Proceedings of ICSLP'96*, Vol. 3 (pp. 1864–1867). Philadelphia.

Dobnikar, A. (1997). Defining the intonation contours for Slovene TTS system. Unpublished PhD thesis, University of Ljubljana, Slovenia.

Fujisaki, H. (1993). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In O. Fujimura (ed.), *Vocal Physiology: Voice Production, Mechanisms and Functions* (pp. 347–355). Raven.

Fujisaki, H. and Ohno, S. (1993). Analysis and modeling of fundamental frequency contour of English utterances. *Proceedings of EUROSPEECH'95*, Vol. 2 (pp. 985–988). Madrid.

Hirst, D.J. (1994). Prosodic labelling tools. *MULTEXT LRE Project 62–050 Report*. Centre National de la Recherche Scientifique, Université de Provence, Aix-en-Provence.

Hirst, D.J., and Di Cristo, A. (1995). *Intonation Systems: A Survey of 20 Languages*. Cambridge University Press.

Hirst, D.J., Di Cristo, A., Le Besnerais, M., Najim, Z., Nicolas, P., and Roméas, P. (1993). Multi-lingual modelling of intonation patterns. *Proceedings of ESCA Workshop on Prosody*, Working Papers 41 (pp. 204–207). Lund University.

Hirst, D.J., and Espesser, R. (1994). Automatic modelling of fundamental frequency. *Travaux de l'Institut de Phonétique d'Aix, 15* (pp. 71–85). Centre National de la Recherche Scientifique, Université de Provence, Aix-en-Provence.

Ladd, D.R. (1987). A phonological model of intonation for use in speech synthesis by Rule. *Proceedings of EUROSPEECH*, Vol. 2 (pp. 21–24). Edinburgh.

Llisterri, J. (1994). *Prosody Encoding Survey, WP 1 Specifications and Standards, T1.5 Markup Specifications, Deliverable 1.5.3, MULTEXT – LRE Project 62–050*. Universitat Autonoma de Barcelona.

Monaghan, A.I.C. (1991). *Intonation in a Text-to-Speech Conversion System*. PhD thesis, University of Edinburgh.

Pierrehumbert, J.B. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). TOBI: A standard for labeling English prosody. *Proceedings of ICSLP'92* (pp. 867–870). Banff, Alberta, Canada.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

15

Representing Speech Rhythm

Brigitte Zellner Keller and Eric Keller
LAIP, IMM, University of Lausanne,
1015 Lausanne, Switzerland
Brigitte.ZellnerKeller@imm.unil.ch, Eric.Keller@imm.unil.ch

Introduction

This chapter is concerned with the search for relevant primary parameters that allow the formalisation of speech rhythm. In human speech, rhythm usually designates a complex physical and perceptual parameter. It involves the coordination of various levels of speech production (e.g. breathing, phonatory and articulatory gestures, kinaesthetic control) as well as a multi-level cognitive treatment based on the synchronised activation of various cortical areas (e.g. motor area, perception areas, language areas). Defining speech rhythm thus remains difficult, although it constitutes a fundamental prosodic feature.

The acknowledged complexity of what encompasses rhythm partly explains that the common approach to describing speech rhythm is based on a few parameters (such as stress, energy, duration), which are the represented parameters. However, current speech synthesisers show that phonological models do not satisfactorily model speech rhythmicity. In this chapter, we argue that our formal 'tools' are not powerful enough and that they reduce our capacity to understand phenomena such as rhythmicity.

The Temporal Component in Speech Rhythm

To our mind, the insufficiencies in the description and synthesis of rhythm are partly related to the larger issue of how speech temporal structure is modelled in current phonological theory. Prosody modelling has often been reduced to the description of accentual and stress phenomena, and temporal issues such as pausing, varying one's speech rate or 'time-interpreting' the prosodic structures have not yet been as extensively examined and formalised. It is claimed that the status of the temporal component in a prosodic model is a key issue for two reasons.

First, it enters into the understanding of the relations between the temporal and the melodic components in a prosodic system. Second, it enters into the modelling of different styles of speech, which requires prosodic flexibility.

Relations between the Temporal and the Melodic Components

Understanding how the temporal component relates to the melodic component within a prosodic system is of fundamental importance, either from a theoretical or an engineering point of view. This issue is further complicated by the fact that there is no evidence that timing-melody relations are stable and identical across languages (Sluijter and Var Heuven, 1995; Zellner, 1996a, 1998), or across various speech styles. Moreover, our work on French indicates that the tendency of current prosodic theories to invariably infer timing-melody relations solely from accentual structures leads to an inflexible conception of the temporal component of speech synthesis systems.

Flexible Prosodic Models

An additional difficulty is that for the modelling of different styles of reading speech (running texts, lists, addresses, etc.), current rhythmic models conceived for declarative speech would not be appropriate. Does each speech style require an entirely different rhythmic model, a new structure with new parameters? If so, how would such different rhythmic models be related to each other within the same overall language structure? The difficulty of formalising obvious and coherent links between various rhythmic models for the same language may well impede the development of a single dynamic rhythmic system for a given language.

In other words, we suggest that more *explicitness* in the representation of the features contributing to the perception of speech rhythm would facilitate the scientific study of rhythm. If we had such a formalism at our disposal, it would probably become easier to define and understand the exact nature of relations between intonation – i.e., model of melodic contours – and temporal features.

In the following sections, current concepts of speech rhythm will be discussed in more detail. Subsequently, two non-speech human communicative systems will be examined, dance and music notation, since they also deal with rhythm description. These non-verbal systems were chosen because of their long tradition in coding events which contribute to rhythm perception. Moreover, as one system is mainly based on body language perception and the other one is mainly based on auditory perception, it is interesting to look for 'invariants' of the two systems when coding determinants to rhythm. Looking at dance and music notation may help us better understand which information is still missing in our formal representations.

Representing Rhythm in Phonology

Currently in Europe, prosodic structures are integrated into phonological models, and two principal types of abstract structures have essentially been proposed. *Tonal prominence* is assumed to represent pitch accent, and *metrical prominence* is

assumed to represent temporal organisation and rhythm (cf. among others Pierrehumbert, 1980; Selkirk, 1984; Nespor & Vogel, 1986; Gussenhoven, 1988). Rhythm in the metrical approach is expressed in terms of prominence relations between syllables. Selkirk (1984) has proposed a metrical grid to assign positions for syllables, and others like Kiparsky (1979) have proposed a tree structure. Variants of these original models have also been proposed (for example, Hayes, 1995). Beyond their conceptual differences, these models all introduce an arrangement in prosodic constituents and explain the prominence relations at the various hierarchical levels.

Inflexible Models

These representations are considered here to be insufficient, since they generally assume that the prominent element in the phonetic chain is *the* key element for rhythm. In these formalisations, durational and dynamic features (the temporal patterns formed by changes in durations and tempo) are either absent or underestimated. This becomes particularly evident when listening to speech synthesis systems implementing such models. For example, the temporal interpretation of the prosodic boundaries usually remains the same, whatever the speech rate. However, Zellner (1998) showed that the 'time-interpretation' of the prosodic boundaries is dependent on speech rate, since not all prosodic boundaries are phonetically realised at all speech rates. Also, speech synthesisers speak generally faster by compressing linearly the segmental durations. However, it has been shown that the segmental durational system should be adapted to the speech rate (Vaxelaire, 1994; Zellner, 1998). Segmental durations will change not only in terms of their intrinsic durations but also in terms of their relations within the segmental system since all the segments do not present the same 'durational elasticity'. A prosodic model should take into account these different strategies for the realisation of prosodic boundaries.

Binary Models

Tajima (1998) pointed out that 'metrical theory has reduced time to nothing more than linear precedence of discrete grid columns, making an implicit claim that serial order of relatively strong and weak elements is all that matters in linguistic rhythm' (p. 11). This 'prominence approach' shared by many variants of the metrical model leads to a rather rudimentary view of rhythm. It can be postulated that if speech rhythm was really as simple and binary in nature, adults would not face as many difficulties as they do in the acquisition of rhythm of a new language. Also, the lack of clarity on how the strong–weak prominence should be phonetically interpreted leads to an uncertainty in phonetic realisation, even at the prominence level (Coleman, 1992; Local, 1992; Tajima, 1998). Such a 'fuzzy feature' would be fairly arduous to interpret in a concrete speech synthesis application.

Natural Richness and Variety of Prosodic Patterns

After hearing one minute of synthetic speech, it is often easy to conjecture what the prosodic pattern of various speech synthesisers will sound like in subsequent

utterances, suggesting that commonly employed prosodic schemes are too simplistic and too repetitive. Natural richness and variety of prosodic patterns probably participate actively in speech rhythm, and models need enrichment and differentiation before they can be used to predict a more natural and fluid prosody for different styles of speech. In that sense, we should probably take into account not only perceived stress, but also the hierarchical temporal components making up an utterance.

We propose to consider the analysis of rhythm in other domains where this aspect of temporal structure is vital. This may help us identify the formal requirements of the problem. Since the first obstacle speech scientists have to deal with is indeed the *formal representation* of rhythm, it may be interesting to look at dance and music notation systems, in an attempt to better understand what the missing information in our models may be.

Representing Rhythm in Dance and Music

Speaking, dancing and playing music are all time-structured objects, and are thus all subject to the same fundamental interrogations concerning the notation of rhythm. For example, dance can be considered as a frame of actions, a form that progresses through time, from an identifiable beginning to a recognisable end. Within this overall organisation, many smaller movement segments contribute to the global shape of a composition. These smaller form units are known as 'phrases', which are themselves composed of 'measures' or 'metres', based on 'beats'.

The annotation of dance and music has its roots in antiquity and demonstrates some improvements over current speech transcriptions. Even though such notations generally allow many variants – which is the point of departure for artistic expression – they also allow the retrieval of a considerable portion of rhythmic patterns. In other words, even if such a system cannot be a totally accurate mirror of the intended actions in dance and music, the assumption is that these notations permit a more detailed capture and transmission of rhythmic components. The next sections will render more visible these elements by looking at how rhythm is encapsulated in dance and music notation.

Dance Notation

In dance, there are two well-known international notation systems: The Benesh system of Dance Notation[1] and Labanotation.[2] Both systems are based on the same lexicon that contains around 250 terms. An interesting point is that this common lexicon is hierarchically structured.

A first set of terms designates static positions for each part of the body. A second set of terms designates patterns of steps that are chained together. These dynamic sequences thus contain an intrinsic timing of gestures, providing a primary rhythmic structure. The third set of terms designates spatial information with

[1] Benesh system: www.rad.org.uk/index_benesh.htm
[2] Labanotation: www.rz.unifrankfurt.de/~griesbec/LABANE.HTML

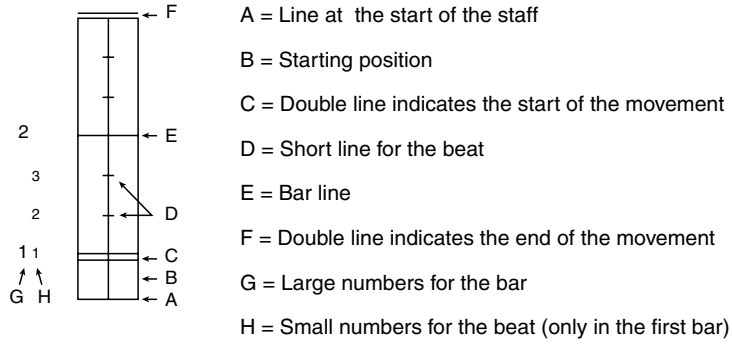


Figure 15.1 (© 1996 Christian Griesbeck, Frankfurt/M)

different references, such as pointing across the stage or to the audience, or references from one to another part of body. The fourth level occasionally used in this lexicon is the 'type' of dance, the choreographic form: a rondo, a suite, a canon, etc.

Since this lexicon is not sufficient to represent all dance patterns, more complex choreographic systems have been created. Among them, a sophisticated one is the Labanotation system, which permits a computational representation of dance. For example Labanotation is a standardised system for transcribing any human motion. It uses a vertical staff composed of three columns Figure 15.1. The score is read from the bottom to the top of the page (instead of left to right like in music notation). This permits noting on the left side of the staff anything that happens on the left side of the body and vice versa for the right side. In the different columns of the staff, symbols are written to indicate in which direction the specific part of the body should move. The length of the symbol shows the time the movement takes, from its very beginning to its end. To record if the steps are long or small, space measurement signs are used. The accentuation of a movement (in terms of prominence) is described with 14 accent signs. If a special overall style of movement is recorded, key signatures (e.g. ballet) are used. To write a connection between two actions, Labanotation uses bows (like musical notation). Vertical bows show that actions are executed simultaneously, they show phrasing.

In conclusion, dance notation is based on a structured lexicon that contains some intrinsic rhythmic elements (patterns of steps). some further rhythmic elements may be represented in a spatial notation system like Labanotation, such as the length of a movement – equivalent to the length of time, the degree of a movement (the quantity), the accentuation, the style of movement, and possibly the connection with another movement.

Music Notation

In music, rhythm affects how long musical notes last (duration), how rapidly one note follows another (tempo), and the pattern of sounds formed by changes in duration and tempo (rhythmic changes). Rhythm in Western cultures is normally formed by changes in duration and tempo (the non-pitch events): it is normally metrical, that is, notes follow one another in a relatively regular pattern at some specified rate.

The standard music notation currently used (five-line staffs, keynotes, bar lines, notes on and between the lines, etc.) was developed in the 1600s from an earlier system called 'mensural' notation. This system permits a fairly detailed transcription of musical events. For example, *pitch* is indicated both by the position of the note and by the clef. *Timing* is given by the length of the note (colour and form of the note), by the time signature and by the tempo. The time signature is composed of bar-lines (' ends a rhythmic group), coupled with a figure placed after the clef (e.g., 2 for 2 beats per measure), and below this figure is the basic unit of time in the bar (e.g., 4 for a quarter of a note, a crotchet). Thus, '2/4' placed after the clef means 2 crotchets per measure. Then comes the tempo which covers all variations of speed (e.g. *lento* to *prestissimo*, number of beats per minute). These movements may be modified with expressive characters (e.g., *scherzo*, *vivace*), rhythmic alterations (e.g., *animato*) or accentual variations (e.g., *legato, staccato*).

In summary, music notation is based on a spatial coding – the staff. A spatially sophisticated grammar permits specifying temporal information (length of a note, time-signature, tempo) as well as the dynamics between duration and tempo. These features are particularly relevant for capturing rhythmic patterns in Western music, and from this point of view, an illustration of the success of this notation system is given by mechanical music as well as by the rhythmically adequate preservation of a great proportion of the musical repertoire of the last few centuries, with due allowance being made for differences to personal interpretation.

Conclusion on these Notations

In conclusion, dance notation and music notation have shown that elements which contribute to the perception of rhythm are represented at various levels of the time-structured object. Much rhythmic information is given by temporal elements at various levels such as the 'rhythmic unit' (duration of the note or the step), the lexical level (patterns of steps), the measure level (time-signature), the phrase level (tempo), as well as by the dynamics between duration and tempo (temporal patterns). Therefore both types of notation represent much more information than only prominent or accentual events.

Proposal of Representation of Rhythm in Speech

Dance and music notations, as shown in the former sections, differ strikingly with respect to the extensive amount of temporal information, typically absent in our models. It is thus proposed to enrich our representations of speech. If rhythm perception results from multidimensional 'primitives', our assumption would be that the richer prosodic formalisms are, the better speech rhythmical determinants will be. In this view, three kinds of temporal information need to be retained: tempo, dynamic patterns and durations.

Tempo determines how fast syllabic units are produced: slow, fast, explicit (i.e., fairly slow, overarticulated), etc. *Tempo* is given at the utterance level (as long as it

doesn't change), and should provide all variations of speed. In our mind, the preliminary establishment of a speech rate in a rhythmic model is important for three reasons.

First, speech rate gives the temporal span by setting the average number of syllables per second. Second, in our model, it also involves the *selection* of the adequate intrinsic segmental durational system, since the segmental durational system is deeply restructured with changes of speaking rate. Third, some phonological structurings related to a specific speech rate can then be modelled: for example in French, schwa treatment or precise syllabification (Zellner, 1998).

Dynamic patterns specify how are related various groups of units, i.e., temporal patterns formed by changes in duration and tempo: word grouping and types of 'temporal boundaries' as defined by Zellner (1996a, 1998). In this scheme, *temporal patterns* are automatically furnished at the phrasing level, thanks to a text parser (Zellner, 1998) and are interpreted according to the applicable tempo (global speech rate). For example, for slow speech rate, an initial minor temporal boundary is interpreted at the syllabic level as a minor syllabic shortening, and a final minor temporal boundary is interpreted as a minor syllabic lengthening. This provides the 'temporal skeleton' of the utterance.

Durations indicate how long units last: durations for syllabic and segmental speech units. This component is already present in current models. *Durations* are specified according to the preceding steps 1 and 2, at the syllabic and segmental levels.

The representation of the three types of temporal information should permit a better modelling and better understanding of speech rhythmicity.

Example

In this section, the suggested concepts are illustrated with a concrete example taken from French. The sentence is 'The village is sometimes overcrowded with tourists'. 'Ce village est parfois encombré de touristes.'

1 Setting the Tempo: fast (around 7 syllables/s)

Since the tempo chosen is fairly fast, some final schwas may be 'reduced' – see next step (Zellner, 1998).

2a Automatic Prediction of the Temporal Patterns

Temporal patterns are initially formed according to the temporal boundaries (m: minor boundary, M: major boundary). These boundaries are predicted on the basis of a text parser (e.g., Zellner, 1996b; Keller & Zellner, 1996) which is adapted depending of the speech rate (Zellner, 1998).

2b Interpretation of the boundaries and prediction of the temporal skeleton

For French, the interpretation of the predicted temporal boundaries depends on the tempo (Zellner, 1998).

'Ce villag(e) est parfois encombré d(e) touristes.'	
M	M

The temporal boundaries are expressed in levels (see below) according to an average syllabic duration (which varies with the tempo). For example, for fast speech rate: a final major boundary (level 3) is interpreted as a major lengthening of the standard syllabic duration. Within the sentence, a pre-pausal phrase boundary or a major phrase boundary is interpreted at the end of the phrase as a minor lengthening of the standard syllabic duration (level 2). Level 0 indicates a shortening of the standard syllabic duration as for the beginning of the sentence. All other cases are realised on the basis of the standard syllabic duration (level 1).

Figures 15.2 and 15.3 show the results of our boundary interpretation according to the fast and to the slow speech rate. Each curve represents the utterance symbolised in levels of syllabic durations. This gives a 'skeleton' of the temporal structure.

3. Computation of the durations

Once the temporal skeleton is defined, the following step consists of the computation of the segmental and syllabic durations of the utterance, thanks to a statistical durational model used in a speech synthesiser. Figures 15.4 and 15.5 represent the obtained temporal curve for the two examples, as calculated by our durational model (Keller & Zellner, 1995, 1996) on the basis of the temporal skeleton. The primitive temporal skeletons are visually clearly related to this higher step. These two figures show the proximity of the predicted curves to the natural ones. Notice that the sample utterance was randomly chosen from 50 sentences.

This example shows to what extent combined changes in tempo, temporal boundaries, and durations impact the whole temporal structure of an utterance, which in turn may affect the rhythmic structure. It is thus crucial to incorporate this temporal information into explicit notations to improve the comprehension of speech rhythm. Initially, tempo could be expressed as syllables per second, dynamic patterns probably require a complex relational representation and duration can be expressed in milliseconds. At a more complex stage, these three components might well be formalisable as an integrated mathematical expression of some generality.

The final step in the attempt to understand speech rhythm would involve the comparison of those temporal curves with traditional intonational contours. Since the latter are focused on prominences, this comparison would illuminate relationships between prominence structures and rhythmic structures.

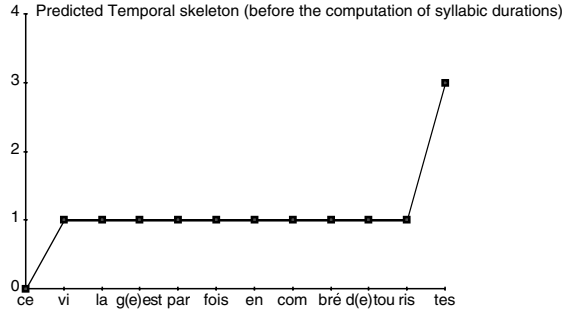


Figure15.2 Predicted temporal skeleton for fast speech rate: '*Ce village est parfois encombré de touristes*'

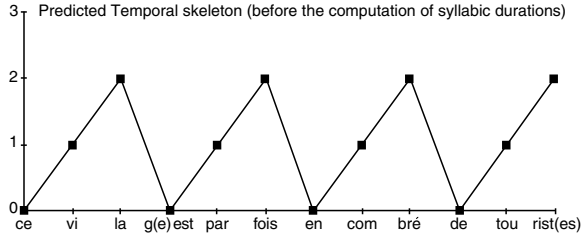


Figure 15.3 Predicted temporal skeleton for slow speech rate: '*Ce village est parfois encombré de touristes*'

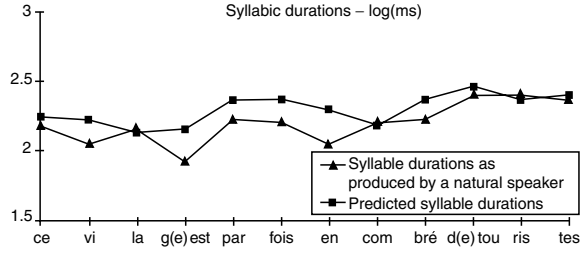


Figure 15.4 Predicted temporal curve and empirical temporal curve for fast speech rate: '*Ce village est parfois encombré de touristes*'

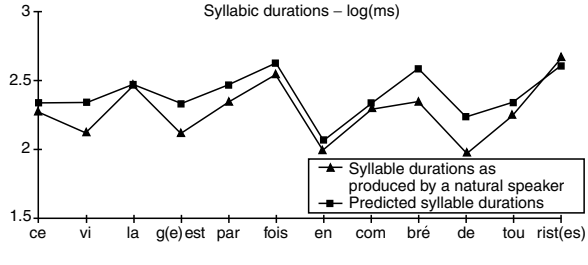


Figure 15.5 Predicted temporal curve and empirical temporal curve for slow speech rate: '*Ce village est parfois encombré de touristes*'

Conclusion

Rhythmic poverty in artificial voices is related to the fact that determinants of rhythmicity are not sufficiently captured with our current models. It was shown that the representation of rhythm is in itself a major issue.

The examination of dance notation and music notation suggests that rhythm coding requires an enriched temporal representation. The present approach offers a general, coherent, coordinated notational system. It provides a representation of the temporal variations of speech at the segmental level, at the syllabic level and at the phrasing level (with the temporal skeleton). In providing tools for the representation of essential information that has till now remained under-represented, a more systematic approach towards understanding speech rhythmicity may well be promoted. In that sense, such a system offers some hope for improving the quality of synthetic speech. If speech synthesis sounds more natural, then we can hope that it will also become more pleasant to listen to.

Acknowledgements

Our grateful thanks to Jacques Terken for his stimulating and extended review. Cordial thanks go also to our colleagues Alex Monaghan and Marc Huckvale for their helpful suggestions on an initial version of this paper. This work was funded by the University of Lausanne and encouraged by the European COST Action 258.

References

Coleman, J. (1992). 'Synthesis by rule' without segments or rewrite-rules. G. Bailly *et al.* (eds), *Talking Machines: Theories, Models, and Designs* (pp. 43–60). Elsevier Science Publishers.

Gussenhoven, C. (1988). Adequacy in intonation analysis: The case of Dutch. In N. Smith & H. Van der Hulst (eds), *Autosegmental Studies on Pitch Accent* (pp. 95–121). Foris.

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. University of Chicago.

Keller, E. and Zellner, B. (1995). A statistical timing model for French. *XIIIth International Congress of Phonetic Sciences, 3* (pp. 302–305). Stockholm.

Keller, E. and Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics, 17*, 53–75. University of York. (Available from http://www.unil.ch/imm/docs/LAIP/Zellnerdoc.html).

Kiparsky, P. (1979). Metrical structure assignement is cyclic. *Linguistic Inquiry, 10*, 421–441.

Local, J.K (1992). Modelling assimilation in a non-segmental, rule-free phonology. In G.J. Docherty and D.R. Ladd (eds), *Papers in Laboratory Phonology*, Vol. II (pp.190–223). Cambridge University Press.

Nespor, M. and Vogel, I. (1986). *Prosodic Phonology*. Foris.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. MIT Press.

Selkirk, E.O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press.

Sluijter, A.M.C. and van Heuven, V.J. (1995). Effects of focus distribution, pitch accent and lexical stress on the temporal organisation of syllables in Dutch. *Phonetica, 52*, 71–89.

Tajima, K. (1998). Speech rhythm in English and Japanese. Experiments in speech cycling. Unpublished PhD. Dissertation. Indiana University.

Vaxelaire, B. (1994). Variation de geste et débit. Contribution à une base de données sur la production de la parole, mesures cinéradiographiques, groupes consonantiques en français. *Travaux de l'Institut de Phonétique de Strasbourg, 24*, 109–146.

Zellner, B. (1996a). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée, 1*, 7–23. Paris.

Zellner, B. (1996b). Relations between the temporal and the prosodic structures of French, a pilot study. *Proceedings of Annual Meeting of the Acoustical Society of America*. Honolulu, HI. (Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm).

Zellner, B. (1998). Caractérisation et prédiction du débit de parole en français. Une étude de cas. Unpublished PhD thesis. Faculté des Lettres, Université de Lausanne. (Available from: http://www.unil.ch/imm/docs/LAIP/Zellnerdoc.html).

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

16

Phonetic and Timing Considerations in a Swiss High German TTS System

Beat Siebenhaar, Brigitte Zellner Keller, and Eric Keller
Laboratoire d'Analyse Informatique de la Parole (LAIP)
Université de Lausanne, CH-1015 Lausanne, Switzerland
Beat.Siebenhaar-Rolli@imm.unil.ch, Brigitte.ZellnerKeller@imm.unil.ch, Eric.Keller@imm.unil.ch

Introduction

The linguistic situation of German-speaking Switzerland shows many differences from the situation in Germany or in Austria. The Swiss dialects are used by everybody in almost every situation – even members of the highest political institution, the Federal Council, speak their local dialect in political discussions on TV. By contrast, spoken Standard German is not a high-prestige variety. It is used for reading aloud, in school, and in contact with people who do not know the dialect. Thus spoken Swiss High German has many features distinguishing it from German and Austrian variants. If a TTS system respects the language of the people to whom it has to speak, this will improve the acceptability of speech synthesis. Therefore a German TTS system for Switzerland has to consider these peculiarities. As the prestigious dialects are not generally written, the Swiss variant of Standard German is the best choice for a Swiss German TTS system.

At the Laboratoire d'analyse informatique de la parole (LAIP) of the University of Lausanne, such a Swiss High German TTS system is under construction. The dialectal variant to be synthesised is the implicit Swiss High German norm such as might be used by a Swiss teacher. In the context of the linguistic situation of Switzerland this means an adaptation of TTS systems to linguistic reality. The design of the system closely follows the French TTS system developed at LAIP since 1991, LAIPTTS-F.[1] On a theoretical level the goal of the German system, LAIPTTS-D, is to see if the assumptions underlying the French system are also

[1] Information on LAIPTTS-F can be found at http://www.unil.ch/imm/docs/LAIP/LAIPTTS.html

applicable to other languages, especially to a lexical stress language such as German. Some considerations on the phonetic and timing levels in designing LAIPTTS-D will be presented here.

The Phonetic Alphabet

The phonetic alphabet used for LAIPTTS-F corresponds closely to the SAMPA[2] convention. For the German version, this convention had to be extended (a) to cover Swiss phonetic reality; and (b) to aid the transcription of stylistic variation:

1. Long and short variants of vowels represent distinct phonemes in German. There is no simple relation to change long into short vowels. Therefore they are treated as different segments.
2. Lexical stress has a major effect on vowels, but again no simple relation with duration could be identified. Consequently, stressed and non-stressed vowels are treated as different segments, while consonants in stressed or non-stressed syllables are not. Lexical stress, therefore, is a segmental feature of vowels.
3. The phonemes /əl/, /əm/, /ən/ and /ər/ are usually pronounced as syllabic consonants [l̩], [m̩], [n̩] and [ɐ̯]. These are shorter than the combination of /ə/ and the respective consonant, but longer than the consonant itself.[3] In formal styles, schwa and consonant replace most syllabic consonants, but this is not a 1:1 relation. These findings led to the decision to define the syllabic consonants as special segments.
4. Swiss speakers tend to be much more sensitive to the orthographic representation than German speakers are. On the phonetic level, the phonetic set had to be enlarged by a sign for an open /ɛ̝/ that is the normal realisation of the grapheme <ä> (Siebenhaar, 1994).

These distinctions result in a phonetic alphabet of 83 distinct segments: 27 consonants, 52 vowels and 4 syllabic consonants. That is almost double the 44 segments used in the French version of LAIPTTS.

The Timing Model

As drawn up for French (Zellner, 1996; Keller *et al.*, 1997), the LAIP approach to TTS synthesis is first to design a timing model and only then to model the fundamental frequency. The task of the timing component is to compute the temporal structure from an annotated phonetic string. In the case of LAIPTTS-D, this string contains the orthographic punctuation marks, marks for word stress, and the distinction between grammatical and lexical words. The timing model has two components. The first one groups prosodic phrases and identifies pauses; the other calculates segmental durations.

[2] Specifications at http://www.phon.ucl.ac.uk/home/sampa/home.htm

[3] [ə+n] mean = 110.2 ms, [n̩] mean = 90.4 ms; [ə+m] mean = 118.3 ms, [m̩] mean = 86.8 ms; [ə+l] mean = 100.1 ms, [l̩] mean = 80.9 ms; [ə+r] mean = 84.4 ms, [ɾ̩] mean = 58.5 ms

The Design of French LAIPTTS and its Adaptation to German

A series of experiments involving multiple general linear models (GLM) for determinants of French segment duration established seven significant factors that could easily be derived from text input: (a) the durational class of the preceding segment; (b) the durational class of the current segment; (c) the durational class of the subsequent segment; (d) the durational class of the next segment but one; (e) the position in the prosodic group of the syllable containing the current segment; (f) the grammatical status of the word containing the current segment; and (g) the number of segments in the syllable containing the current segment. 'Durational class' refers to one of nine clusters of typical durations for segmental duration. These factors have been implemented in LAIPTTS-F. In the move to a multilingual TTS Synthesis, LAIPTTS-D should optimally be based on a similar analysis. Nevertheless, some significant changes had to be considered. The general structure of the German system and its differences from the French system are discussed below.

Database

Ten minutes of read text from a single speaker were manually labelled. The stylistic variants of the text were news, addresses, isolated phrases, fast and slow reading. As the raw segment duration is not normally distributed, the log transformation was chosen for the further calculations. This gave a distribution that was much closer to normal.

Factors Affecting Segmental Duration

To produce a general linear model for timing, the factors with statistical relevance were established in a parametric regression. Most of the factors mentioned in the literature were considered. Step-wise non-significant factors were excluded. Table 16.1 shows the factors finally retained in the model of segmental duration in German, compared to the French system.

The Segmental Aspect

Most TTS systems base their analysis and synthesis of segment durations on phonetic characteristics of the segments and on supra-segmental aspects. For the segmental aspects of LAIPTTS-F, Keller and Zellner (1996) chose a different approach. They grouped the segments according to their mean durations and their articulatory definitions. Zellner (1998, pp. 85 ff.) goes one step further and leaves out the articulatory aspect. This grouping is quite surprising. There are categories containing only one segment, for example [ʃ] in fast speech or [o] in normal speech, which have a statistically different length from all other segments. Other groups contain segments as different as [e, a, b, m *and* t].

Table 16.1 Factors affecting segmental duration in German and French

German	French
Durational class of the current segment	Durational class of the current segment
Type of segment preceding the current segment	Durational class of the segment preceding the current segment
Type of subsequent segment	Durational class of the subsequent segment
–	Durational class of the next segment but one
Type of syllable containing the current segment.	Number of segments in the syllable containing the current segment.
Position of the segment in the syllable	Position of the segment in the syllable
Lexical stress	Syllable containing Schwa
Grammatical status of the word containing the current segment	Grammatical status of the word containing the current segment
Location of the syllable in the word	–
Position in the prosodic group of the syllable containing the current segment	Position in the prosodic group of the syllable containing the current segment

For three reasons, this classification could not be applied directly to German: First, there are more segments in German than in French. Second, there are the phonological differences of long and short vowels. Third, there are major differences in German between stressed and unstressed vowels. Therefore a more traditional approach of using phonetically different classes was employed initially. Any segment was defined by two parameters, containing 17 or 14 phonetic categories (cf. Riedi, 1998, pp. 50–2). Using these segmental parameters and the parameters for the syllable, word, minor and major prosodic group, a general linear model was built to obtain a timing model. Comparing the real values and the values predicted by the model, a correlation of $r = .71$ was found. With only 4 500 segments, the main problem comes from sparsely populated cells. The generalisation of the model was therefore not apparent. There were two ways to rectify this situation: one was to record quite a bit more data, and the other was to switch to the Keller/Zellner model and to group the segments only by their duration. It was decided to do both.

Some 1500 additional segments were recorded and manually labelled. The whole set was then clustered according to segment durations. Initially, an analysis of the single segments was conducted. Then, step by step, segments with no significant difference were included in the groups. At first articulatory definitions were considered significant, but it emerged – as Zellner (1998) had found – that this criterion could be dropped, and only the confidence intervals between the segments were taken into account. In the end, there were 7 groups of segments, and 1 for pauses. Table 16.2 shows these groups.

There is no 1:1 relation between stressed and non-stressed vowels. In group seven, stressed and unstressed diphthongs coincide: stressed [ˈaː] and [ˈɛː] are in this group, while the unstressed versions are in different groups ([aː] is in group six, [ɛ̗ː] in group five). There is also no 1:1 relation between long and short vowels. Unaccented long and short [a] and [ɛ] show different distributions. Short [a] and [ɛ] are both in group three, but [aː] is in group six while [ɛː] is in group five.

Table 16.2 Phoneme class with mean, standard deviation, coefficient of variation, court, percentage

Group	Segments	Mean	Standard deviation	Coefficient of variation	Count	%
1	[r, ɐ]	36.989	16.463	0.445	363	6.09
2	[ɛ, ɪ, i, o, ʊ, u, ʏ, y, ə, j, d, l, ʔ, v, w]	50.174	23.131	0.461	1 634	27.39
3	[ˈɪ, ˈʏ, ˈʊ, ˈiː, ˈyː, ɔ, e, ɐ̯, a, œ, ø, ɐ̯, h, ŋ, n]	64.797	23.267	0.359	1 119	18.76
4	[ˈa, ˈɐ̯, ˈɛ, ˈɔ, ˈœ, iː, uː, g, b]	73.955	22.705	0.307	553	9.27
5	[ˈiː, ˈyː, ɐ̯ː, eː, øː, oː, uː, m̩, n̩, l̩, t, s, z, f, ʃ, ʒ, x]	91.337	35.795	0.392	1 288	21.59
6	[ˈeː, ˈøː, ˈoː, ˈuː, aː, ç, p, k]	111.531	38.132	0.342	384	6.44
7	[ˈaʊ̯, ˈaɪ̯, ˈɔɪ̯, ˈaː, ˈɐ̯ː, ˈãː, ˈɛ̃ː, ˈœ̃ː, ˈõː, aʊ̯, aɪ̯, ɔɪ̯, ãː, ɛ̃ː, œ̃ː, õː, pf, ts]	126.951	41.414	0.326	412	6.91
8	Pause	620.542	458.047	0.738	212	3.55

Keller and Zellner (1996) use the same groups for the influence of the previous and the following segments, as do other systems for input into neural networks. Doing the same with the German data led to an overfitting of the model. Most classes showed only small differences and these were not significant, so the same step-by-step procedure for establishing significant factors as for the segmental influence was performed for the influence of the previous and the following segment. Four classes for the previous segment were distinguished, and three for the following segment:

1. For the previous segment the following classes were distinguished: (a) vowels; (b) affricates and pauses; (c) fricatives and plosives; (d) nasals, liquids, syllabic consonants.
2. The following segment showed influences for (a) pauses; (b) vowels, syllabic consonants and affricates; (c) fricatives, plosives, nasals and liquids.

These three segmental factors explain only 49.5% of the variation of the segments, and 62.1% of the variation including pauses. The model's predicted segmental durations correlated with the measured durations at $r = 0.703$ for the segments only, or at $r = 0.788$ including pauses. This simplified model fits as well as the first model with the articulatory definitions of the segments, but it has the advantage that it has only three instead of six variables, and every variable only has three to eight classes, as compared to 14 to 17 of the first model. The second model is therefore more stable.

The last segmental aspect taken into consideration was the segment's position in the syllable. Besides the position relative to the nucleus, Riedi (1998, p. 52) considers the absolute position as relevant. The data used for present study indicate that this absolute position is not significant. Three positions with significant differ-

ences were found: nucleus, onset, offset. A slightly better fit was achieved when liquids and nasals were considered as belonging to the nucleus.

Aspects at the Syllable Level

For French, the number of segments in the syllable is a relevant factor. For German this aspect was not significant, but it was found that the structure of the syllable containing the current segment is important for every segment. Each of the traditional linguistic distinctions V, CV, VC, CVC was significantly distinct from all others.

Although stress was defined as a segmental feature of vowels, it appeared that a supplementary variable at the syllable level was also significant. For French LAIPTTS-F distinguishes syllables containing a schwa (^{0}) from those with other vowels (^{1}) as nucleus:

Ce vi1lage est parfois encombré de touristes.
$Ce^0/vi^1/llage^1/est^1/par^1/fois^1/en^1/com^1/bré^1/de^0/tou^1/ristes^1$

In addition to the French distinction, a distinction between stressed and unstressed vowels was considered resulting in three stress classes. LAIPTTS-D distinguishes syllables with schwa (^{0}), non-stressed syllables (^{1}) and stressed syllables (^{2}):

Dieses Dorf ist manchmal überschwemmt von Touristen.
$Die^1/ses^0/Dorf^2/ist^1/manch^2/mal^1/ü^1/ber^0/schwemmt^2/von^1/Tou^1/ris^2/ten^0$

This is not as differentiated as other systems because only the main lexical stress is considered, while others also consider stress levels based on syntactic analysis (Riedi, 1998, p. 53; van Santen, 1998, p. 124).

While Riedi (1998, p. 53) considers the number of syllables in the word and the absolute position of the syllable, this was not significant in the present data. The relative position of the syllable was taken into account: monosyllabic words, first, last and medial syllables of polysyllabic words were distinguished.

The marking of the grammatical status of the word containing the current segment is identical to the French system which simply distinguishes lexical and grammatical words. Articles, pronouns, prepositions and conjunctions, modal and auxiliary verbs are considered as grammatical words, all others are lexical words. This distinction is the basis for the definition of minor prosodic groups.

Position of the Syllable Relative to Minor and Major Breaks

LAIPTTS does not perform syntactic analysis beyond the simple phrase. Only the grammatical status of words and the length of the prosodic group define the boundaries of prosodic groups. This approach means that the temporal hierarchy is independent of accent and fundamental frequency effects. It is generally agreed that the first of a series of grammatical words normally marks the beginning of a prosodic group. A prosodic break between a grammatical and a lexical word is

unlikely except for the rare postpositions. The relation between syllables and minor breaks was analysed, revealing three significantly different positions: (a) the first syllable of a minor prosodic group; (b) the last syllable of a minor prosodic group; and (c) a neutral position. These classes are the same as in French. In both languages, segments in the last syllable are lengthened and segments in the first syllable are shortened.

These minor breaks define only a small part of the rhythmic structure. The greater part is covered by the position of syllables in relation to major breaks. A first set of major breaks is defined by punctuation marks, and others are inserted to break up longer phrases. Grosjean and Collins (1979) found that people tend to put these major breaks at the centre of longer phrases.[4] The maximal number of syllables within a major prosodic group is 12, but for different speaking rates, this value has to be adapted. In the French system, there are five pertinent positions: first, second, neutral, penultimate and last syllable in a major phrase. In the German data the difference between the second and neutral syllables was not significant. There are thus four classes in German: (a) shortened first syllables, (b) neutral syllables, (c) lengthened second to last syllables, and (d) even more lengthened last syllables.

Reading Styles

Speaking styles influence many aspects of speech, and should therefore be modelled by TTS systems to improve the naturalness of synthetic speech. For this analysis news, short sentences, addresses, slow and fast reading were recorded. To start with, the analysis distinguished all of these styles, but only the timing of fast and slow reading differed significantly from normal reading. Not all segments differ to the same extent between the two speech rates (Zellner, 1998), and only consonants and vowels were distinguished here: this crude distinction needs to be refined in future studies.

Type of Pause

The model was also intended to predict the length of pauses. These were included in the analysis, with four classes based on the graphic representation of the text: (a) pauses at paragraph breaks; (b) pauses at full stops; (c) pauses at commas; (d) pauses inserted at other major breaks. This coarse classification produces quite good results. As a further refinement, pauses at commas marking the beginning of a relative clause were reduced to pauses of the fourth degree (d), a simple adjustment that can be done at the text level.

Results

The model achieves a reasonable explanation of segment durations for this speaker. The Pearson correlation reaches a value of $r = 0.844$, explaining 71.2% of the

[4] Grosjean confirmed these findings in several subsequent articles with various co-authors.

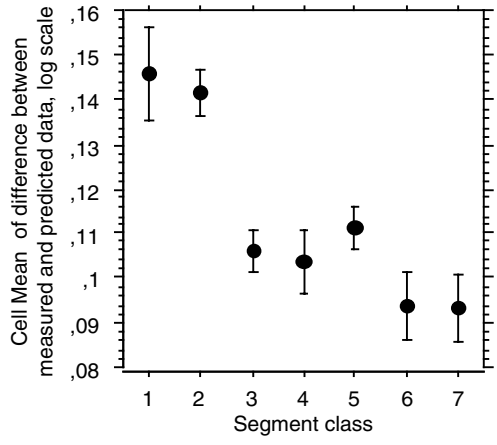


Figure 16.1 Interaction line plot of differences between predicted and measured data (mean and 95% confidence interval), by segment class

overall variance. If pauses are excluded, these values drop to a correlation of $r = 0.763$ and an explained variance of 58.2%. Compared with the values for the segmental information only, this shows that the main information lies in the segment itself, and that a large amount of the variation is still not explained. The correlations of Riedi (1998) and van Santen (1998) are somewhat better. This might be explained by the fact that (a) they have a database that is three to four times larger; (b) their speakers are professionals who may read more regularly; (c) the input for their database is more structured due to syntactically-based stress values; (d) the neural network approach handles more exceptions than a linear model. The model proposed here produces acceptable durations, although it still needs considerable refinement.

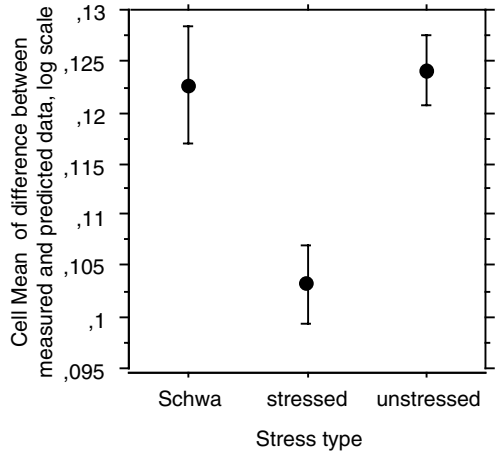


Figure 16.2 Interaction line plot of differences between predicted and measured data (mean and 95% confidence interval), by stress

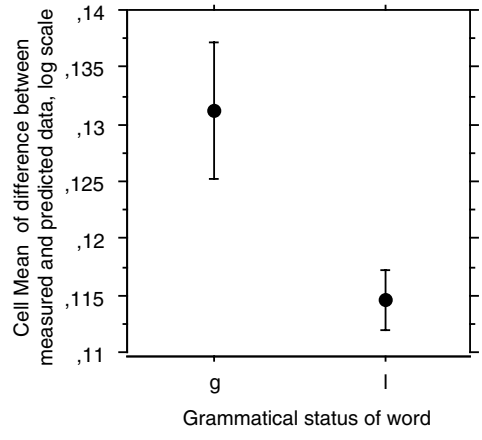


Figure 16.3 Interaction line plot of differences between predicted and measured data (mean and 95% confidence interval), by grammatical status of the word containing the segment

Comparing predicted and actual durations, it seems that the longer segment classes are modelled better than the shorter segment classes (Figure 16.1). Segments in stressed syllables are modelled better than those in unstressed syllables (Figure 16.2), and segments in lexical words are modelled better than those in grammatical words (Figure 16.3). It appears that the different styles or speaking rates can all be modelled in the same manner (Figure 16.4). This approach also predicts the number of pauses and their position quite well, although compared to the natural data it introduces more pauses and in some cases a major break is placed too early.

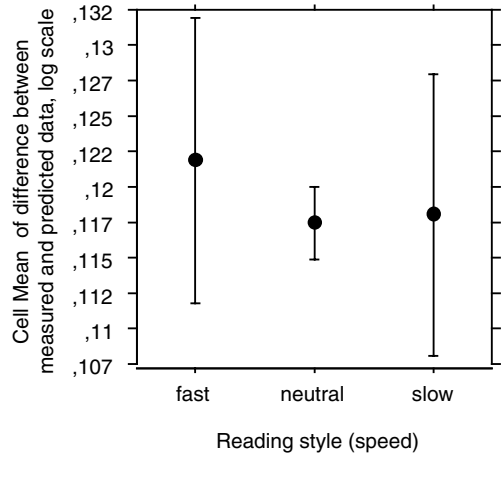


Figure 16.4 Interaction line plot of differences between predicted and measured data (mean and 95% confidence interval), by style

Conclusion

For the timing component of a TTS system, the psycholinguistic approach of Keller and Zellner for French can be transferred to German with minor modifications.

The results show that refinement of the model should focus on specific aspects. On the one hand, extending the database may improve the results generally. On the other hand, only specific parts of the model need be refined. Particular attention should be given to intrinsically short segments, and perhaps different timing models could be used for stressed and non-stressed syllables, or for lexical and grammatical words.

Preliminary tests show that the chosen phonetic alphabet makes it easy to produce different styles by varying the extent of assimilation in the phonetic string: there is no need to build completely different timing models for different speaking styles. The integration of different reading speeds into a single timing model already marks an improvement over the linear shortening of traditional approaches (cf. the accompanying audio examples). The fact that LAIP does not yet have its own diphone database and still uses a Standard German MBROLA database forces us to translate our sophisticated output into a cruder transcription for the sound output. This obscures some contrasts we would have liked to illustrate.

First results of the implementation of this TTS system are available at www.unil.ch/imm/docs/LAIP/LAIPTTS_D_SpeechMill_dl.htm.

Acknowledgements

This research was supported by the BBW/OFES, Berne, in conjunction with the COST 258 European Action.

References

Grosjean, F. and Collins, M. (1979). Breathing, pausing, and reading. *Phonetica, 36*, 98–114.

Keller, E. (1997). Simplification of TTS architecture vs. operational quality. *Proceedings of EUROSPEECH '97*. Paper 735. Rhodes, Greece. September 1997.

Keller, E. and Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics, 17*, 53–75.

Keller, E., Zellner, B. and Werner, S. (1997). Improvements in prosodic processing for speech synthesis. *Proceedings of Speech Technology in the Public Telephone Network: Where are we Today?* (pp. 73–76) Rhodes, Greece.

Riedi, M. (1998). *Controlling Segmental Duration in Speech Synthesis Systems*. Doctoral thesis. Zürich: ETH-TIK.

Siebenhaar, B. (1994). Regionale Varianten des Schweizerhochdeutschen. *Zeitschrift für Dialektologie und Linguistik, 61*, 31–65.

van Santen, J. (1998). Timing. In R. Sproat (ed.), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach* (pp. 115–139). Kluwer.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: la communication parlée, 1*, 7–23.

Zellner, B. (1998). Caractérisation et prédiction du débit de parole en français. Une étude de cas. Unpublished doctoral thesis, University of Lausanne. Available: www.unil.ch/imm/docs/LAIP/ps.files/DissertationBZ.ps

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

17

Corpus-based development of prosodic models across six languages

Justin Fackrell,[1] Halewijn Vereecken,[2] Cynthia Grover,[3] Jean-Pierre Martens[2] and Bert Van Coile[1,2]

[1]*Lernout and Hauspie Speech Products NV*
Flanders Language Valley 50
8900 Ieper, Belgium
[2]*Electronics and Information Systems Department, Ghent University*
Sint-Pietersnieuwstraat 41
9000 Gent, Belgium
[3]*Currently affiliated with Belgacom NV, E. Jacqmainlaan 177, 1030 Brussels, Belgium.*

Introduction

High-quality speech synthesis can only be achieved by incorporating accurate *prosodic models*. In order to reduce the time-consuming and expensive process of making prosodic models manually, there is much interest in techniques which can make them automatically. A variety of techniques has been used for a number of prosodic parameters – among these neural networks and statistical trees have been used for modelling word prominence (Widera *et al*., 1997), pitch accents (Taylor, 1995) and phone durations (Mana and Quazza, 1995; Riley, 1992). However, the studies conducted to date have nearly always concentrated on one particular language, and most frequently, one technique. Differences between languages and corpus designs make it difficult to compare published results directly. By developing models to predict three prosodic variables for six languages, using two different automatic learning techniques, this chapter attempts to make such comparisons. The prosodic parameters of interest are prosodic boundary strength (PBS), word prominence (PROM) and phone duration (DUR). The automatic prosodic modelling techniques applied are multi-layer perceptrons (MLPs) and regression trees (RTs).

The two key variables which encapsulate the prosody of an utterance are intonation and duration. Similar to the work performed at IKP Bonn (Portele and

Heuft, 1997), we have introduced a set of intermediate variables. These permit the prosody prediction to be broken into two independent steps:

1. The prediction of the intermediate variables from the text.
2. The prediction of duration and intonation from the intermediate variables in combination with variables derived from the text.

The intermediate variables used in the current work are PBS and PROM (Figure 17.1). PBS describes the strength of the prosodic break between two words, and is measured on an integer scale from 0 to 3. PROM describes the prominence of a word relative to the other words in the sentence, and is measured on a scale from 0 to 9 (details of the experiments used to choose these scales are given in Grover *et al.*, 1997).

The ultimate aim of this work is to find a way of going from recordings to prosodic models fully automatically. Hence, we need automatic techniques for quickly and accurately adding phonetic and prosodic labels to large databases of speech. Previously, an automatic phonetic segmentation and labelling algorithm was developed (Vereecken *et al.*, 1997; Vorstermans *et al.*, 1996). More recently, we have added an automatic prosodic labelling algorithm as well (Vereecken *et al.*, 1998). In order to allow for a comparison between the performance of our prosodic labeller and our prosodic predictor we will review the prosodic labelling algorithm here as well.

In the next section, we will describe the architecture of the system used for the automatic labelling of PBS and PROM. For labelling, the speech signal and its orthography are mapped to a series of acoustic and linguistic features, which are then mapped to prosodic labels using MLPs. The acoustic features include pitch, duration and energy on various levels; the linguistic ones include part-of-speech labels, punctuation and word frequency. For modelling PBS, PROM and DUR, the same strategy is applied, obviously using only linguistic features. Here, the classifiers can either be RTs or MLPs. We then present labelling and modelling results.

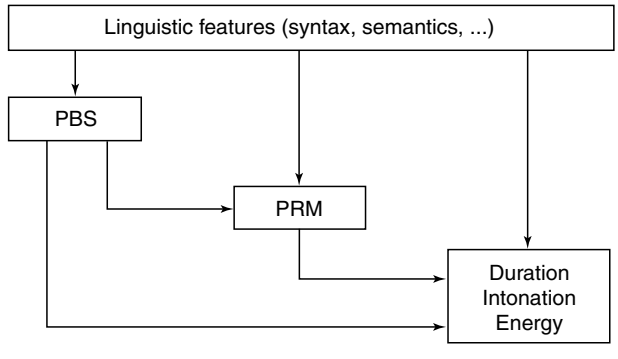


Figure 17.1 Architecture of the TTS prosody prediction

Prosodic Labelling

Introduction

Automatic prosodic labelling is often viewed as a standard recognition problem involving two stages: feature extraction followed by classification (Kiessling *et al.*, 1996; Wightman and Ostendorf, 1994). The feature extractor maps the speech signal and its orthography to a time sequence of feature vectors that are, ideally, good discriminators of prosodic classes. The goal of the classification component is to map the sequence of feature vectors to a sequence of prosodic labels. If some kind of language model describing acceptable prosodic label sequences is included, an optimisation technique like Viterbi decoding is used for finding the most likely prosodic label sequence. However, during preliminary experiments we could not find a language model for prosodic labels that caused a sufficiently large reduction in perplexity to justify the increased complexity implied by a Viterbi decoder. Therefore we decided to skip the language model, and to reduce the prosodic labelling problem to a 'static' classification problem (Figure 17.2).

Feature Extraction and Classification

For the purpose of obtaining *acoustic* features, the speech signal is analysed by an auditory model (Van Immerseel and Martens, 1992). The corresponding orthography is supplied to the grapheme-to-phoneme component of a TTS system, yielding a phonotypical phonemic transcription. Both the transcription and the auditory model outputs (including a pitch value every 10 ms) are supplied to the automatic phonetic segmentation and labelling (annotation) tool, which is described in detail in Vereecken *et al.* (1997) and Vorstermans *et al.* (1996). The phonetic boundaries and labels are used by the prosodic feature extractor to calculate pitch, duration and energy features on various levels (phone, syllable, word, sentence). A linguistic analysis is performed to produce *linguistic* features

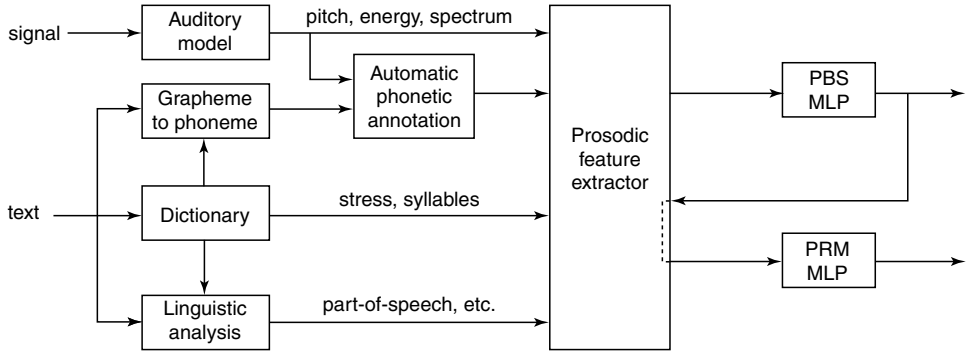


Figure 17.2 Automatic labelling of prosodic boundary strength (PBS) and word prominence (PRM): acoustic and linguistic feature extraction, and feature classification using multilayer perceptrons (MLPs)

such as part-of-speech information, syntactic phrase type, word frequency, accentability (something like the content/function word distinction) and position of the word in the sentence. Syllable boundaries and lexical stress markers are provided by a dictionary. Both acoustic and linguistic features are combined to form one feature vector for each word (PROM labelling) or word boundary (PBS labelling). An overview of the acoustic and linguistic features can be found in Vereecken *et al.* (1998) and Fackrell *et al.* (1999) respectively.

The classification component of the prosodic labeller starts by mapping each PBS feature vector to a PBS label. Since phrasal prominence is affected by prosodic phrase structure, the PBS labels are used to provide phrase-oriented features to the word prominence classifier, such as the PBS before and after the word and the position of the primary stressed syllable in the prosodic phrase. Both classifiers are fully connected MLPs of sigmoidal units, with one hidden layer. The PBS MLP has four outputs, each one corresponding to one PBS value. The PROM MLP has one output only. In this case, PROM values are mapped to the (0:1) interval. The error-backpropagation training of the MLPs proceeds until a maximum performance on some hold-out set is obtained. The automatic labels are rounded to integers.

Prosodic Modelling

The strategy for developing models to predict the prosodic parameters is very similar to that used to label the same parameters. However, there is an important difference, namely that no acoustic features can be used as input features since they are unavailable at the time of prediction. We have adopted a cascade model of prosody in which high-level prosodic parameters (PBS, PROM) are predicted first, and used as input features in the prediction of the low-level prosodic parameter duration (DUR). So, while DUR was input to the PBS and PROM labeller (Figure 17.2), the predicted PBS and PROM are in turn input to the DUR predictor (Figure 17.1). Two separate cascade predictors of phone duration were developed during this work, one using a cascade of MLPs and the other using a cascade of RTs. For each technique, the PBS model was trained first, and its predictions were subsquently used as input features to the PROM model. Both the PBS and the PROM model were then used to add features to the DUR training data.

The MLPs used in this part of the work are two-layer perceptrons. The RTs were grown and pruned following the algorithm of Breiman *et al.* (1984).

Experimental Evaluation

Prosodic Databases

We evaluated the performance of the automatic prosody labeller and the automatic prosody predictors on six databases corresponding to six different languages: Dutch, English, French, German, Italian and Spanish. Each database contains about 1400 *isolated* sentences representing about 140 minutes of speech. The

sentences include a variety of text styles, syntax patterns and sentence lengths. The recordings were made with professional native speakers (one speaker per language). All databases were carefully hand-marked on a prosodic level. About 20 minutes (250 sentences) of each database was hand-marked on a phonetic level as well. Further details on these corpora are given in Grover *et al.* (1998).

The automatic prosodic labelling technique described above has been used to add PBS and PROM labels to the databases. Furthermore, the automatic phonetic annotation (Vereecken *et al.*, 1997; Vorstermans *et al.*, 1996) has been used to add DUR information. However, in this chapter we wish to concentrate on the comparison between MLPs and RTs for modelling, and so we use *manually* rather than *automatically* labelled data as training and reference material. This makes it possible to also compare the performance of the prosody *labeller* with the performance of the prosody *predictor*.

The available data we divided into four sets A, B, C and D. Set A is used for training the PBS and PROM labelling/modelling tools, while set B is used for verifying them. Set C is used to train the DUR models, and set D is held out from all training processes for final evaluation. The sizes of the sets A:B:C:D are in the approximate proportions 15:3:3:1 respectively. The smallest set (D) contains approximately 60 sentences. Sets C+D span the 20-minute subset of the database for which manual duration labels are available, while sets A+B span the remaining 120 minutes. Thus, the proportion of the available data used for training the PBS and PROM models is much larger than that used for training the DUR models. This is a valid approach since the data requirements of the models are different as well: DUR is a phone-level variable whereas PBS and PROM are word-level variables.

Prosodic Labelling Results

In this section we present labelling performances using (1) only acoustic features; and (2) acoustic plus linguistic features. Prosodic labelling using only linguistic features is actually the same as prosodic prediction, the results of which are presented in the next subsection. The training of the prosodic labeller proceeds as follows:

1. A PBS labeller is trained on set A and is used to provide PBS labels for sets A and B.
2. Set A, together with the PBS labels, is used to train the PROM labeller. The PROM labeller is then used to provide PROM labels for sets A and B.

The labelling performance is measured by calculating on each data set the correlation, mean square error and confusion matrix between the automatic and the hand-marked prosodic labels.

The results for PBS and PROM on set B are shown in Tables 17.1 and 17.2 respectively. Since the database contains just sentences, the PBS results apply to within-sentence boundaries only. As the majority of the word boundaries have PBS=0, we have also included the performance of a baseline predictor always

Table 17.1 PBS labelling performance (test set B) of the baseline predictor (PBS=0), an MLP labeller using acoustic features (AC) and an MLP labeller using acoustic plus linguistic features (AC+LI): exact identification (%) and correlation

Language	'PBS=0'	AC	AC + LI
Dutch	70.1	76.4 (0.79)	78.4 (0.82)
English	60.5	74.6 (0.79)	75.0 (0.80)
French	75.2	77.4 (0.74)	78.7 (0.78)
German	70.0	79.0 (0.84)	81.7 (0.87)
Italian	79.6	87.7 (0.88)	88.5 (0.90)
Spanish	86.9	91.6 (0.84)	92.6 (0.86)

Table 17.2 PROM labelling performance (test set B): exact identification $\pm$ 1 (%) and correlation

Language	AC	AC + LI
Dutch	79.1 (0.81)	80.6 (0.82)
English	69.7 (0.82)	76.7 (0.87)
French	76.1 (0.75)	81.7 (0.81)
German	73.6 (0.80)	79.1 (0.84)
Italian	74.6 (0.80)	84.1 (0.89)
Spanish	80.2 (0.83)	92.6 (0.92)

yielding PBS=0. There appears to be a correlation between the complexity of the task (measured by the performance of the baseline predictor) and the labelling performance.

Adding linguistic features does improve the prosodic labelling performance significantly. The PROM labelling is improved dramatically; the improvements for PBS are smaller, but taken as a whole they are significant too. Hence, there seems to be some vital information contained in the linguistic features. This could indicate that the manual labellers were to some extent influenced by the text, which is of course inevitable. The correlations from Tables 17.1 and 17.2 compare favourably with the inter-transcriber agreements mentioned in (Grover *et al.*, 1997).

Prosodic Modelling

The training of the cascade prosody model proceeds as follows:

1. A PBS model is trained on Set A and is then used to make predictions for all four data sets A–D.
2. Set A, together with the PBS predictions, is used to train a PROM model. The PROM model is then used to make predictions for all four data sets. Set B is

used as hold-out set. The double use of set A in the training procedure, albeit for different prosodic parameters, does carry a small risk of overtraining.
3. Set C, together with the predictions of the PBS and PROM models, was used to train a DUR model.
4. Set D, which was not used at any time in the training procedure, was used to evaluate the DUR model.

Tables 17.3, 17.4, 17.5 and 17.6 compare the performance of the MLP and the RT models at each stage in the cascade, compared to manual labels of PBS, PROM and DUR respectively.

Table 17.3 PBS predicting performance (test set B) of baseline, MLP and RT predictors: exact identification (%)

Language	'PBS=0'	MLP	RT
Dutch	70.1	72.3	72.7
English	60.5	65.2	65.6
French	75.2	74.2	71.4
German	70.0	74.8	72.7
Italian	79.6	78.2	79.1
Spanish	86.9	88.7	89.7

Table 17.4 PBS predicting performance (test set B) of baseline, MLP and RT predictors: exact identification $\pm$ 1 (%)

Language	'PBS=0'	MLP	RT
Dutch	85.6	94.9	94.7
English	85.0	95.5	94.7
French	81.4	91.0	91.3
German	85.3	96.3	96.3
Italian	87.2	97.0	97.4
Spanish	93.2	97.3	97.3

Table 17.5 PROM predicting performance (test set B) of cascade MLP and RT predictors: exact identification $\pm$ 1 (%)

Language	MLP	RT
Dutch	72.1	72.8
English	69.9	72.9
French	76.9	81.4
German	74.5	74.8
Italian	80.0	80.3
Spanish	90.8	92.2

Table 17.6 DUR predicting performance (test set D) of cascade MLP and RT predictors: correlation between the predictions of the model and the manual durations

Language	MLP	RT
Dutch	0.80	0.79
English	0.78	0.75
French	0.73	0.69
German	0.78	0.75
Italian	0.84	0.83
Spanish	0.75	0.72

The prediction results in Table 17.3 show that, as far as exact prediction performance is concerned, all models predict PBS more accurately than the baseline predictor, with the exceptions of French and Italian. However, if a margin of error of $\pm$ 1 is allowed (Table 17.4), then all models perform much better than the baseline predictor. The difference between the performance of MLP and RT is negligible in all cases.

Table 17.5 shows that the RT model is slightly better than the MLP model in all cases, at predicting PROM. As in Tables 17.3 and 17.4, English has some of the lowest prediction rates, while Spanish has the highest.

Note that the PBS modelling results are worse than the corresponding labelling results (Table 17.1), which is to be expected since the labeller has access to acoustic (AC) information as well. However, for PROM the labelling results based on AC features alone (Table 17.2) seem to be worse or comparable to the MLP PROM modelling results (Table 17.5) most of the time. This suggests that for these languages the manual labellers are influenced more strongly by linguistic evidence than by acoustic evidence. This also explains why there is such a big improvement in PROM labelling performance when using all the available features (AC+LI).

Table 17.6 shows that although the RT model performs best at PROM prediction, the MLP models for DUR outperform the RT models for each language, albeit slightly. One possible explanation for this is that although DUR, PBS and PROM are all measured on an interval-scale, PBS and PROM can take only a limited number of values, whereas DUR can take *any* value between certain limits.

Conclusion

In this chapter the automatic labelling and modelling of prosody were described. During labelling, the speech signal and the text are first transformed to a series of acoustic and linguistic variables, including duration. Next, these variables are used to label the prosodic structure of the utterance (in terms of boundary strength and word prominence). The prediction of duration from text alone proceeds in reverse order: the prosodic structure is predicted and serves as input to the duration prediction. A comparison between regression trees and multi-layer perceptrons seems

to suggest that whilst the RT is capable of outperforming the MLP in the PROM and PBS tasks, it performs worse than the MLP in the prediction of DUR.

More recently, a perceptual evaluation of these duration models (Fackrell *et al.*, 1999) has suggested that they are at least as good as hand-crafted models, and sometimes even better. Furthermore, using the automatic labelling techniques to prepare the training data, rather than using the manual labelling, seemed to have no negative impact on the model performance.

Acknowledgments

This research was performed with support of the Flemish Institute for the Promotion of the Scientific and Technological Research in the Industry (contract IWT/AUT/950056). COST Action 258 is acknowledged for providing a useful platform for scientific discussions on the topics treated in this chapter. The authors would like to acknowledge the contributions made to this research by Lieve Macken and Ellen Stuer.

References

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International.

Fackrell, J., Vereecken, H., Martens, J.-P., and Van Coile, B. (1999). Multilingual prosody modelling using cascades of regression trees and neural networks. *Proceedings of Eurospeech* (pp. 1835–1838). Budapest.

Grover, C., Fackrell, J., Vereecken, H., Martens, J.-P., and Van Coile, B. (1998). Designing prosodic databases for automatic modelling in 6 languages. *Proceedings of ESCA/COCOSDA Workshop on Speech Synthesis* (pp. 93–98). Jenolan Caves, Australia.

Grover, C., Heuft, B., and Van Coile, B. (1997). The reliability of labeling word prominence and prosodic boundary strength. *Proceedings of ESCA Workshop on Intonation* (pp. 165–168). Athens, Greece.

Kiessling, A., Kompe, R., Batliner, A., Niemann, H., and Nöth, E. (1996). Classification of boundaries and accents in spontaneous speech. *Proceedings of the 3rd CRIM/FORWISS Workshop* (pp. 104–113). Montreal.

Mana, F. and Quazza, S. (1995). Text-to-speech oriented automatic learning of Italian prosody. *Proceedings of Eurospeech* (pp. 589–592). Madrid.

Portele, T. and Heuft, B. (1997). Towards a prominence-based synthesis system. *Speech Communication*, *21,* 61–72.

Riley, M.D. (1992). Tree-based modelling of segmental durations, in In G. Bailly, C. Benoit, and T.R. Sawallis (eds), *Talking Machines: Theories, Models, and Designs* (pp. 265–273). Elsevier Science.

Taylor, P. (1995). Using neural networks to locate pitch accents, *Proceedings of Eurospeech* (pp. 1345–1348). Madrid.

Van Immerseel, L. and Martens, J.-P. (1992). Pitch and voiced/unvoiced determination using an auditory model. *Journal of the Acoustical Society of America*, *91(6)*, 3511–3526.

Vereecken, H., Martens, J.-P., Grover, C., Fackrell, J., and Van Coile, B. (1998). Automatic prosodic labeling of 6 languages. *Proceedings of ICSLP* (pp. 1399–1402). Sydney.

Vorstermans, A., Martens, J.-P., and Van Coile, B. (1996). Automatic segmentation and labelling of multi-lingual speech data. *Speech Communication*, *19,* 271–293.

Vereecken, H., Vorstermans, A., Martens, J.-P., and Van Coile, B. (1997). Improving the phonetic annotation by means of prosodic phrasing. *Proceedings of Eurospeech* (pp. 179–182). Rhodes, Greece.

Widera, C., Portele, T., and Wolters, M. (1997). Prediction of word prominence. *Proceedings of Eurospeech* (pp. 999–1002). Rhodes, Greece.

Wightman, C. and Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, *2(4)*, 469–481.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

18

Vowel Reduction in German Read Speech

Christina Widera
Institut für Kommunikationsforschung und Phonetik (IKP), University of Bonn, Germany
cwi@ikp.uni-bonn.de

Introduction

In natural speech, a lot of inter- and intra-subject variation in the realisation of vowels is found. One factor affecting vowel reduction is speaking style. In general, spontaneous speech is regarded to be more reduced than read speech. In this chapter, we examine whether in read speech vowel reduction can be described by discrete levels and how many levels are reliably perceived by subjects. The reduction of a vowel was judged by matching stimuli to representatives of reduction levels (prototypes). The experiments show that listeners can reliably discriminate up to five reduction levels depending on the vowel and that they use the prototypes speaker-independently.

In German 16 vowels (monophthongs) are differentiated: eight tense vowels, seven lax vowels and the reduced vowel 'schwa'. /i:/, /e:/, /ɛ:/, /a:/, /u:/, /o:/, /y:/, and /ø:/ belong to the group of tense vowels. This group is opposed to the group of lax vowels (/ɪ/, /ɛ/, /a/, /ʊ/, /ɔ/, /ʏ/, and /œ/). In a phonetic sense the difference between these two groups is a qualitative as well as a quantitative one (/i:/ vs. /ɪ/, /e:/ and /ɛ:/ vs. /ɛ/, /u:/ vs. /ʊ/, /o:/ vs. /ɔ/, /y:/ vs. /ʏ/, and /ø:/ vs. /œ/). However, the realisation of the vowel /a/ differs in quantity: qualitative differences are negligible ([a:] vs. [a]; cf. Kohler, 1995a).

Vowels spoken in isolation or in a neutral context are considered to be ideal vowel realisations with regard to vowel quality. Vowels differing from the ideal vowel are described as reduced. Vowel reduction is associated with articulators not reaching the canonical target position (target undershoot; Lindblom, 1963). From an acoustic point of view, vowel reduction is described by smaller spectral distances between the sounds. Perceptually, reduced vowels sound more like 'schwa'.

Vowel reduction is related to prosody and therefore to speaking styles. Depending on the environment (speaker-context-listener) in which a discourse takes place, different speaking styles can be distinguished (Eskénazi, 1993). Read speech tends to be more clearly and carefully pronounced than spontaneous speech

(Kohler, 1995b), but inter- and intra-subject variation in the realisation of vowels is also found.

Previous investigations of perceived vowel reduction show that the inter-subject agreement is quite low. Subjects had to classify vowels according to their vowel quality into two (full vowel or 'schwa'; van Bergem, 1995) or three groups (without any duration information; Aylett and Turk, 1998). The question addressed here is whether in read speech listeners can reliably perceive several discrete reduction levels on the continuum from unreduced vowels to the most reduced vowel ('schwa'), if they use representatives of reduction levels as reference.

In this approach vowels at the same level are considered to exhibit the same degree of reduction: differences in quality between them can be ignored. A description of reduction in terms of level allows statistical analyses of reduction phenomena and the prediction of reduction level. This is of interest for vowel reduction modelling in speech synthesis to increase the naturalness of synthesised speech and for an adaptation to different speaking styles.

Database

The database from which our stimuli were taken ('Bonner Prosodische Datenbank') consists of isolated sentences, question and answer pairs, and short stories read by three speakers (two female, one male; Heuft *et al.*, 1995). The utterances were labelled manually (SAMPA, Wells, 1996). There are 2 830 tense and 5 196 lax vowels. Each vowel is labelled with information about its duration. For each vowel, the frequencies of the first three formants were computed every 5 ms (ESPS 5.0). The values of each formant for each vowel were estimated by a third order polynominal function. The polynomial is fitted to the formant trajectory. The formant frequency of a vowel is defined here as the value in the middle of that vowel (Stöber, 1997). The formant values (Hz- and mel-scaled) within each phoneme class of a speaker were standardised with respect to the mean value and the deviation (z-scores).

Perceptual Experiments

The experiments are divided into two main parts. In the first part, we examined how many reduction levels exist for the eight tense vowels of German. The tense vowels were grouped by mean cluster analysis. It was assumed that the clustering of the vowels would indicate potential prototypes of reduction levels. In perception experiments subjects had to arrange vowels according to their strength of reduction. Then, the relevance of the prototypes for reduction levels was tested by assigning further vowels to these prototypes. The results of this classification showed that not all prototypes can be regarded as representative of reduction levels. These prototypes were excluded and the remaining prototypes were evaluated by further experiments. In the second part reduction phenomena of the seven lax German vowels were investigated using the same method as for the tense vowels.

Tense Vowels

Since the first two formant frequencies (F1, F2) are assumed to be the main factors determining vowel quality (Pols *et al.*, 1969), the F1 and F2 values (mel-scaled and standardised) of the tense vowels of one speaker (Speaker 1) were clustered by mean cluster analysis. The number of clusters varied from two to seven for each of the eight tense vowels.

In a pre-test, a single subject judged perceptually the strength of the reduction of vowels in the same phonetic context (open answer form). The perceived reduction levels were compared with the groups of the different cluster analyses. The results show a higher agreement between perceptual judgements and the cluster analysis with seven groups for the vowels [i:], [y:], [a:], [u:], [o:] and with six groups for [e:], [ɛ:], and [ø:] than between the judgements and the classifications of the other cluster analyses.

For each cluster, one prototype was determined whose formant values were closest to the cluster centre. Within a cluster, the distances between the formant values (mel-scaled and standardised) and the cluster centre (mel-scaled and standardised) were computed by:

$$d = (ccF1 - F1)^2 + (ccF2 - F2)^2 \quad (1)$$

where *ccF1* stands for mean F1 value of the vowels of the cluster; *F1* is the F1 value of a vowel of the same cluster; *ccF2* stands for mean F2 value of the vowels of the same cluster; *F2* is the F2 value of a vowel of the same cluster. The hypothesis that these prototypes are representatives for different reduction levels, is tested with the following method.

Method

Perceptual experiments were carried out for each of the eight tense vowels separately. The task was to arrange the prototypes by strength of reduction from unreduced to reduced. The reduction level of each prototype was defined by the modal value of the subjects' judgements. Nine subjects participated in the first perception experiment. All subjects are experienced in labelling speech. The prototypes were presented on the computer screen as labels. The subjects could listen to each prototype as often as they wanted via headphones.

In a second step, subjects had to classify stimuli based on their perceived qualitative similarity to these prototypes. Six vowels from each cluster (if available) whose acoustical values are maximally different as well as the prototypes were used as stimuli. The test material contained each stimulus twice (for [i:], [o:], [u:] $n = 66$; for [a:] $n = 84$; for [e:] $n = 64$; for [y:] $n = 48$; for [ɛ:] $n = 40$; for [ø:] $n = 36$; where n stands for the number of vowels judged in the test). Each stimulus was presented over headphones together with the prototypes as labels on the computer screen. The subjects could listen to the stimuli within and outside their syllabic context and could compare each prototype with the stimulus as often as they wanted. Assuming that a stimulus shares its reduction level with the pertinent prototype, each stimulus received the reduction level of its prototype. The overall reduction

level (ORL) of each judged stimulus was determined by the modal value of the reduction levels of the individual judgements.

Results

Prototype stimuli were assigned to the prototypes correctly in most of the cases (average value of all subjects and vowels: 93.6%). 65.4% of all stimuli (average value of all subjects and vowels) were assigned to the same prototype in the repeated presentation. The results indicate that the subjects are able to assign the stimuli more or less consistently to the prototypes, but it is a difficult task due to the large number of prototypes.

The relevance of a prototype for the classification of vowels was determined on the basis of a confusion matrix. The prototypes themselves were excluded from the analysis. If individual judgements and ORL agreed in more than 50% and more than one stimulus was assigned to the prototype, then the prototype was assumed to represent one reduction level. According to this criterion the number of prototypes was reduced to five for [i:], [u:] as well as for [e:], and three for the other vowels. The resulting prototypes were evaluated in further experiments with the same design as used before.

Evaluation of prototypes

Eight subjects were asked to arrange the prototypes with respect to their reduction and to transcribe them narrowly using the IPA system. Then they had to classify the stimuli using the prototypes. Stimuli were vowels with maximally different syllabic context. Each stimulus was presented twice in the test material (for [i:] $n = 82$; for [o:] $n = 63$; for [u:] $n = 44$; for [a:] $n = 84$; for [e:] $n = 68$; for [y:] $n = 52$; for [ɛ:] $n = 34$; for [ø:] $n = 30$).

For [i:] it was found that two prototypes are frequently confused. Since those prototypes sound very similar, one of them was excluded. The results are based on four prototypes evaluated in the next experiment (cf. section on speaker-independent reduction levels).

The average agreement between individual judgements and ORL (stimuli with two modal values were excluded) is equal to or greater than 70% for all vowels (Figure 18.1). χ^2-tests show a significant relation between the judgements of any two subjects for most vowels (for [i:], [u:], [e:], [o:], [y:] $p < .01$; for [a:] $p < .02$; for [ɛ:] $p < .05$). Only for [ø:], nine non-significant ($p > .05$) inter-subject judgements are found, most of them (six) due to the judgement of one subject.

To test whether the agreement has improved because the prototypes are good representatives of reduction levels or only because of the decrease in their number, the agreement between individual judgements and ORL was computed with respect to the number of prototypes (Lienert and Raats, 1994):

$$agreement\ \ (pa) = n\ (ra) - \frac{n\ (wa)}{n\ (pa) - 1} \tag{2}$$

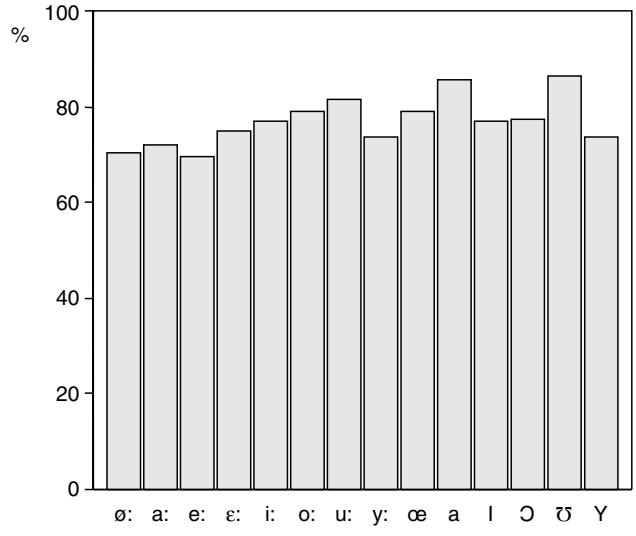


Figure 18.1 Average agreement between individual judgements and overall reduction level for each vowel

where *n(ra)* is the number of matching answers between ORL and individual judgements (right answers); *n(wa)* is the number of non-matching answers between the two values (wrong answers); *n(pa)* is the number of prototypes (possible answers).

In comparison to the agreement between individual judgements and ORL in the first experiment, the results have indeed improved (Figure 18.2). It can be assumed that the prototypes represent reduction levels, and the assigned stimuli

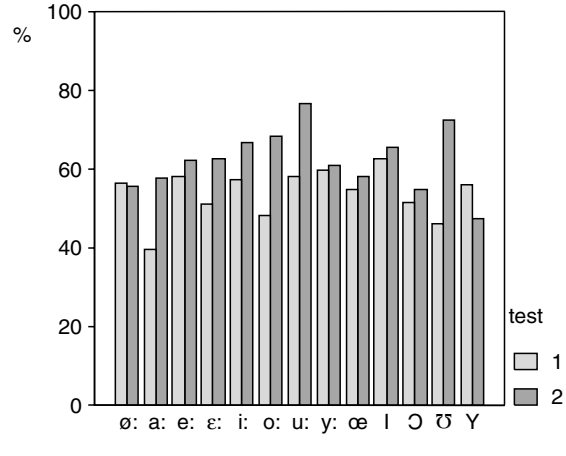


Figure 18.2 Agreement between individual judgments and overall reduction level with respect to the number of prototypes of the first (1) and second (2) experiment for each vowel

can be regarded as classified with respect to their reduction. This is supported by the inter-subject agreement of judgements for most vowels. The average correlationbetween any two subjects is significant at the .01– level for the vowels [i:], [e:], [u:], [o:], [y:] and at the .04-level for [ɛ:]. For [a:] and [ø:], the inter-subject correlation is low but significant at the .02 or at the .05-level, respectively (Figure 18.3).

Speaker-independent reduction levels

A further experiment investigated whether the reduction levels and their prototypes can be transferred to other speakers. Eight subjects had to judge five stimuli for each speaker and for each reduction level. The same experimental design as in the other perception experiments was used. The comparison of individual judgements and ORL shows that independently of the speaker, the average agreement between these values is quite similar (76.4% for Speaker 1; 73.1% for Speaker 2; 76.5% for Speaker 3; Figure 18.4).

In general, the correlation of any two subjects' judgements is comparable to the correlation of the last set of experiments (Figure 18.3). These results show that within this experiment subjects compensate for speaker differences. They are able to use the prototypes speaker-independently.

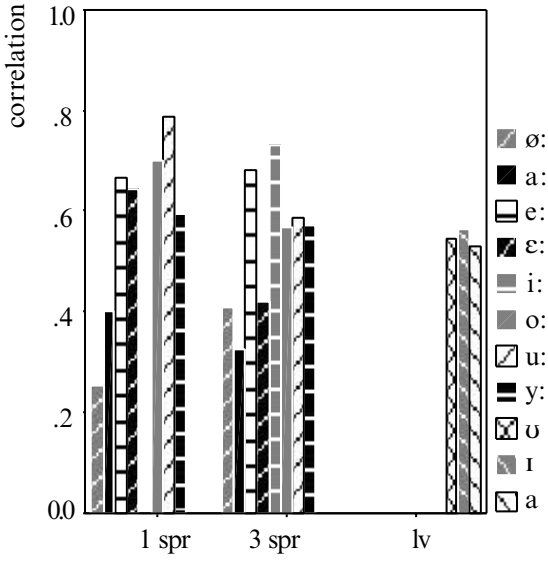


Figure 18.3 Correlation for each vowel grouped by experiments. Correlation between subjects of the test with tense vowels of one speaker (1 spr; correlation for [i:] was not computed for 1 spr, cf. section on Evaluation of prototypes) and of three speakers (3 spr); correlation between subjects of the test with lax vowels (lv)

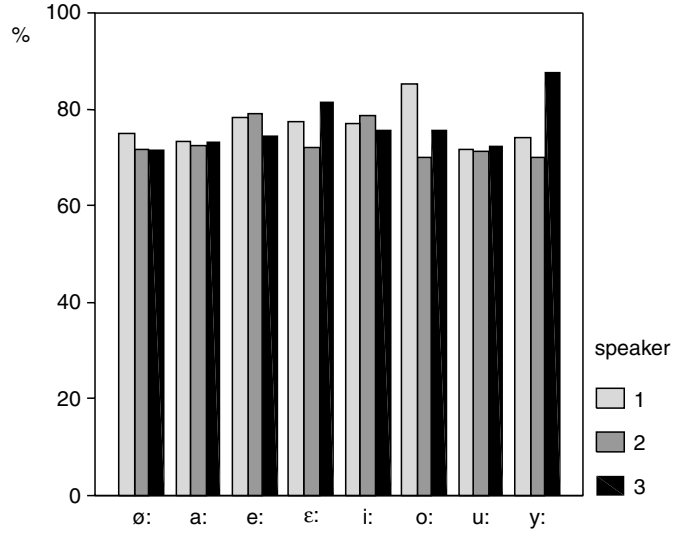


Figure 18.4 Average agreement between individual judgements and overall reduction level depending on the speaker for each tense vowel

Lax Vowels

Method

On the basis of this speaker-independent use of prototypes, the F1 and F2 values (mel-scaled and standardised) of lax vowels of all three speakers were clustered. The number of clusters fits the number of the resulting prototypes of the tense counterpart: four groups for [ɪ] and three groups for [ɛ], [a], [ɔ], [œ], and [ʏ]. For [ʊ] only three groups are taken, because two of the five prototypes of [u:] are limited to a narrow range of articulatory context. From each cluster, one prototype was derived (cf. section on tense vowels, equation 1). The number of prototypes of [ɛ] and of [a] is decreased to two, because the clusters of these prototypes only contain vowels with unreliable formant values.

As in the perception experiments for the tense vowels, eight subjects had to arrange the prototypes by their strength of reduction and to judge the reduction by matching stimuli to prototypes according to their qualitative similarity. Stimuli were vowels with maximally different syllabic context (for [ɪ] $n = 60$; for [ʊ] $n = 71$; for [œ] $n = 43$; for [ɛ] $n = 29$; for [ʏ], [ɔ], [a] $n = 45$; where n stands for the number of vowels presented in the test).

Results

The results show that the number of prototypes has to be decreased to three for [ɪ] due to a high confusion rate between two prototypes, and to two for [ʊ], [ɔ], [œ], and [ʏ] because of non-significant relations between the judgements of any two subjects (χ^2-tests, $p > .05$). These prototypes were tested in a further experiment.

For [ɛ] with two prototypes no reliably perceived reduction levels are found ($p > .05$). For [a], there is an agreement between individual judgements and ORL of 85.4% (Figure 18.1). χ^2-tests indicate a significant relation between the inter-subject judgements ($p < .02$).

A follow-up experiment was carried out with the decreased number of prototypes and the same stimuli used in the previous experiment. Figure 18.1 shows the agreement between individual judgements and ORL. The agreement between individual judgements and ORL with respect to the number of prototypes is improved by the decrease of prototypes for [ɪ], [ʊ], [ɔ], and [œ] (Figure 18.2). However, χ^2-tests only indicate significant relations between the judgements of any two subjects for [ɪ] and [ʊ] ($p < .01$).

The results indicate three reliably perceived reduction levels for [ɪ] and two reduction levels for [ʊ] and [a]. For the other four lax vowels [ɛ], [ɔ], [œ], and [ʏ] no reliably perceived reduction levels can be found. This contrasts sharply with the finding that subjects are able to discriminate reduction levels for all tense vowels. For [ɪ], [ʊ], and [a] the average agreement with respect to the number of prototypes (69.7 %) is comparable to that of tense vowels (63.8 %). The mean correlation between any two subjects is significant for [ʊ] ($p < .01$), [ɪ] ($p < .05$), and [a] ($p < .03$; Figure 18.3), but on average it is lower than those of the tense vowels. One possible reason for this effect could be duration. The tense vowels (mean duration: 80.1 ms) are longer than the lax vowels (mean duration: 57.6 ms). However, within the group of lax vowels, duration does not affect the reliability of discrimination (mean duration of lax vowels with reduction levels: 56.1 ms and of lax vowels without reliably perceived reduction levels: 59.3 ms).

Conclusion

The aim of this research was to investigate a method for labelling vowel reduction in terms of levels. Listeners judged the reduction by matching stimuli to prototypes according to their qualitative similarity. The assumption is that vowel realisations have the same reduction level as their chosen prototypes. The results were investigated according to inter-subject agreement.

These experiments indicate that a description of reduction in terms of levels is possible and that listeners use the prototypes speaker-independently. However, the number of reduction levels depends on the vowel. For the tense vowels reliably perceived reduction levels could be found. In contrast, reduction levels can only be assumed for three of the seven lax vowels, [ɪ], [ʊ], and [a].

The results can be explained by the classical description of the vowels' place in the vowel quadrilateral. According to the claim that in German the realisation of the vowel /a/ predominantly differs in quantity ([a:] vs. [a]; cf. Kohler, 1995a), the vowel system can be described by a triangle (cf. Figure 18.5). The lax vowels are closer to the 'schwa' than the tense vowels. Within the set of lax vowels [ɪ], [ʊ], and [a] are at the edge of the triangle. Listeners only discriminate reduction levels for these vowels, and their number of reduction levels is lower than those of their tense counterparts [i:], [u:], and [a:].

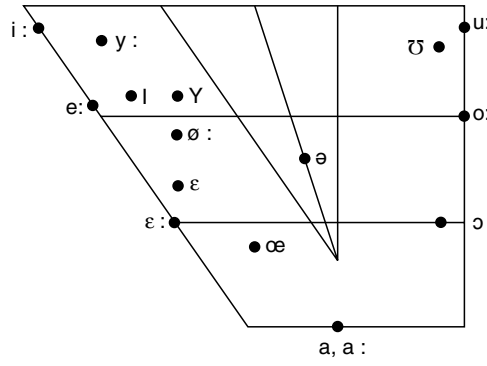


Figure 18.5 Phonetic realisation of German monophthongs (from Kohler, 1995a, p. 174)

The transcription (IPA) of the prototypes indicates that a reduced tense vowel is perceived as its lax counterpart (i.e. reduced /u/ is perceived as [ʊ]), with the exception of [o:], where the reduced version is associated with a decrease in rounding. Between reduced tense vowels perceived as lax and the most reduced level, labelled as centralised or as schwa, no further reduction level is discriminated. This is also observed for the three lax vowels. However, in comparison to the lax vowels, listeners are able to discriminate reliably between a perceived lax vowel quality and a more centralised (schwa like) vowel quality for all tense vowels. The question is whether the reduced versions of the tense vowels [ɛ:], [o:], [y:], and [ø:] which are perceived as lax are comparable with the acoustic quality of their lax counterparts ([ɛ], [ɔ], [ʏ], and [œ]).

On the one hand, for [ɛ:] and [o:] spectral differences (mean of standardised values of F1, F2, F3) between the vowels perceived as lax and the most reduced level can be found, and the reduced versions of [y:] differ according to their duration (mean value), whereas there are no significant differences between both reduction levels for [ø:]. The latter accounts for the low agreement between listeners' judgements. On the other hand, the lax vowels without reliably perceived reduction levels [ɛ], [ɔ], and [œ] show no significant differences according to their spectral properties from the reduced tense vowels associated with lax vowel quality. Only for [ʏ] can differences (F1, F3) be established. Furthermore, the spectral properties of [ɛ], [œ], and [ʏ] do not differ from those of the reduced tense vowels associated with centralised vowel quality, but [ɔ] does show a difference here with respect to F2 values.

This analysis indicates that spectral distances between reduced tense vowels perceived as lax and tense vowels associated with a schwa-like quality are greater than those within the group of lax vowels. The differences between reduced (lax-like) tense vowels and unreduced lax vowels are not perceptually significant. Therefore, lax vowels can be regarded as reduced counterparts of tense vowels.

The labelling of the reduction level of [i:] and of [e:] indicates that listeners discriminate between a long and short /i/ and /e/. However, both reduction levels differ in duration as well as in their spectral properties, so that the lengthening can be interpreted in terms of tenseness. This might account for the great distance to their counterparts, i.e. [ɪ] is closer to [e:] than to [i:] (cf. Figure 18.5). One reduction level of [e:] is associated with [ɪ].

In conclusion, then, the reduction of vowels can be considered as centralisation. Its perception is affected by the vowel and its distance to stronger centralised vowel qualities as well as to the 'schwa'. Preliminary studies indicate that the strength of reduction correlates with different prosodic factors (i.e. pitch accent, perceived prominence; Widera and Portele, 1999). However, further work is required to examine vowel reduction in different speaking styles. Spontaneous speech is thought to be characterised by stronger vowel reduction. One question we have to address is whether these reduction levels are sufficient to describe vowel reduction in spontaneous speech.

Because of the relation of vowel reduction and prosody, vowel reduction is highly relevant to speech synthesis. A multi-level approach allows a classification of units of a speech synthesis system with respect to vowel quality and strength of reduction. The levels can be related to prosodic parameters of the system.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG) under grant HE 1019/9–1. It was presented at the COST 258 meeting in Budapest 1999. I would like to thank all participants for fruitful discussions and helpful advice.

References

Aylett, M. and Turk, A. (1998). Vowel quality in spontaneous speech: What makes a good vowel? [Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm]. *Proceedings of the 5th International Conference on Spoken Language Processing* (Paper 824). Sydney, Australia.

Eskénazi, M. (1993). Trends in speaking styles research. *Proceedings of Eurospeech*, *1* (pp. 501–509). Berlin.

ESPS 5.0 [Computer software]. (1993). Entropic Research Laboratory, Washington.

Heuft, B., Portele, T., Höfer, F., Krämer, J., Meyer, H., Rauth, M., and Sonntag, G. (1995). Parametric description of F0-contours in a prosodic database. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, *2* (pp. 378–381). Stockholm.

Kohler, K.J. (1995a). *Einführung in die Phonetik des Deutschen* (2nd edn). Erich Schmidt Verlag.

Kohler, K.J. (1995b). Articulatory reduction in different speaking styles. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, *1* (pp. 12–19). Stockholm.

Lienert, G.A. and Raats, U. (1994). *Testaufbau und Testananlyse* (5th edn). Psychologie Verlags Union.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*, 1773–1781.

Pols, L.C.W., van der Kamp, L.J.T., and Plomp, R. (1969). Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America*, *46*, 458–467.

Stöber, K.-H. (1997). Unpublished software.

van Bergem, D.R. (1995). Perceptual and acoustic aspects of lexical vowel reduction, a sound change in progress. *Speech Communication*, *16*, 329–358.

Wells, J.C. (1996). SAMPA – computer readable phonetic alphabet. Available at: http://www.phon.ucl.ac.uk/home/sampa/german.htm.

Widera, C. and Portele, T. (1999). Levels of reduction for German tense vowel. *Proceedings of Eurospeech*, *4* (pp. 1695–1698). Rhodes, Greece.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Part III

Issues in Styles of Speech

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

19

Variability and Speaking Styles in Speech Synthesis

Jacques Terken
Technische Universiteit Eindhoven
IPO, Center for User-System Interaction
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
j.m.b.terken@tue.nl

Introduction

Traditional applications in the field of speech synthesis are mainly in the field of text-to-speech conversion. A characteristic feature of these systems is the lack of possibilities for variation. For instance, one may choose from a limited number of voices, and for each individual voice only a few parameters may be varied. With the rise of concatenative synthesis, where utterances are built from fragments that are taken from natural speech recordings that are stored in a database, the possibilities for variation have further decreased. For instance, the only way to get convincing variation in voice is by recording multiple databases. More possibilities for variation are provided by experimental systems for parametric synthesis, which allow researchers to manipulate up to 50 parameters for research purposes, but knowledge about how to synthesise different speaking styles has been lacking.

Progress both in the domains of language and speech technology and of computer technology has given rise to the emergence of new types of applications including speech output, such as multimedia applications, tutoring systems, animated characters or embodied conversational agents, and dialogue systems. One of the consequences of this development has been an increased need for possibilities for variation in speech synthesis as an essential condition for meeting quality requirements.

Within the speech research community, the issue of speaking styles has raised interest because it addresses central issues in the domain of speech communication and speech synthesis. We only have to point to several events in the last decade, witnessing the increased interest in speaking styles and variation both in the speech recognition and the speech synthesis communities:

- The ESCA workshop on the Phonetics and Phonology of Speaking Styles, Barcelona (Spain) 1991;

- The recent ISCA workshop on Speech and Emotion, Newcastle (Northern Ireland) 2000;
- Similarly, the COST 258 Action on 'Naturalness of Synthetic Speech' has designated the topic of speaking styles to one of the main action lines in the area of speech synthesis.

Obviously, the issue of variability and speaking styles can be studied from many different angles. However, prosody was chosen as the focus in the COST 258 action line on speaking styles because it seems to constitute a principal means for achieving variation in speaking style in speech synthesis.

Elements of a Definition of 'Speaking Styles'

Before turning towards the research contributions in this part we need to address the question of what constitutes a speaking style. The notion of speaking style is closely related to that of variability, although the notion of variability is somewhat broader. For instance, variability may also include diachronic variation and differences between closely related languages. Looking at the literature, there appears to be no agreed upon definition or theoretical framework for classifying speaking styles, if there is a definition at all. Instead, we see that authors just mention a couple a speaking styles that they want to investigate. For instance, Bladon, Carlson, Granström, Hunnicutt and Karlson (1987) link speaking styles to the casual-formal dimenson. Abe (1997) studies speaking styles for a literary novel, an advertisement and an encyclopedia paragraph. Higuchi, Hirai and Sagisaka (1997) study hurried, angry and gentle style in contrast with unmarked style (speaking free of instruction). Finally, the *Handbook on Standards and Resources for Spoken Language Systems* (Gibbon *et al.*, 1997) mentions speaking styles such as read speech and several kinds of spontaneous speech; elsewhere, it links the notion of speaking style directly to observable properties such as speaking rate and voice height. Apparently, authors hesitate to give a characterisation of what they mean by speaking style.

An analogy may be helpful. In the domain of furniture, a style, e.g. the Victorian style, consists of a set of formal, i.e., observable characteristics by which experts may identify a particular piece of furniture as belonging to a particular period and distinct from pieces belonging to different periods ('formal' is used here in the sense of 'concerning the observable form'). The style embodies a set of ideas of the designer about the way things should look like. Generalising these considerations, we may say that a style contains a descriptive aspect ('what are the formal characteristics') and an explanatory aspect ('the explanation as to why this combination of observable properties makes a good style'). Both aspects also take on a normative character: the descriptive aspect specifies the observable properties that an object should have to be considered as an instantiation of the particular style; the explanatory aspect defines the aesthetic value: objects or collections of objects that do not exhibit particular combinations of observable properties are considered to have low aesthetic value.

When we apply these considerations to the notion of style in speech, we may say that a speaking style consists of a set of observable properties by which we may

identify particular speaking behaviour as tuned to a particular communicative situation. The descriptive aspect concerns the observable properties that make different samples of speech to be perceived as representing distinct speaking styles. The explanatory aspect concerns the appropriateness of the manner of speaking in a particular communicative situation: a particular speaking style may be appropriate in one situation but completely inappropriate in another one.

The communicative situation to which the speaker tunes his speech and by virtue of which these formal characteristics will differ may be characterised in terms of at least three dimensions: the content, the speaker and the communicative context.

- With respect to the content, variation in speaking style may arise due to the content that has to be transmitted (e.g., isolated words, numerals or texts) and the source of the materials: is it spontaneously produced, rehearsed or read aloud?
- With respect to the speaker, variation in speaking style may arise due to the emotional-attitudinal state of the speaker. Furthermore, speaker habits and the speaker's personality may affect the manner of speaking. Finally, language communities may encourage particular speaking styles. Well-known opposites are the dominant male speaking style of Southern California and the submissive speaking style of Japanese female speakers.
- With respect to the situation, we may draw a distinction between the external situation and the communicative situation. The external situation concerns factors such as the presence of loud noise, the need for confidentiality, the size of the audience and the room. These factors may give rise to Lombard speech or whispered speech. The communicative situation has to do with factors such as monologue versus dialogue (including turn-taking relations), error correction utterances in dialogue versus default dialogue behaviour, rhetorical effects (convince/persuade, inform, enchant, hypnotise, and so on) and listener characteristics, including the power relations between speaker and listener (in most cultures different speaking styles are appropriate for speaking to peers and superiors).

From these considerations, we see that speaking style is essentially a multidimensional phenomenon, while most studies address only a select range of one or a few of these dimensions. Admittedly, not all combinations of factors make sense and certainly the different dimensions are not completely independent. Thus, a considerable amount of work needs to be done to make this framework more solid. However, in order to get a full understanding of the phenomenon of speaking styles we need to relate the formal characteristics of speaking styles to these or similar dimensions. One outcome of this exercise would be that we are able to predict which will be appropriate prosodic characteristics for speech in a particular situation even if the speaking style has not been studied yet.

Guide to the Chapters

The chapters in this section present research on variability and speaking styles that was done in the framework of the COST 258 action on Naturalness of Synthetic Speech, relating to the framework introduced above in various ways.

The chapter by López Gonzalo,Villar Navarro and Hernández Córtez addresses the notion of variability in connection with differences between dialects/languages. It describes an approach to the problem of obtaining a prosodic model for a particular target language and poses the question whether combining this model with the segmental synthesis for a closely related language will give acceptable synthesis of the 'accent' of the target language. They find that perception of 'accent' is strongly influenced by the segmental properties, and conclude that acceptable synthesis of the 'accent' for the target language quite likely requires access to the segmental properties of the target language as well.

Five chapters address the prosodic characters of particular speaking styles. Duez investigates segmental reduction and assimilation in conversational speech and discusses the implications for rule-based synthesisers and concatenative approaches in terms of the knowledge that needs to be incorporated in these systems. Zei Pollermann and Archinard, Ní Chasaide and Gobl, Gustafson and House, and Montero, Gutiérrez-Arriola, de Cordoba, Enríquez and Pardo all investigate affective speaking styles. Ní Chasaide and Gobl and Gustafson and House apply an analysis-by-synthesis methodology to determine settings of prosodic parameters that elicit judgements of particular emotions or affective states. Ní Chasaide and Gobl study the relative contributions of pitch and voice quality for different emotions and affective states. Gustafson and House concentrate on one particular speaking style, and aim to find parameter settings for synthetic speech that will make an animated character being perceived as funny by children. Zei Pollermann and Archinard, and Montero, Gutiérrez-Arriola, de Cordoba, Enríquez and Pardo investigate the prosodic characteristics of 'basic' emotions. Both Ní Chasaide and Gobl, Zei Pollermann and Archinard, and Montero, Gutiérrez-Arriola, de Cordoba, Enríquez and Pardo provide evidence that the usual focus on pitch and temporal properties will lead to limited success in the synthesis of the different emotions. Certainly, variation that relates to voice source characteristics needs to be taken into consideration to be successful.

Whereas all the chapters above focus on the relation between prosodic characteristics of speaking styles and communicative dimensions, three further chapters focus on issues in the domain of linguistic theory and measurement methodology. Such studies tend to make their observations in controlled environments or laboratories, and with controlled materials and specific instructions to trigger particular speaking styles directly. Gobl and Ní Chasaide present a brief overview of work on the modelling of glottal source dynamics and discuss the relevance of glottal source variation for speech synthesis. Zellner-Keller and Keller and Monaghan instruct speakers to speak fast or slow, in order to get variation of formal characteristics beyond what is obtained in normal communicative situations and to get a clearer view of the relevant parameters. This research sheds light on the question how prosody is restructured if a speaker changes the speaking rate. These findings are directly relevant to the question of how prosodic structure can be represented such that prosodic restructuring can be easily and elegantly accounted for and modelled in synthesis.

References

Abe, M. (1997). Speaking styles: Statistical analysis and synthesis by a text-to-speech system, In J. Santen, R. Sproat, J. Olive, and J. Hirschberg (eds), *Progress in Speech Synthesis* (pp. 495–510). Springer-Verlag.

Bladon, A., Carlson, R., Granström, B., Hunnicutt, S., and Karlsson, I. (1987). A text-to-speech system for British English, and issues of dialect and style. In J. Laver and M. Jack (eds), *European Conference on Speech Technology*, Vol. I (pp. 55–58). Edinburgh: CEP Consultants.

Cowie, R., Douglas-Cowie, E., and Schröder, M. (eds) (2000). Speech and emotion: A conceptual Framework for Research. *Proceedings of the ISCA workshop on Speech and Emotion.* Belfast: Textflow.

Gibbon, D., Moore, R., and Winski, R. (eds) (1997). *Handbook on Standards and Resources for Spoken Language Systems*. Mouton De Gruyter.

Higuchi, N., Hirai, T., and Sagisaka, Y. (1997). Effect of speaking style on parameters of fundamental frequency contour. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (eds), *Progress in Speech Synthesis* (pp. 417–427). Springer-Verlag.

Llisteri, J. and Poch, D. (eds) (1991). *Proceedings of the ESCA workshop on the Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication*. Barcelona: Universitad Autonoma de Barcelona.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

20

An Auditory Analysis of the Prosody of Fast and Slow Speech Styles in English, Dutch and German

Alex Monaghan
Aculab Plc, Lakeside, Bramley Road
Mount Farm, Milton Keynes MK1 1PT, UK
Alex.Monaghan@aculab.com

Introduction

In April 1999, a multilingual speech database was recorded as part of the COST 258 work programme. This database comprised read text from a variety of genres, recorded by speakers of several different European languages. The texts obviously differed for each language, but the genres and reading styles were intended to be the same across all language varieties. The objective was to obtain comparable data for different styles of speech across a range of languages. More information about these recordings is available from the COST 258 web pages.[1] One component of this database was the recording of a passage of text by each speaker at two different speech rates. Speakers were instructed to read first slowly, then quickly, and were given time to familiarise themselves with the text beforehand. The resulting fast and slow versions from six speakers provided the data for the present study.

Speech in English, Dutch, and four varieties of German was transcribed for accent location, boundary location and boundary strength. Results show a wide range of variation in the use of these aspects of prosody to distinguish fast and slow speech, but also a surprising degree of consistency within and across languages.

[1] http://www.unil.ch/imm/docs/LAIP/COST_258/

Methodology

The analysis reported here was purely auditory. No acoustic measurements were made, no visual inspection of the waveforms was performed. The procedure involved listening to the recordings on CD-ROM, through headphones plugged directly into a PC, and transcribing prosody by adding diacritics to the written text. The transcriber was a native speaker of British English, with near-native competence in German and some knowledge of Dutch, who is also a trained phonetician and a specialist in the prosody of the Germanic languages.

Twelve waveforms were analysed, corresponding to fast and slow versions of the same text as read by native speakers of English, Dutch, and four standard varieties of German (as spoken in Bonn, Leipzig, Austria and Switzerland: referred to below as GermanB, GermanL, GermanA and GermanS, respectively). There were five different texts: the texts for the Leipzig and Swiss speakers were identical. There was one speaker for each language variety. The English, Austrian and Swiss speakers were male: the other three were not.

Three aspects of prosody were chosen as being readily transcribable using this methodology:-

- accent location
- boundary location
- boundary strength

Accent location in the present study was assessed on a word-by-word basis. There were a few cases in the Dutch speech where compound words appeared to have more than one accent, but these were ignored in the analysis presented here: future work will examine these cases more closely.

Boundary location in this data corresponds to the location of well-formed prosodic boundaries between intonation phrases. As this is fluent read speech, there are no hesitations or other spurious boundaries.

Boundary strength was transcribed according to three categories:

- major pause (Utt)
- minor pause (IP)
- boundary tone, no pause (T)

The distinction between major and minor pauses here corresponds intuitively to the distinction between inter-utterance and intra-utterance boundaries, hence the label Utt for the former. In many text-to-speech synthesisers, this would be the difference between the pause associated with a comma in the text and that associated with a sentence boundary. However, at different speech rates the relations between pausing and sentence boundaries can change (see below), so a more neutral set of labels is required. Unfortunately, the aspiring ToBI labelling standard[2] does not label boundaries above the intonation phrase and makes no mention of pausing:

[2] http://ling.ohio-state.edu/phonetics/E_ToBI

while all our T boundaries would correspond to ToBI break index 4, not all 4s would correspond to our Ts since a break index of 4 may be accompanied by a pause in the ToBI system. We have thus chosen to use the label T to denote an intonational phrase boundary marked by tonal features but with no pause, and the label IP to denote the co-occurrence of an intonational phrase boundary with a short pause. We assume that there is a hierarchy of intonational phrase boundaries, with T being the lowest and Utt being the highest in our present study.

There was no attempt made to transcribe different degrees of accent strength or different accent contours in the present study, for two reasons. First, different theories of prosody allow for very different numbers of distinctions of accent strength and contour, ranging from two (e.g. Crystal, 1969) to infinity (e.g. Bolinger, 1986; Terken, 1997). Second, there was no clear auditory evidence of any systematic use of such distinctions by speakers to distinguish between fast and slow speech, with the exception of an increase in the use of linking or 'flat hat' contours (see 't Hart *et al.*, (1990); Ladd (1996)) in fast speech: this tendency too will be investigated in future analyses.

The language varieties for which slow and fast versions were analysed, and the sex of the speaker for each, are given in Table 20.1.[3] As mentioned above, the text files for Leipzig German and Swiss German were identical: all others were different.

Results

General Characteristics

An examination of the crudest kind (Table 20.2) shows that the texts and readings were not as homogeneous as we had hoped. Text length varied from 35 words to 148 words, and although all texts were declarative and informative in style they ranged from weather reports (English) through general news stories to technical news items (GermanA). These textual differences seem to correlate with some prosodic aspects discussed below.

More importantly, the meaning of 'slow' and 'fast' seems to vary considerably: the proportional change (Fast/Slow) in the total duration of each text between the slow and fast versions varies from 25% to 45%. It is impossible to say whether this variation is entirely due to the interpretation of 'slow' and 'fast' by the different speakers, or whether the text type plays a role: text type cannot be the whole story,

Table 20.1 Language varieties and sexes of the six speakers

Speakers	
English (M)	Dutch (F)
Austrian German (M)	Bonn German (F)
Swiss German (M)	Leipzig German (F)

[3] The texts and transcription files for all six varieties are available on the accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm., or from http://www.compapp.dcu.ie/~alex/cost258.html

Table 20.2 Length in words, and duration to the nearest half second, of the six fast and slow versions

	Text length			
	Words	Fast	Slow	Fast/Slow
English	35	11.5s	17.0s	0.68
Dutch	75	24.0s	42.0s	0.57
GermanA	148	54.5s	73.0s	0.75
GermanB	78	28.0s	51.0s	0.55
GermanL	63	27.0s	49.0s	0.55
GermanS	63	25.5s	38.5s	0.66

however, as the same text produced different rate modifications for GermanL (45%) and GermanS (34%). The questions of the meaning of 'fast' and 'slow', and of whether these are categories or simply points on a continuum, are interesting ones but will not be addressed here.

Accents

Table 20.3 shows the numbers of accents transcribed in the fast and slow versions for each language variety. Although there are never more accents in the fast version than in the slow version, the overlap ranges from 100% to 68%. This is a true overlap, as all accent locations in the fast version are also accent locations in the slow version: in other words, nothing is accented in the fast version unless it is accented in the slow version. Fast speech can therefore be characterised as a case of accent deletion, as suggested in our previous work (Monaghan, 1990; 1991a; 1991b). However, the amount of deletion varies considerably, even within the same text (68% for GermanL, 92% for GermanS): thus, it seems likely that speakers apply different amounts of deletion either as a personal characteristic or as a result of differing interpretations of slow and fast. This variation does not appear to correlate with the figures in Table 20.2 for overall text durations: in the cases of GermanB and GermanL in particular, the figures are very similar in Table 20.2 but quite different in Table 20.3.

Table 20.3 Numbers of accents transcribed, and the overlap between accents in the two versions for each language variety

	Accent location		
	Fast	Slow	Overlap
English	21	21	21 (100%)
Dutch	34	43	34 (79%)
GermanA	74	78	74 (95%)
GermanB	35	42	35 (83%)
GermanL	28	41	28 (68%)
GermanS	33	36	33 (92%)

The case of the English text needs some comment here. This text is a short summary of a weather forecast, and as such it contains little or no redundant information. It is therefore very difficult to find any deletable accents even at a fast speech rate. However, it should not be taken as evidence against accent deletion at faster speech rates: as always, accents are primarily determined by the information content of the text and therefore may not be candidates for deletion in certain cases (Monaghan, 1991a; 1993).[4]

Boundaries

Table 20.4 shows the numbers and types of boundary transcribed. As with accents, the number of boundaries increases from fast to slow speech but there is a great deal of variation in the extent of the increase. The total increase ranges from 30% (GermanA) to 230% (GermanL). All types of boundary are more numerous in the slow version, with the exception of IPs in the case of GermanS. There is a large amount of variation between GermanL and GermanS, with the possible exception of the Utt boundaries: the two Utt boundaries in the fast version of GermanL are within the title of the text, and may therefore be indicative of a different speech rate or style. If we reclassify those two Utt boundaries as IP boundaries[5], then there is

Table 20.4 Correspondence between boundary location and textual punctuation, broken down by boundary category.

	Boundary categories			
Fast	Utt	IP	T	All
English	0	3	5	8
Dutch	0	7	3	10
GermanA	0	13	9	22
GermanB	4	5	3	12
GermanL	*0	*5	2	7
GermanS	0	7	1	8
Subtotal	4	40	23	67
Slow	Utt	IP	T	All
English	3	5	3	11
Dutch	6	8	7	21
GermanA	5	14	10	29
GermanB	6	16	14	36
GermanL	5	10	8	23
GermanS	5	6	4	15
Subtotal	30	59	46	135
TOTAL	34	99	69	202

Note: The figures marked * result from reclassifying the two boundaries in the title of GermanL.

[4] The English text is also problematic as regards the relation between boundaries and punctuation discussed below.

[5] These two boundaries have been reclassified as IP boundaries in the fast version of GermanL for all subsequent tables.

some evidence (GermanL, GermanS, English, Dutch) of correspondence between slow Utt boundaries and fast IP boundaries. In order to investigate this point further, we analysed the demotion and deletion of boundaries: the results are presented in Tables 20.5 and 20.6.

Table 20.5 shows the changes in boundary strength location between the slow and fast speech versions, for every boundary location in the slow versions. The most obvious finding is that the last two columns in Table 20.5 are empty, indicating that there are no increases in boundary strength: this means that no new boundary locations appear in the fast versions, and that no boundaries are stronger in the fast version than in the slow version. Again, this is in line with our previous rules for faster speech rates (Monaghan, 1990; 1991a; 1991b), where fast speech involves the demotion or deletion of boundaries.

Table 20.5 shows a strong tendency for boundaries to be reduced by one category in the change from slow to fast speech (Utt becomes IP, IP becomes T, T is deleted), and a secondary tendency to reduce boundaries by two categories. GermanA is an exception to this tendency, with a secondary tendency to leave boundary strengths unchanged: this may be related to the relatively small overall duration difference for GermanA in Table 20.2.

Table 20.6 shows the correspondence between punctuation in the texts and boundary strength. The details vary considerably, depending on the text. For the GermanA text, the pattern is complex. Boundaries are demoted or deleted in the change from slow to fast speech, but the precise mechanism is unclear.

For the other varieties of German, a pattern emerges. In slow speech, Utt boundaries correlate well with period punctuation, IPs coincide with commas, but T boundaries are apparently not related to punctuation. In fast speech, boundaries are regularly demoted by one category so that almost all boundaries correspond to punctuation. The apparent unpredictability of T boundaries in slow speech could be attributed to rhythmic factors, stylistic constraints on foot length, the need to breathe, or other factors: this remains to be investigated. In fast speech, however, where Ts are rare, boundaries appear to be predictable.

The English and Dutch data present an even clearer picture. In these texts, boundaries in the fast version are predictable from the slow version. In the Dutch

Table 20.5 Changes in categories of boundary, for each boundary location present in the slow versions

	Boundary strength: Changes –Slow to Fast				
	−2	−1	0	+1	+2
English	2	7	2	0	0
Dutch	5	14	2	0	0
GermanA	2	15	12	0	0
GermanB	11	17	8	0	0
GermanL	8	15	0	0	0
GermanS	3	10	2	0	0

Table 20.6 Correspondence between boundary location and textual punctuation.

Boundaries: correspondence with punctuation												
Punctuation = Boundary	English		Dutch		GermanA		GermanB		GermanL		GermanS	
	Fast	Slow	Fast	Slow	Fast	Slow	Fast	Slow	Fast	Slow	Fast	Slow
"." = Utt		2		5		4	4	6		5		5
"." = IP	2		4	*1	9	5	1		5		5	
"." = T			*2				1					
"." = Nil	0	0	0	0	0	0	0	0	0	0	0	0
"-"/":" = Utt				1		1						
"-"/":" = IP			1		1	3						
"-"/":" = T					3							
"-"/":" = Nil	0	0	0	0	0	0	0	0	0	0	0	0
"," = Utt												
"," = IP			2	2	2	2	2	4		2	2	2
"," = T				2	4	1		1				
"," = Nil	0	0	0	0	0	0	0	0	0	0	0	0
None = Utt		1										
None = IP	1	5		5	1	4	2	12			8	4
None = T	5	3	1	7	4	6	1	14	1	8	1	4

Note: The figures marked * include a period punctuation mark which occurs in the middle of a direct quotation

text, fast speech boundaries are also predictable from the punctuation. We suggest that the reason this is not true of the English text is that this text is not sufficiently well punctuated: it consists of 35 words and two periods, with no other punctuation, and this lack of explicit punctuation means that in both fast and slow versions the speaker has felt obliged to insert boundaries at plausible comma locations. The location of commas in English text is largely a matter of choice, so that under- or over-punctuated texts can easily occur: the question of optimal punctuation for synthesis is beyond the scope of this study, but should be considered by those preparing text for synthesisers. One predictable consequence of over- or under-punctuation would be a poorer correspondence between prosodic boundaries and punctuation marks.

There is a risk of circularity here, since a well-punctuated text is one where the punctuation occurs at the optimal prosodic boundaries. However, it seems reasonable to assume an independent level of text structure which is realised by punctuation in the written form and by prosodic boundaries in the spoken form. Given this assumption, a well-punctuated text is one in which the punctuation accurately and adequately reflects the text structure. By the same assumption, a good reading of the text is one in which this text structure is accurately and adequately expressed in the prosody. For synthesis purposes, we must generally assume that the punctuation is appropriate to the text structure since no independent analysis of that structure is available: poorly punctuated texts will therefore result in sub-optimal prosody.

Detailed Comparison

A final analysis compared the detailed location of accents and boundaries in GermanL and GermanS, two versions of the same text. We looked at the locations of accents and of the three categories of boundary, and in each case we took the language variety with the smaller total of occurrences and noted the overlap between those occurrences and the same locations in the other language variety.

As Table 20.7 shows, the degree of similarity is almost 100% for both speech rates: the notable exception is the case of T boundaries, which seem to be assigned at the whim of the speaker. For all other categories, however, the two speakers agree completely on their location. (Note: we have reclassified the Utt boundaries around the title in GermanL.)

If T boundaries are optional, and therefore need not be assigned by rule, then it appears from Table 20.7 that accents and boundaries are predictable from text and that, moreover, boundary locations and strengths are predictable from punctuation. It also appears that speakers agree (for a given text) on which accents will be deleted and which boundaries will be demoted at a faster speech rate.

The differences between the four versions of the text in Table 20.7 could be explained as being the result of at least three different speech rates. The slowest, with the largest number of accents and boundaries, would be GermanL 'slow'. The next slowest would be GermanS 'slow'. The two 'fast' versions appear to be more similar, both in duration and in numbers of accents and boundaries. If this explanation were correct, we could hope to identify several different prosodic realisations of a given text ranging from a maximally slow version (with the largest possible number of accents and boundaries) to a maximally fast version via several stages of deletion of accents and boundaries.

Discussion

Variability

There are several aspects of this data which show a large amount of variation. First, there is the issue of the meaning of 'fast' and 'slow': the six speakers here differed greatly in the overall change in duration (Table 20.2), the degree of accent deletion (Table 20.3), and the extent of deletion or demotion of boundaries (Table 20.4) between their slow and fast versions. This could be attributed to speaker

Table 20.7 Comparison of overlapping accent and boundary locations in an identical text read by two different speakers

	Overlap – GermanL-GermanS			
	Accents	Utt	IP	T
Fast	27/28	0/0	5/5	0/1
Slow	36/36	5/5	6/6	2/4

differences, to differences between language varieties, or to the individual texts. The changes in boundary strength appear to be consistent for the same text (Table 20.5), and indeed the general tendencies seem to apply to all speakers and language varieties.

There is some variation in the mapping from punctuation marks to boundary categories, particularly in the case of the GermanA data. In both fast and slow versions of the GermanA text, IP boundaries were associated with all classes of punctuation (periods, commas, other punctuation, and no punctuation). However, in all other language varieties there was a very regular mapping from punctuation to prosodic boundaries: those irregularities which occurred, such as the realisation of the period boundary in a direct quotation in the Dutch text, can probably be explained by other factors of text structure.

The only remaining unexplained variation is in the number and location of T boundaries. In Table 20.6 these usually occur in the absence of punctuation but are occasionally aligned with all classes of punctuation marks. Again in Table 20.7, these are the only aspects of prosody on which the two speakers do not agree. It was suggested above that T boundaries may be optional and therefore need not be assigned by a synthesis system: however, since they constitute more than a third of the total boundaries, it seems unlikely that they could be omitted without affecting the perceived characteristics of the speech. A more promising strategy might be to assume that they are assigned according to rather variable factors, and that therefore their placement is less critical than that of higher level boundaries: heuristic rules might be suggested on the basis of factors such as maximal or optimal numbers of syllables (accented, stressed or otherwise) between boundaries, following the proposals of Gee and Grosjean (1983) or Keller and Zellner (1998). Again, an exploration of this issue is beyond the scope of the present study.

Consistency

There are also several aspects of this data which show almost total consistency across language varieties. Table 20.3 shows a 100% preservation of accents between the fast and slow versions, i.e. all accent locations in the fast version also receive an accent in the slow version. There is also a close to 100% result for an increase in the number of boundaries in every category when changing from fast to slow speech (Table 20.4), and a strong tendency to increase the strength of each individual boundary at the same time (Table 20.5). These findings are consistent with our previous rules for synthesising different speech rates by manipulating prosodic accent and boundary characteristics (Monaghan, 1991a; 1991b).

The alignment of boundaries with punctuation, and the demotion of these boundaries by one category when changing from slow to fast speech, is also highly consistent as shown in Table 20.6. Indeed, for most of the language varieties studied here, the strength and location of boundaries in fast speech is completely predictable from the slow version. Moreover, for a given text the location of accents and boundaries seems to be almost completely consistent across speakers at a particular speech rate (Table 20.7). The only unpredictable aspect is the location of T boundaries, as discussed above.

The consistency across the six language varieties represented here is surprisingly high. Although all six are from the same sub-family of the Germanic languages, we would have expected to see much larger differences than in fact occurred. The fact that GermanA is an exception to many of the global tendencies noted above is probably attributable to the nature of the text rather than to peculiarities of Standard Austrian German: this is a lengthy and quite technical passage, with several unusual vocabulary items and a rather complex text structure.

One aspect which has not been discussed above is the difference between the data for the three male speakers and the three female speakers. Although there is no conclusive evidence of differences in this quite superficial analysis, there are certainly tendencies which distinguish male speakers from the females. Table 20.2 shows that the female speakers (Dutch, GermanB and GermanL) have a consistently greater difference in total duration between fast and slow versions, and Table 20.3 shows a similarly consistent tendency for the female speakers to delete more accents in the change from slow to fast. Tables 20.4 to 20.6 show the same tendency for a greater difference between fast and slow versions among female speakers (more boundary deletions and demotions), in particular a much larger number of T boundaries in the slow versions.

In contrast, Table 20.7 shows almost uniform results for a male and a female speaker reading the same text. However, when we look at Table 20.7 we must remember that the results are for overlap between versions rather than identity: thus, GermanS has many fewer IP boundaries than GermanL in the slow versions but the overlap with the locations in the GermanS version is total.

The most obvious superficial explanation for these differences and similarities between male and female speakers appears to be that female 'slow' versions are slower than male 'slow' versions. The data for the fast versions for GermanL and GermanS are quite similar, especially if we reclassify the two Utt boundaries in the fast version of GermanL. This explanation builds on the suggestion above that there is a range of possible speech rates for a given text and that speakers agree on the prosodic characteristics of a specific speech rate. It also suggests an explanation for the unpredictability of T boundaries and their apparently optional nature: the large number of T boundaries produced by female speakers in the slow versions is attributable to the extra slow speech rate, and these boundaries are not required at the less slow speech rate used by the male speakers.

Conclusion

This is clearly a small and preliminary investigation of the relation between prosody and speech rate. However, several tentative conclusions can be drawn about the production of accents and boundaries in this data, and these are listed below. Since the object of this investigation was to characterise fast and slow speech prosody, some suggestions are also given as to how these speech rates might be synthesised.

Accents

For a given text and speech rate, speakers agree on the location of accents (Table 20.7). Accent location is therefore predictable, and its prediction does not require telepathy, but the factors which govern it are still well beyond the capabilities of automatic text analysis (Monaghan, 1993).

At faster speech rates, accents are progressively deleted (Table 20.3). This is again similar to our proposals (Monaghan, 1990; 1991a; 1991b) for automatic accent and boundary location at different speech rates: these proposals also included the progressive deletion and/or demotion of prosodic boundaries at faster speech rates (see below). It is not clear how many different speech rates are distinguishable on the basis of accent location, but from the figures for GermanL and GermanS in Tables 20.3 and 20.7 it seems that if speech rate is categorial then there are at least three categories.

Boundaries

For a given text and speech rate, speakers agree on the location and strength of Utt and IP boundaries (Table 20.7). In fast speech, these boundaries seem to be predictable on the basis of punctuation marks (Table 20.6).

Boundaries of all types are more numerous at slower speech rates (Table 20.4). They are regularly demoted at faster speech rates (Tables 20.5 and 20.6), which is once again consistent with our previous proposals (Monaghan, 1991a; 1991b).

T boundaries do not appear to be predictable from punctuation (Table 20.6), but appear to characterise slow speech rates. They may therefore be important to the perception of speech rate, but must be predicted on the basis of factors not yet investigated.

Fast and Slow Speech

The main objective of the COST 258 recordings, and of the present analysis, was to improve the characterisation of different speech styles for synthetic speech output systems. In the ideal case, the results of this study would include the formulation of rules for the generation of fast and slow speech styles automatically.

We can certainly characterise the prosody of fast and slow speech based on the observations above, and suggest rules accordingly. There are, however, two non-trivial obstacles to the implementation of these rules for any particular synthesis system. The first obstacle is that the rules refer to categories (e.g. Utt, IP, T) which may not be explicit in all systems and which may not be readily manipulated: in a system which assigns a minor pause every dozen or so syllables, for instance, it is not obvious how this strategy should be modified for a different speech rate. The second obstacle is that systems' default speech rates probably differ, and an important first step for any speech output system is to ascertain where the default speech rate is located on the fast-slow scale: this is not a simple matter, since the details of accent placement heuristics and duration rules amongst other things will affect the perceived speech rate. Assuming that these obstacles can be overcome,

the following characterisations of fast and slow speech prosody should allow most speech synthesis systems to implement two different speech rates for their output.

Fast speech is characterised by the deletion of accents, and the deletion or demotion of prosodic boundaries. Major boundaries in fast speech seem to be predictable from textual punctuation marks, or alternatively from the boundaries assigned at a slow speech rate. T boundaries may be optional, or may be based on rhythmic or metrical factors.

Accent deletion and the demotion/deletion of boundaries operate in a manner similar to that proposed in Monaghan (1990; 1991a; 1991b). Unfortunately, as discussed in Monaghan (1993), accent location is not predictable without reference to factors such as salience, givenness and speaker's intention: however, given an initial over-assignment of accents as specified in Monaghan (1990), their deletion appears to be quite simple.

Slow speech is characterised by the insertion of pauses (Utt and IP boundaries) at punctuation marks in the text, and by the placement of non-pause boundaries (T boundaries in our model) based on factors which have not been determined for this data. At slow speech rates, as proposed in Monaghan (1990; 1991b), accents may be assigned to contextually salient items on the basis of word class information: this will result in an 'over-accented' representation which is similar to the slow versions in the present study.

Candidate heuristics for the assignment of T boundaries at slow speech rates would include upper and lower thresholds for the number of syllables, stresses or accents between boundaries (rhythmic criteria); correspondence with certain syntactic boundaries (structural criteria); and interactions with the relative salience of accented items such that, for instance, more salient items were followed by a T boundary and thus became nuclear in their IP (pragmatic criteria). Such heuristics have been successfully applied in the LAIPTTS system (Siebenhaar-Rölli *et al.*, Chapter 16, this volume; Keller and Zellner, 1998 and references therein) for breaking up unpunctuated stretches of text.

The rules proposed here are based on small samples of read speech, and may therefore require refinement particularly for other genres. Nonetheless, the tendencies in most respects are clear and universal for these language varieties.

The further investigation of T boundaries in the present data, and the subclassification of accents into types including the flat hat contour, are the next tasks in this programme of research. It would also be interesting to extend this analysis to larger samples, and to other languages. In a rather different study on Dutch only, Caspers (1994) found similar results for boundary deletion (including unpredictable T boundaries), but much less evidence of accent deletion. This suggests that not all speakers treat accents in the same way, or that Caspers' data was qualitatively different.

Conclusion

This study presents an auditory analysis of fast and slow read speech in English, Dutch, and four varieties of German. Its objective was to characterise the prosody of different speech rates, and to propose rules for the synthesis of fast and slow

speech based on this characterisation. The data analysed here are limited in size, being only about seven and a half minutes of speech (just under 1000 words) with only one speaker for each language variety. Nevertheless, there are clear tendencies which can form the basis of initial proposals for speech rate rules in synthesis systems.

The three aspects of prosody which were investigated in the present study (accent location, boundary location and boundary strength) show a high degree of consistency across languages at both fast and slow speech rates. There are reliable correlations between boundary location and textual punctuation, and for a given text and speech rate the location of accents and boundaries appears to be consistent across speakers.

The details of prosodic accent and boundary assignment in these data are very similar to our previous Rhythm Rule and speech rate heuristics (Monaghan, 1990; 1991b, respectively). Although the location of accents is a complex matter, their deletion at faster speech rates seems to be highly regular. The demotion or deletion of boundaries at faster speech rates appears to be equally regular, and their location in the data presented here is largely predictable from punctuation.

We hope that these results will provide inspiration for the implementation of different speech rates in many speech synthesis systems, at least for the Germanic languages. The validation and refinement of our proposals for synthetic speech output will require empirical testing in such automatic systems, as well as the examination of further natural speech data.

The purely auditory approach which we have taken in this study has several advantages, including speed, perceptual filtering and categoriality of judgements. Its results are extremely promising, and we intend to continue to apply auditory analysis in our future work. However, it obviously cannot produce all the results which the prosody rules of a synthesis system require: the measurement of minimum pause durations for different boundary strengths, for instance, is simply beyond the capacities of human auditory perception. We will therefore be complementing auditory analysis with instrumental measures in future studies.

References

Bolinger, D. (1986). *Intonation and its Parts*. Stanford University Press.

Caspers, J. (1994). *Pitch Movements Under Time Pressure*. Doctoral dissertation, Rijksuniversiteit Leiden.

Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge University Press.

Gee, J.P. and Grosjean, F. (1983). Performance structures. *Cognitive Psychology*, *15*, 411–458.

't Hart, J., Collier, R., and Cohen, A. (1990). *A Perceptual Study of Intonation*. Cambridge University Press.

Keller, E. and Zellner, B. (1998). Motivations for the prosodic predictive chain. *Proceedings of the 3rd ESCA Workshop on Speech Synthesis* (pp. 137–141). Jenolan Caves, Australia.

Ladd, D.R. (1996). *Intonational Phonology*. Cambridge University Press.

Monaghan, A.I.C. (1990). Rhythm and stress shift in speech synthesis. *Computer Speech and Language*, *4*, 71–78.

Monaghan, A.I.C. (1991a). *Intonation in a Text to Speech Conversion System*. PhD thesis, University of Edinburgh.

Monaghan, A.I.C. (1991b). Accentuation and speech rate in the CSTR TTS System. *Proceedings of the ESCA Research Workshop on Phonetics and Phonology of Speaking Styles* (pp. 411–415). Barcelona, September–October.

Monaghan, A.I.C. (1993). What determines accentuation? *Journal of Pragmatics*, *19*, 559–584.

Terken, J.M.B. (1997). Variation of accent prominence within the phrase: Models and spontaneous speech data. In Y. Sagisaka *et al.* (eds), *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 95–116). Springer-Verlag.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

21

Automatic Prosody Modelling of Galician and its Application to Spanish

Eduardo López Gonzalo, Juan M. Villar Navarro and Luis A. Hernández Gómez

Dep. Señales Sistemas y Radiocomunicaciones;
E.T.S.I. de Telecomunicación
Universidad Politécnica de Madrid
Ciudad Universitaria S/N.
28040 Madrid (Spain)
eduardo, juanma, luis @gaps.ssr.upm.es
http://www.gaps.ssr.upm.es/tts

Introduction

Nowadays, there are a number of multimedia applications that require accurate and specialised speech output. This fact is directly related to improvements in the area of prosodic modelling in text-to-speech (TTS) that make it possible to produce adequate speaking styles.

For a number of years, the construction and systematic statistical analysis of a prosodic database (see, for example, Emerard *et al.*, 1992, for French) have been used for prosodic modelling. In our previous research, we have worked on prosodic modelling (López-Gonzalo and Hernández-Gómez, 1994), by means of a statistical analysis of manually labelled data from a prosodic corpus recorded by a single speaker. This is a subjective, tedious and time-consuming work that must be redone every time a new voice or a new speaking style is generated.

Therefore, there was a need for more automatic methodologies for prosodic modelling that improve the efficiency of human labellers. For this reason, we proposed in López-Gonzalo and Hernández-Gómez (1995) an automatic data-driven methodology to model both fundamental frequency and segmental duration in TTS systems that captures all the characteristic features of the recorded speaker. Two major lines previously proposed in speech recognition were extended to automatic prosodic modelling of one speaker for Text-to-Speech: (a) the work described in Wightman and Ostendorf (1994) for automatic recognition of prosodic boundaries;

and (b) the work described in Shimodaira and Kimura (1992) for prosodic segmentation by pitch pattern clustering.

The prosodic model describes the relationship between some linguistic features extracted from the text and some prosodic features. Here, it is important to define a prosodic structure. In the case of Spanish, we have used a prosodic structure that considers syllables, accent groups (group of syllables with one lexical stress) and breath groups (group of accent groups between pauses). Once these prosodic features are determined, a diphone-based TTS system generates speech by concatenating some diphones with the appropriate prosodic properties.

This chapter presents an approach to cross-linguistic modelling of prosody for speech synthesis with two related, but different languages: Spanish and Galician. This topic is of importance in the European context of growing regional awareness. Results are provided on the adaptation of our automatic prosody modelling method to Galician. Our aim was twofold: on the one hand, we wanted to try our automatic methodology on a different language because it had only been tested for Spanish, on the other, we wanted to see the effect of applying the phonological and phonetic models obtained for the Galician corpus to Spanish. In this way, we expected to get the Galician accent when synthesising text in Spanish, combining the prosodic model obtained for Galician with the Spanish diphones. The interest of this approach lies in the fact that inhabitants of a region usually prefer a voice with its local accent, for example, Spanish with a Galician accent for a Galician inhabitant. This fact has been informally reported to us by a pedagogue specialising in teaching reading aloud (F. Sepúlveda). He has noted this fact in his many courses around Spain.

In this chapter, once the prosodic model was obtained for Galician, we will try two things:

- to generate Galician synthesising a Galician text with the Galician prosodic model and using the Spanish diphones for speech generation;
- to generate Spanish with a Galician accent synthesising a Spanish text with the Galician prosodic model and using the Spanish diphones for speech generation;

The outline of the rest of the chapter is as follows: first, we give a brief summary of the general methodology used, then, we report our work on the adaptation of the corpus; finally we summarise results and conclusions.

Automatic Prosodic Modelling System

The final aim of the method is to obtain a set of data that permits modelling the prosodic behaviour of a given prosodic corpus recorded by one speaker. The automatic method is based on another method developed by one of the authors in his PhD thesis (López-Gonzalo, 1993), which established a processing of prosody in three levels (named acoustic, phonetic and phonological). Basically the same assumptions are made by the automatic method at each level. An overview can be seen in Figure 21.1. The input is a recorded prosodic corpus and its textual representation. The analysis gives a database of syllabic prosodic patterns, and a set of

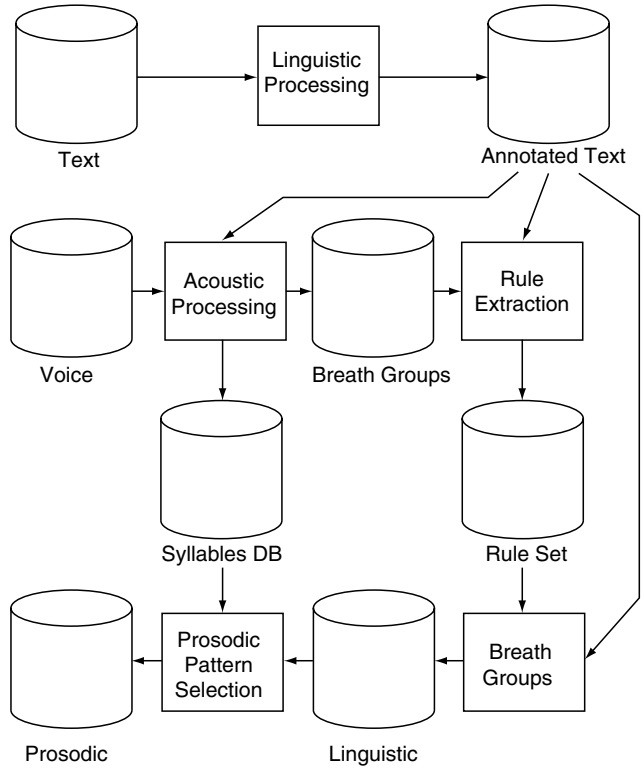


Figure 21.1 General overview of the methodology, both analysis and synthesis

rules for breath group classification. From this, the synthesis part is capable to assign prosodic patterns to a text.

Both methods perform a joint modelling for the fundamental frequency (F0) contour and segmental duration, both assign prosody on a syllable-by-syllable basis and both assign the actual F0 and duration values from a data-base. The difference lies in how the relevant data is obtained for each level: acoustical, phonological and phonetic.

Acoustic Analysis

From the acoustic analysis we obtain the segmental duration of each sound and the pitch contour, and then simplify it. The segmental duration of each sound is obtained in two steps, first a Hidden Markov Model (HMM) recognizer is employed in forced alignment, then a set of tests is performed on selected frontiers to eliminate errors and improve accuracy. The pitch contour estimation takes into account the segmentation, and then tries to calculate the pitch only for the voiced segments. Once a voiced segment is found, the method first calculates some points in the centre of the segment and proceeds by tracking the pitch right and left. Pitch continuity is forced between segments by means of a maximum range that depends on the type of segment and the presence of pauses. Pitch value estimation is accomplished by an analy-

sis–synthesis method explained in Casajús-Quiros and Fernández-Cid (1994). Once we have both duration and pitch, we proceed to simplify the pitch contour. We keep the initial and final values for voiced consonants and three F0 values for vowels. All subsequent references to the original pitch contour are referred to this representation.

Phonological Level

The phonological level is responsible for assigning the pauses and determining the class of each breath group. In the original method, the breath groups could be any of ten related to actual grammatical constructions, like wh-question, enumeration, declarative final, etc. They were linguistically determined. Once we have the breath group and the position of the accents, we can classify each syllable in the prosodic structure. The full set of prosodic contours was obtained by averaging for each vowel the occurrences in a corpus, thus obtaining one intonation pattern for each vowel (syllable), either accented or not and in the final, penultimate or previous position in the accent group. The accent group belongs to a specific breath group and is located in its initial, intermediate or final position.

In the automatic approach, the classes of breath groups are obtained automatically by means of Vector Quantization (VQ). Thus the meaning of each class is lost, because each class is obtained by spontaneous clustering of the acoustical features of the syllables in the corpus. Two approaches have been devised for the breath group classification with similar results, one based on quantising the last syllable and another one by considering the last accented syllable.

Once the complete set of breath groups is obtained from the corpus, it must be synchronised with the linguistic features in order to proceed with an automatic rule generation mechanism. The mechanism works by linking part-of-speech (POS) tags and breath groups in a rule format. For each breath group, a basic rule is obtained taking the maximum context into account. Then all the sub-rules that can be obtained by means of reducing the context are also generated. For example, consider that according to the input sentence, the sequence of POS 30, 27, 42, 28, 32, 9 generate two breath groups (BG), BG 7 for the sequence of POS 30,27,42, and BG 15 for the sequence of POS 28, 32, 9. Then we will form the following set of rules: $\{30, 27, 42, 28, 32, 9\}$-> $\{7, 7, 7, 15, 15, 15\}$; $\{30, 27, 42\}$-> $\{7, 7, 7\}$; $\{27, 42\}$->$\{7, 7\}$; $\{27, 42, 28, 32, 9\}$-> $\{7, 7, 15, 15, 15\}$; $\{42, 28, 32, 9\}$-> $\{7, 15, 15, 15\}$; $\{28, 32, 9\}$-> $\{15, 15, 15\}$; $\{32, 9\}$-> $\{15,15\}$ and so on.

The resulting set of rules for the whole corpus will have some inconsistencies as well as repeated rules. A pruning algorithm eliminates both problems. At this point, two strategies are possible: either eliminate all contradictory rules or decide for the most frequent breath group (when there are several repetitions). In López-Gonzalo *et al.*, (1997), a more detailed description can be found with results on the different strategies.

Phonetic Level

The prosody is modelled with respect to segment and pause duration, as well as pitch contour. So far, intensity is not taken into account, but this is a probable next step. The prosodic unit in this level is the syllable. From the corpus we

proceed to quantise all durations of the pauses, rhyme and onset lengthening, as well as F0 contour and vowel duration. With this quantisation, we form a database of all the syllables in the corpus. For each syllable two types of features are kept, acoustic and linguistic. The stored linguistic features are: the name of nuclear vowel, the position of its accent group (initial, internal, final), the type of its breath group, the distance to the lexical accent and the place in the accent group. The current acoustic features are the duration of the pause (for pre-pausal syllables), the rhyme and onset lengthening, the prosodic pattern of the syllable and the prosodic pattern of the next syllable. It should be noted that the prosodic patterns carry information about both F0 and duration.

Prosody Assignment

As described above, in the original method, we have one prosodic pattern for each vowel with the same linguistic features. Thus obtaining the prosody of a syllable was a simple matter of looking up the right entry in a database.

In the automatic approach, the linguistic features are used to pre-select the candidate syllables and then the two last acoustic features are used to find an optimum alignment from the candidates. The optimum path is obtained by a shortest-path algorithm which combines a static error (which is a function of the linguistic adequacy) and a continuity error (obtained as the difference between the last acoustic feature and the actual pattern of each possible next syllable). This mechanism assures that a perfect assignment of prosody is possible if the sentence belongs to the prosodic corpus. Finally, the output is computed in the following steps: first, the duration of each consonant is obtained from its mean value and the rhyme/onset lengthening factor. Then, the pitch contour and duration of the vowel are copied from the centroid of its pattern. And finally, the pitch value of the voiced consonants is obtained by means of an interpolation between adjacent vowels, or by maintaining the level if they are adjacent to an unvoiced consonant.

Adaptation to the Galician Corpus

The corpus used for these experiments contains 80 sentences that cover a wide variety of intonation structures. There are 34 declarative (including 3 incomplete enumerations, 8 complete enumerations, and 10 parenthetical sentences), 21 exclamations (of which, 7 imperative) and 25 questions (10 or-questions, 6 yes–no questions, 2 negative and 7 wh-questions). For testing purposes there was a set of 10 sentences not used for training. There was at least one of each broad class.

The corpus was recorded at the University of Vigo by a professional radio speaker and hand-segmented and labelled by a trained linguist. Its mean F0 was 87 Hz, with a maximum of 190 Hz and a minimum of 51 Hz. Figure 21.2 shows the duration and mean F0 value of the vowels of the corpus as produced by the speaker. The F0 range for Galician is generally larger than for Spanish, in our previous corpus, the F0 range was about one octave, in this corpus there were

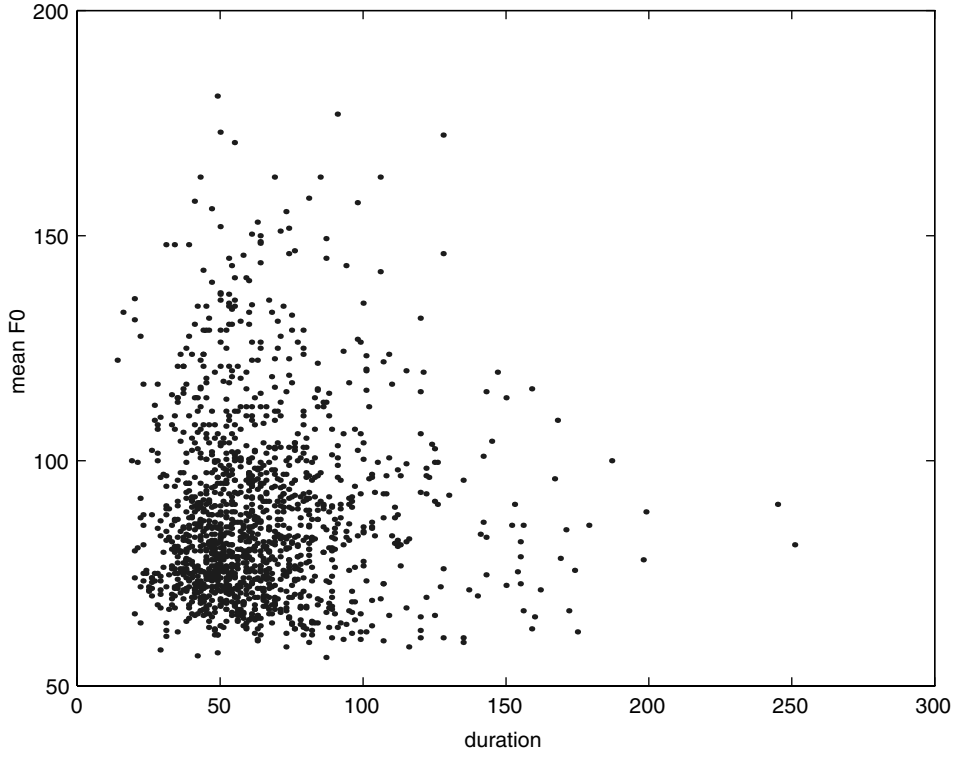


Figure 21.2 Scatter plot of the duration and mean F0 values of the vowels in the corpus

almost two octaves of range. In our previous recordings, speakers were instructed to produce speech without any emphasis. This led to a low F0 range in the previous recordings. Nevertheless, the 'musical accent' of the Galician language may result in a corpus with an increased F0 range.

The corpus contains many mispronunciations. It is interesting to note that some of them can be seen as 'contaminations' from Spanish (as 'prexudicial' in which the initial syllable is pronounced as in 'perjudicial', the Spanish word). Some others are typically Galician as the omission of plosives preceding fricatives ('ocional' instead of 'opcional'). The remaining ones are quite common in speech (joining of contiguous identical phonemes and even sequences of two phonemes as in 'visitabades espectaculos' which becomes 'visitabespectaculos').

The mismatch in pronunciation can be seen either as an error or a feature. Seen as an error, one could argue that in order to model prosody or anything else, special attention should be taken during the recordings to avoid erroneous pronunciations (as well as other accidents). On the other hand, mispronunciation is a very common effect, and can even be seen as a dialectal feature. As we intend to model a speaker automatically, we finally faced the problem of synchronising text and speech in the presence of mispronunciation (when it is not too severe, i.e. up to one deleted, inserted or swapped phoneme).

Experiments and Results

Experiments were conducted with different levels of quantisation. For syllables in the final position of breath groups, we tried 4 to 32 breath-group classes, although there was not enough data to cope with the 32 groups case. Pauses were classified into 16 groups, while it was possible to reach 32 rhyme and onset lengthening groups. For all syllables, 64 groups were obtained, with a good distribution coverage. For breath groups, we tried 4 to 32 breath-group classes, but there was not enough data to cope with the 32 groups case. Pauses were classified into 16 groups, while it was possible to reach 32 rhyme and onset lengthening groups. The number of classes was increased until the distortion ceased to decrease significantly.

Figure 21.3 shows the distribution of final vowels in a breath group with respect to duration and mean F0 value for the experiment with 16 breath groups. Each graph shows all the vowels pertaining to each group, as found in the corpus, together with the centroid of the group (plotted darker).

As can be seen, some of the groups (C1, C3) consist of a unique sample, thus only the centroid can be seen. Group C4 is formed by one erroneous sample (due to the pitch estimation method) and some correct ones. The resulting centroid averages them and this averaging is crucial when not dealing with a pitch stylisation mechanism and a supervised pitch estimation. The fact that some classes consist of only one sample makes further subdivision impossible. It should be noted that the centroid carries information about both F0 and duration, and that the pitch contour is obtained by interpolating F0 between the centroids. Therefore not only the shape of the F0 in each centroid, but also its absolute F0 level are relevant.

We performed a simple evaluation to compare the synthetic prosody produced with 16 breath groups with 'natural' Galician prosody. The ten sentences not used for training were analysed, modelled and quantised with the codebook used for prosodic synthesis. The synthetic prosody sentences were produced from the mapping rules obtained in the experiment. They were then synthesised with a PSOLA Spanish diphone synthesiser. The 'natural' prosody was obtained by quantising the original pitch contour by the 64 centroids obtained with all the syllables in the training corpus. The pairs of phrases were played in random order to five listeners. They were instructed to choose the preferred one of each pair.

The results show an almost random choice pattern, with a slight trend towards the original prosody ones. This was expected because the prosodic modelling method has already shown good performance with simple linguistic structures. Nevertheless, the 'Galician feeling' was lost even from the 'natural' prosody sentences. It seems that perception was dominated by the segmental properties contained in the Spanish diphones. A few sentences in Spanish with the Galician rules and prosodic database showed that fact without need for further evaluation.

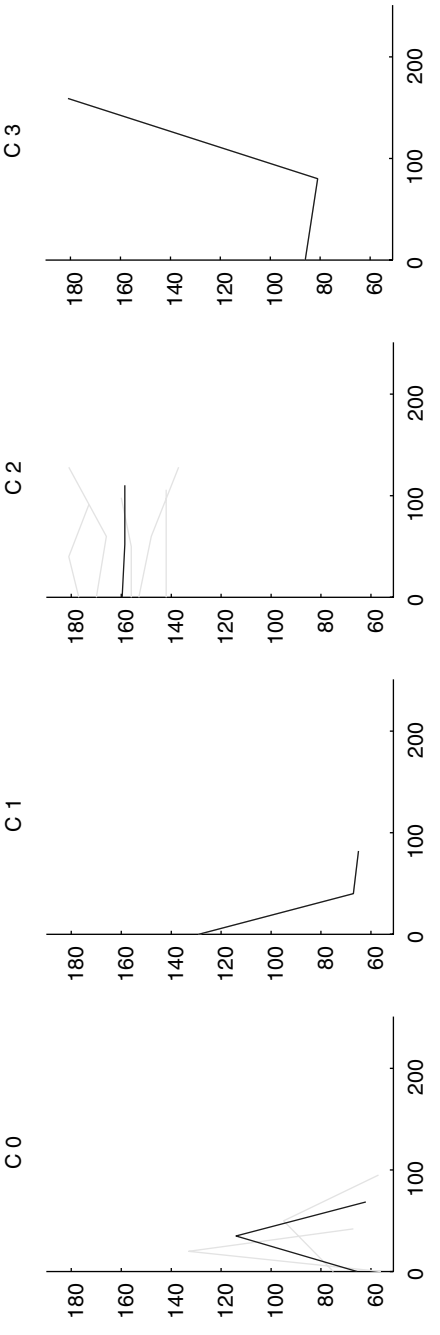
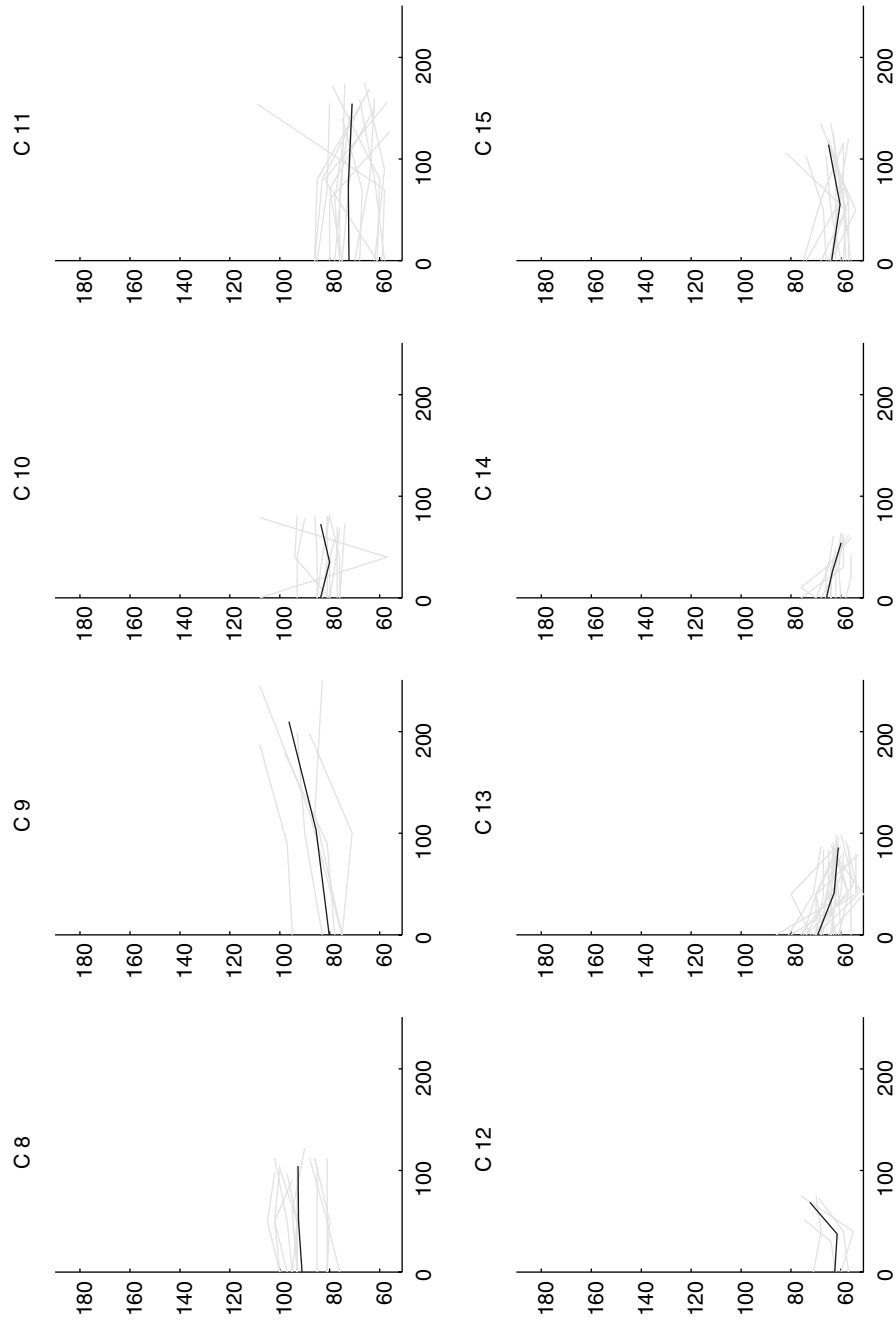


Figure 21.3 The 16 breath groups and the underlying vowels they quantise For each class (C0–C15) the x axis represents time in ms and y axis frequency in Hertz


C 11
C 10
C 9
C 8
C 15
C 14
C 13
C 12
180
160
140
120
100
80
60
0
100
200

Conclusion

First of all, we have found that our original aim was based on a wrong assumption, namely to produce a Galician accent by means of applying Galician prosody to Spanish. The real reason remains unanswered but several lines of action seem interesting: (a) use of the same voice for synthesis (to see if voice quality is of importance); (b) use of synthesiser with the complete inventory of Galician diphones (there are two open vowels and two consonants not present in Spanish). What is already known is that we can adapt the system to a prosodic corpus when the speaker has recorded both the diphone inventory and the prosodic database.

From the difficulties found we have refined our programs. Some of the problems are still only partially solved. It seems quite interesting to be able to learn the pronunciation pattern of the speaker (his particular phonetic transcription). Using the very same voice (in a unit selection concatenative approach) may achieve this result.

Regarding our internal data-structure, we have started to open it (see Villar-Navarro, *et al*., 1999). Even so, a unified prosodic-linguistic standard and a mark-up language would be desirable in order to keep all the information together and synchronised, and to be able to use a unified set of inspection tools, not to mention the possibility of sharing data, programs and results with other researchers.

References

Casajús-Quiros, F.J. and Fernández-Cid, P. (1994). Real-time, loose-harmonic matching fundamental frequency estimation for musical signals. *Proceedings of ICASSP '94*, (pp. II.221–224). Adelaide, Australia.

Emerard, F., Mortamet, L., and Cozannet A. (1992). Prosodic processing in a TTS synthesis system using a database and learning procedures. In G. Bailly and C. Benoit (eds), *Talking Machines: Theories, Models and Applications* (pp. 225–254). Elsevier.

López-Gonzalo, E. (1993). *Estudio de Técnicas de Procesado Lingüistico y Acústico para Sistemas de Conversión Texto – Voz en Español Basados en Concatenación de Unidades*. PhD thesis, E.T.S.I. Telecomunicación Universidad Politécnica de Madrid.

López-Gonzalo, E. and Hernández-Gómez, L.A. (1994). Data-driven joint f0 and duration modelling in text to speech conversion for Spanish. *Proceedings of ICASSP '94* (pp. I.589–592). Adelaide, Australia.

López-Gonzalo E. and Hernández-Gómez, L.A. (1995). Automatic data-driven prosodic modelling for text to speech. *Proceedings of EUROSPEECH '95* (pp. I.585–588). Madrid.

López-Gonzalo, E., Rodríguez-García, J.M., Hernández-Gómez, L.A., and Villar, J.M. (1997). Automatic corpus-based training of rules for prosodic generation in text-to-speech. *Proceedings of EUROSPEECH '97* (pp. 2515–2518). Rhodes, Greece.

Shimodaira, H. and Kimura, M. (1992). Accent phrase segmentation using pitch pattern clustering. *Proceedings of ICASSP '92* (pp. I–217–220). San Francisco.

Villar-Navarro, J.M., López-Gonzalo, E., and Relaño-Gil, J. (1999). A mixed approach to Spanish prosody. *Proceedings of EUROSPEECH '99* (pp. 1879–1882). Madrid.

Wightman, C.W. and Ostendorf, M. (1994). Automatic labeling of prosodic phrases. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, *4*, 469–481.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

22

Reduction and Assimilatory Processes in Conversational French Speech Implications for Speech Synthesis

Danielle Duez
Laboratoire Parole et Langage, CNRS ESA 6057, Aix en Provence, France
duez@lpl.univ-aix.fr

Introduction

Speakers adaptively tune phonetic gestures to the various needs of speaking situations (Lindblom, 1990). For example, in informal speech styles such as conversations, speakers speak fast and hypoarticulate, decreasing the duration and amplitude of phonetic gestures and increasing their temporal overlap. At the acoustic level, hypoarticulation is reflected by a higher reduction and context-dependence of speech segments: Segments are often reduced, altered, omitted, or combined with other segments compared to the same read words.

Hypoarticulation does not affect speech segments in a uniform way: It is ruled by a certain number of linguistic factors such as the phonetic properties of speech segments, their immediate context, their position within syllables and words, and by lexical properties such as word stress or word novelty. Fundamentally, it is governed by the necessity for the speaker to produce an auditory signal which possesses sufficient discriminatory power for successful word recognition and communication (Lindblom, 1990).

Therefore the investigation of reduction and contextual assimilation processes in conversational speech should allow us to gain a better understanding of the basic principles that govern them. In particular, it should allow us to find answers to the questions as why certain modifications occur and others do not, and why they take particular directions. The implications would be of great interest for the improvement of speech synthesis. It is admitted that current speech-synthesis systems are principally able to generate highly intelligible output. However, there are still difficulties with naturalness of synthetic speech, which is strongly dependent on con-

textual assimilation and reduction modelling (Hess, 1995). In particular, it is crucial for synthesis quality and naturalness to manipulate speech segments in the right manner and at the right place.

This chapter is organised as follows. First, perceptual and spectrographic data obtained for aspects of assimilation and reduction in oral vowels (Duez, 1992), voiced stops (Duez, 1995) and consonant sequences (Duez, 1998) in conversational speech are summarised. Reduction means here a process in which a consonant or a vowel is modified in the direction of lesser constriction or weaker articulation, such as a stop becoming an affricate or fricative, or a fricative becoming a sonorant, or a closed vowel becoming more open. Assimilation refers to a process that increases the similarity between two adjacent (or next-to-adjacent) segments. Then, we deal with the interaction of reduction and assimilatory processes with factors such as the phonetic properties of speech sounds, immediate adjacent context (vocalic and consonantal), word class (grammatical or lexical), position in syllables and words (initial, medial or final), position in phrases (final or non-final). The next section summarises some reduction-and-assimilation tendencies. The final section raises some problems of how to integrate reduction and contextual assimilation in order to improve naturalness of speech-synthesis and proposes a certain number of rules derived from results on reduction and assimilation are discussed.

Reduction and contextual assimilation

Vowels

Measurements of the second formant measured in CV syllables occurring in conversational speech and read speech showed that the difference in formant frequency between the CV boundary (locus) and the vowel nucleus (measured in the middle of the vowel) was smaller in conversational speech. The frequency change was also found to be greater for the nucleus than for the locus. Moreover, loci and nuclei did not change in the same direction. The results were interpreted as reflecting differences in coarticulation, both an anticipatory effect of the subsequent vowel on the preceding consonant, and/or formant undershoot (as defined by Lindblom, 1963).

Voiced stops

Perceptual and acoustic data on voiced stops extracted from the conversational speech produced by two speakers revealed two consistent tendencies: (1) There was a partial or complete nasalisation of /b/'s and /d/'s in a nasal vowel context, that is, a preceding and/or a succeeding nasal vowel: at the articulatory level, there was the velum-lowering gesture partially or totally overlapped with the closing gesture (for an illustration of complete nasalisation, see the following example):

pendant ('during')	
Phonological	Identified
/pãdã/	/pãnã/

(2) There was a weakening of /b/ into the corresponding approximant fricative /β/, semivowel /w/ and approximant (labial) and the weakening of /d/ into the corresponding fricative /z/, sonorant /l/, approximant /dental/, or its complete deletion. These changes were assumed to be the result of a reduction in the magnitude of the closure gesture. The deletion of the consonant was viewed as reflecting the complete deletion of the closure gesture. Interestingly, assimilated or reduced consonants tended to keep their place of articulation, suggesting that place of articulation is one of the consonantal invariants.

Consonant Sequences

A high number of heterosyllabic [C_1#C_2] and homosyllabic [C_1C_2] consonant sequences were different from their phonological counterparts. In most cases, C_1's were changed into another consonant or omitted. Voiced or unvoiced fricatives and occlusives were devoiced or voiced, reflecting the anticipatory effect of an unvoiced or voiced C_2. Voiced or unvoiced occlusives were nasalized when preceded by a nasal vowel, suggesting a total overlapping of the velum-lowering gesture of the nasal vowel with the closure gesture. Similar patterns were observed for a few C_2's. There were also some C_1's and C_2's with only one or two features identified: Voicing, devoicing and nasalisation were incomplete, reflecting partial contextual assimilation. Other consonants, especially sonorants, were omitted, which may be the result an extreme reduction process. An illustration of C_1-omission can be seen in the following example:

Il m'est arrivé ('it happened to me')

Phonological	Identified
/ilmɛtaʁive/	/imɛtaʁive/

In some cases, there was a reciprocal assimilation of C_1 to C_2. It was particularly obvious in C_1C_2's, where the manner and place features of C_1 coalesced the voicing feature of C_2 to give a single consonant (/sd/ ⇒ /z/, /js/= /ʃ/, /sv/ ⇒ /z/, /fz/ ⇒ /z/, /tv/ ⇒ /d/). An illustration can be found in the following example:

Une espèce de ('a kind of')

Phonological	Identified
/ynɛspɛsdə/	/ynɛspɛzə/

Thus, two main trends in assimilation characterised consonant sequences: (1) Assimilation of C_1 and C_2 to nasal vowel context; and (2) voicing assimilation of C_1 to C_2, and/or C_2 to C_1. In all cases, C_1 and C_2 tended each to keep their place of articulation.

Factors Limiting Reduction-and-Assimilation Effects

Segment Properties and Adjacent Segments

Vowels as well as consonants underwent different degrees of reduction and assimilation. The loci and the nuclei of the front vowels were lowered, while those of the back vowels were raised, and there was little change for vowels with

mid-frequency F_2 Nucleus-frequency differences exhibited greater changes for back vowels than for front vowels, for labials as well as for dentals. Data obtained for voiced stops revealed a higher identification rate for dentals than for labials, suggesting that the former resist reduction and assimilatory effects more than the latter. This finding may be due to the fact that the degree of freedom is greater for the lips than the tongue which is submitted to a wide range of constraints. Consonant sequences also revealed a different behaviour for the different consonant types. Omitted consonants were mostly sonorants Moreover, differences were observed within a same category. The omitted sonorants were /l, or m/, those reported as different were /n/ changed into /m/ before / p/.

The above findings suggest a lesser degree of resistance to reduction and assimilatory effects for sonorants than for occlusives and fricatives. Sonorants are consonants with a formantic structure: They are easily changed into vowels or completely deleted. Similarly, voiced occlusives are less resistant than unvoiced occlusives which have more articulatory force (Delattre, 1966). The resistance of speech-segments to the influence of reduction and contextual assimilation should be investigated in various languages: The segments which resist more are probably those which in turn exert a stronger influence on their neighbours.

Syllable Structure and Position in a Syllable.

Mean identification scores were higher for homosyllabic C_1C_2's than for heterosyllabic ones. The highest identification scores were for sequences consisting of a fricative plus a sonorant, the lowest scores for sequences composed of two occlusives. In heterosyllabic sequences, the C_1's not equal to their phonological counterparts were mostly in coda position. Moreover, in C_1C_2-onset sequences there was a slight tendency for C_2's to be identified as a different consonant. The data suggest a stronger resistance of onset-speech segments, which is in total conformity with the results found for articulatory strength (Straka, 1964). Moreover, onset segments have a higher signalling value for a listener in word recognition.

Word Class

Word class had no significant effect on the identification of voiced plosives, but a significant effect on the identification of C_1's in consonant sequences. Grammatical words did not react in the same way to the influence of reduction and assimilatory processes. For example, the elided article or preposition (de ⇒/dø/) was often omitted in $C_1\#C_2$'s as C_1 as well as C_2. It was also often changed into an /n/ when it was an intervocalic voiced stop preceded by a nasal vowel. On the opposite, in phrases consisting of je ⇒ /ʒə/ (personal pronoun) + verb (lexical word), the /ʒ/ was maintained while the first consonant of the verb was mostly reported as omitted, or at least changed into another consonant.

et je vais te dire (and I am going to tell you)

Phonological	Identified
/ɛʒvɛtədi/	/ɛʒɛtədi/

Final Prominence

In French, the rhythmic pattern of utterances mainly relies on the *prominence* given to final syllables at the edge of a breath group (Vaissière, 1991). As final prominence is largely signalled by lengthening, final-phrase syllables tend to be long, compared to non-final phrase syllables. Phrase-final segments resist the influence of reduction and assimilatory processes which are partly dependent on duration (Lindblom, 1963). Prominent syllables showed a larger formant excursion from the locus to the nucleus than non-prominent ones. Voiced plosives and consonant sequences perceived as phonological were located within prominent syllables.

Tendencies in Reduction and Assimilation

Natural speech production is a balance between an articulatory-effort economy on the part of the speaker and the ability to perceive and understand on the part of the listener. These two principles operate, at different degrees, in all languages, in any discourse and everywhere in the discourse, within syllables, words, phrases and utterances. Thus, the acoustic structure of the speech signal is characterised by a continuous succession of (more or less) overlapping and reduced segments, the degree and the extent of overlapping and reduction being dependent on speech style and information. Reduction and assimilatory processes are universal since they reflect basic articulatory mechanisms, but they are also language-dependent to the extent that they are ruled by phonological and prosodic structures of languages. Interestingly, the regularities observed here suggest some tendencies in reduction and contextual assimilation specific to French.

Nasalisation

There is a universal tendency for *nasality* to spread from one segment to another, although the details vary greatly from one language to another and nasalisation is a complex process that operates in different stages. For example, the normal path of emergence of distinctive nasal vowels begins with the non-distinctive nasalisation of vowels next to consonants. This stage is followed by the loss of the nasal consonants and the persistence of vowel nasalisation, which therefore becomes distinctive (Ferguson, 1963; Greenberg, 1966). Interestingly, investigations of patterns of nasalisation in modern French revealed different nasalisation-timing patterns (Duez, 1995; Ohala and Ohala, 1991) and nasalisation degrees depending on consonant permeability (Ohala and Ohala, 1991). The fact that nasal vowels may partially or completely nasalise adjacent occlusives has implications for speech synthesis since sequences containing voiced or unvoiced occlusives preceded by a nasal vowel are frequent in common adverbs and numbers.

C_2 Dominance

In languages such as French, the peak of intensity coincides with the vowel while in some other languages, it occurs earlier in the syllable and tends to remain constant. In the first case, the following consonant tends to be weak and may drop while in the other case, it tends to be reinforced. This characteristic partly explains the evolution of French (for example, the loss of the nasal consonant in the process of nasalisation) and the predominance of CV syllables (Delattre, 1969). It also gives an explanation to the strong tendency for occlusive or fricative C_1's to be voiced or devoiced under the anticipatory effect of a subsequent unvoiced or voiced occlusive or fricative, and for sonorants to be vocalised or omitted.

Resistance of Prominent Final-Phrase Syllables

In French, prominent syllables are components of a hierarchical prosodic structure, and boundary markers. They are information points which predominantly attract the listener's attention (Hutzen, 1959), important landmarks which impose a cadence on the listener for integrating information (Vaissière, 1991). They are crucial for word recognition (Grosjean and Gee, 1987) and the segmentation of the speech stream into hierarchical syntactic and discourse units. Thus, the crucial role of the prominence pattern in speech perception and production may account for its effect on the reduction and contextual assimilation of speech segments.

Implications for Speech Synthesis

The fact that speech production is at a same time governed by an effort-economy principle and perceptual needs has crucial implications for speech-synthesis. Perceived naturalness has proven to strongly depend on the fit to natural speech, listeners being responsive to an incredible number of acoustic details and performing best when the synthesis contains all known regularities (Klatt, 1987). As a consequence, the improvement of synthetic naturalness at the segmental level requires detailed acoustic information, which implies in turn a fine-grained knowledge of linguistic processes operating at different levels in the speech hierarchy, and in particular a good understanding of reduction and assimilation processes in languages.

Concatenation-Based Synthesisers

There are two types of synthesisers: formant and spectral-domain synthesisers, and concatenation-based synthesisers. Concatenation-based synthesisers are based on the concatenation of natural speech units of various sizes (diphones, demi-syllables, syllables and words) recorded from a human speaker. They present a certain number of advantages and disadvantages mainly related to the size of units. For example, small units such as diphones and demisyllables do not need much memory but do not contain all the necessary information on assimilation and reduction phenomena. Diphones which are units extending from the central point of the steady part of a phone to the central point of the following phone contain

information on consonant/vowel and vowel-consonant transitions but do not cover coarticulation effects in consonant sequences. In contrast, demisyllables which result from the division of a syllable into an initial and a final demisyllable (Fujimura, 1976) cover most coarticulation effects in onset and coda consonant sequences actually present in words but not in sequences resulting from the elision of an optional /ə/. Systems based on the concatenation of larger units such as syllables and words (Lewis, and Tatham, 1999; Stöber *et al*., 1999) solve some of the above problems since they contain many coarticulatory and reduction effects. However, they also need to be context-knowledge based. For example, Lewis and Tatham (1999) described how syllables have to be modified for concatenation in contexts other than those from which they were excised. Stöber *et al*. (1999) proposed a system using words possessing the inherent prosodic features and the right pronunciation. In concatenation-based systems, the quality and naturalness of synthesis require the selection of appropriate concatenation units or word instances in the right contexts, which implies the knowledge of regularities in reduction and assimilatory processes. In French consonant sequences, the assimilation of an occlusive to a preceding nasal vowel was shown to depend on its location within syllables (final or initial) and its membership in either homosyllabic or heterosyllabic sequences. Coarticulatory effects were also found to be limited by final prominence. Thus, different timing patterns of nasalisation can be obtained for occlusives by integrating in the corpus different instances of the same syllables or words produced in both phrase-final and phrase-internal positions. Similarly, the problem of grammatical words which tend to sound 'too loud and too long' (Hess, 1995) can be solved by recording different instances of these words in different contexts. This procedure should be particularly useful for grammatical words whose prominence depends on their location within phrases. The personal pronoun (il $\Rightarrow$ / il/) may be clearly articulated in phrase-final position, on the opposite, the /l/ is deleted when /i l/ is followed by a verb, that is, in phrase-internal position. In the latter case, it constitutes with the verb a single prosodic word. Some verbal phrases consisting of the personal pronoun (je /ʒə/ + verb) were also shown to present considerable and complex reduction. In some cases there was elision of /ə/ and assimilation of voicing of /ʒ/ to the following consonant. In other cases, there was deletion of /ə/ and partial or complete reduction of the verb-initial consonant. As verbal phrases are frequently used, different instances as a function of context and styles might be added in the corpus.

Rule-Based Synthesis: Rules of Reduction and Assimilation

In formant and spectral-domain synthesisers where the generation of the acoustic signal is derived from a set of segmental rules which model the steady state properties of phoneme realisation and control the fusion of strings of phonemes into connected speech (Klatt, 1987), output can be improved (at least partly) by the incorporation of reduction and contextual-assimilation rules in the text-to-speech system. For example, the present results suggest that we should include the following rules for consonants located in non-prominent syllables: (1) rules of nasalisation for voiced intervocalic occlusives followed and/or preceded by a nasal vowel, and for unvoiced and voiced syllable-final plosives preceded

by a nasal vowel and followed by another consonant; (2) rules of devoicing or voicing for voiced or unvoiced syllable-final obstruents before an unvoiced or voiced syllable-initial obstruent; (3) rules of vocalisation for syllable-final sonorants in heterosyllabic sequences and (4) rules of deletion of / l/ into personal pronouns.

An illustration can be found in the following tentative rules. The formalism of these rules follows that of Kohler (1990) and has the following characteristics. Rules are of the form $X \Rightarrow Y / W__Z$ where X is rewritten Y after left-hand context W and before right-hand context Z, respectively. In the absence of Y, there is a deletion rule. Each symbol is composed of a phonetic segment, V and C for vowels and consonants, respectively, and # for syllable boundary. Vowels and consonants are defined as a function of binary features. +/− FUNC means function/non-function word marker. As assimilated, reduced and omitted consonants were mostly located into non-prominent syllables, the feature /-PROM/ is not represented.

Change of intervocalic voiced plosives into their nasal counterparts after a nasal vowel

$$\begin{bmatrix} \text{C} \\ -\text{nas} \\ +\text{voice} \\ +\text{occl} \end{bmatrix} \Rightarrow [+\text{nas}] \; / \; \begin{bmatrix} \text{V} \\ +\text{nas} \end{bmatrix} \; \text{———} \; \text{V}$$

Nasalisation of voiced-or-unvoiced-stops before any syllable-initial consonant

$$\begin{bmatrix} \text{C} \\ -\text{nas} \\ +\text{occl} \end{bmatrix} \Rightarrow [+\text{nas}] \; / \; \begin{bmatrix} \text{V} \\ +\text{nas} \end{bmatrix} \; \text{———} \; \#\,\text{C}$$

Voicing of unvoiced obstruents before syllable-initial voiced obstruents

$$\begin{bmatrix} \text{C} \\ -\text{voice} \\ +\text{obst} \end{bmatrix} \Rightarrow [+\text{voice}] \; / \; \text{———} \; \# \; \begin{bmatrix} \text{C} \\ +\text{voice} \\ +\text{obst} \end{bmatrix}$$

Devoicing of voiced obstruents before unvoiced syllable-initial obstruents

$$\begin{bmatrix} \text{C} \\ +\text{voice} \\ +\text{obst} \end{bmatrix} \Rightarrow [-\text{voice}] \; / \; \text{———} \; \# \; \begin{bmatrix} \text{C} \\ -\text{voice} \\ +\text{obst} \end{bmatrix}$$

Vocalisation of sonorants before any syllable-initial consonant

$$\begin{bmatrix} \text{C} \\ +\text{son} \end{bmatrix} \Rightarrow \text{V} \; / \; \text{———} \; \#\;\text{C}$$

Deletion of /l/ in the function word /il/ before any syllable-initial consonant

$$\begin{bmatrix} \text{L} \\ +\text{func} \end{bmatrix} \Rightarrow \emptyset \; / \; \text{———} \; \#\;\text{C}$$

References

Delattre, P. (1966). La force d'articulation consonantique en français. *Studies in French and Comparative Phonetics* (pp. 111–119). Mouton.

Delattre, P. (1969). Syllabic features and phonic impression in English, German, French and Spanish, *Lingua, 22*, 160–175.

Duez, D. (1992). Second formant locus-nucleus patterns: An investigation of spontaneous French speech. *Speech Communication, 11*, 417–427.

Duez, D. (1995). On spontaneous French speech: Aspects of the reduction and contextual assimilation of voiced plosives. *Journal of Phonetics, 23*, 407–427.

Duez, D. (1998). Consonant sequences in spontaneous French speech. *Sound Patterns of Spontaneous Speech, ESCA Workshop* (pp. 63–68). La Baume-les-Aix, France.

Ferguson, F.C. (1963). Assumptions about nasals: A sample study in phonological universals. J.H. Greenberg (ed.), *Universals of Language* (pp. 53–60). MIT Press.

Fujimura, O. (1976). Syllable as the Unit of Speech Synthesis. Internal memo. Bell Laboratories.

Greenberg, J.H. (1966). Synchronic and diachronic universals in phonology. *Language, 42*, 508–517.

Grosjean, F. and Gee, P.J. (1987). Prosodic structure and word recognition. *Cognition, 25*, 135–155.

Hess, W. (1995). Improving the quality of speech synthesis systems at segmental level. In C. Sorin, J. Mariani, H. Meloni and J. Schoentgen (eds), *Levels in Speech Communication: Relations and Interactions* (pp. 239–248). Elsevier.

Hutzen, L.S. (1959). Information points in intonation. *Phonetica, 4*, 107–120.

Klatt, D.H. (1987). Review of text-to-text conversion for English. *Journal of the Acoustical Society of America, 82–3*, 737–797.

Kohler, K. (1990). Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W.J. Hardcastle and A. Marchal (eds), *Speech Production and Speech Modelling. NATO ASI Series*, Vol. 55 (pp. 69–92). Kluwer.

Lewis, E. and Tatham, M. (1999). Word and syllable concatenation in text-to-speech synthesis. *Eurospeech*, Vol. 2 (pp. 615–618). Budapest.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America, 35*, 1773–1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. Hardcastle and A. Marchal (eds), *Speech Production and Speech Modelling. NATO ASI Series*, Vol. 55 (pp. 403–439). Kluwer.

Ohala, M. and Ohala, J.J. (1991). Nasal epenthesis in Hindi. *Phonetica, 48*, 207–220.

Stöber, K., Portele, T., Wagner, P., and Hess, W. (1999). Synthesis by word concatenation. *Eurospeech*, Vol. 2 (pp. 619–622). Budapest.

Straka, G. (1964). L'évolution phonétique du latin au français sous l'effet de l'énergie et de la faiblesse articulatoire. *T.L.L., Centre de Philologie Romane, Strasbourg II*, 17–28.

Vaissière, J. (1991). Rhythm, accentuation and final lengthening in French. In J. Sundberg, L. Nord and R. Carlson, *Music, Language and Brain* (pp. 108–120). Macmillan.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

23

Acoustic Patterns of Emotions

Branka Zei Pollermann and Marc Archinard
Liaison Psychiatry, Geneva University Hospitals
51 Boulevard de la Cluse, CH-1205 Geneva, Switzerland
branka.zei@hcuge.ch

Introduction

Naturalness of synthesised speech is often judged by how well it reflects the speaker's emotions and/or how well it features the culturally shared vocal prototypes of emotions (Scherer, 1992). Emotionally coloured vocal output is thus characterised by a blend of features constituting patterns of a number of acoustic parameters related to F0, energy, rate of delivery and the long-term average spectrum.

Using the covariance model of acoustic patterning of emotional expression, the chapter presents the authors' data on: (1) the inter-relationships between acoustic parameters in male and female subjects; and (2) the acoustic differentiation of emotions. The data also indicate that variations in F0, energy, and timing parameters mainly reflect different degrees of emotionally induced physiological arousal, while the configurations of long term average spectra (more related to voice quality) reflect both arousal and the hedonic valence of emotional states.

Psychophysiological Determinants of Emotional Speech Patterns

Emotions have been described as psycho-physiological processes that include cognitions, visceral and immunological reactions, verbal and nonverbal expressive displays as well as activation of behavioural reactions (such as approach, avoidance, repulsion). The latter reactions can vary from covert dispositions to overt behaviour. Both expressive displays and behavioural dispositions/reactions are supported by the autonomic nervous system which influences the vocalisation process on three levels: respiration, phonation and articulation. According to the *covariance* model (Scherer *et al*., 1984; Scherer and Zei, 1988; Scherer, 1989), speech patterns covary with emotionally induced physiological changes in respiration, phonation and articulation. The latter variations affect vocalisation on three levels:

1. suprasegmental (overall pitch and energy levels and their variations as well as timing);
2. segmental (tense/lax articulation and articulation rate);
3. intrasegmental (voice quality).

Emotions are usually characterised along two basic dimensions:

1. activation level (aroused vs. calm), which mainly refers to the physiological arousal involved in the preparation of the organism for an appropriate reaction;
2. hedonic valence (pleasant/positive vs. unpleasant/negative) which mainly refers to the overall subjective hedonic feeling.

The precise relationship between the physiological activation and vocal expression was first modelled by Williams and Stevens (1972) and has received considerable empirical support (Banse and Scherer, 1996; Scherer, 1981; Simonov *et al.*, 1980; Williams and Stevens, 1981). The activation aspect of emotions is thus known to be mainly reflected in the pitch and energy parameters such as mean F0, F0 range, general F0 variability (usually expressed either as SD or the coefficient of variation), mean acoustic energy level, its range and its variability as well as the rate of delivery. Compared with an emotionally unmarked (neutral) speaking style, an angry voice would be typically characterised by increased values of many or all of the above parameters, while sadness would be marked by a decrease in the same parameters. By contrast, the hedonic valence dimension, appears to be mainly reflected in intonation patterns, and in voice quality.

While voice patterns related to emotions have a status of symptoms (i.e. signals emitted involuntarily), those influenced by socio-cultural and linguistic conventions have a status of a consciously controlled speaking style. Vocal output is therefore seen as a result of two forces: the speaker's physiological state and socio-cultural linguistic constraints (Scherer and Kappas, 1988).

As the physiological state exerts a direct causal influence on vocal behaviour, the model based on scalar *covariance* of continuous acoustic variables appears to have high cross-language validity. By contrast the *configuration* model remains restricted to specific socio-linguistic contexts, as it is based on configurations of category variables (like pitch 'fall' or pitch 'rise') combined with linguistic choices. From the listener's point of view, naturalness of speech will thus depend upon a blend of acoustic indicators related, on the one hand, to emotional arousal, and on the other hand, to culturally shared vocal stereotypes and/or prototypes characteristic of a social group and its status.

Intra and Inter-Emotion Patterning of Acoustic Parameters

Subjects and Procedure

Seventy-two French speaking subjects' voices were used. Emotional states were induced through verbal recall of the subjects' own emotional experiences of joy,

sadness and anger (Mendolia and Kleck, 1993). At the end of each recall, the subjects said a standard sentence on the emotion congruent tone of voice. The sentence was: 'Alors, tu acceptes cette affaire' ('So you accept the deal.'). Voices were digitally recorded, with mouth-to-microphone distance being kept constant.

The success of emotion induction and the degree of emotional arousal experienced during the recall and the saying of the sentence were assessed through self-report. The voices of 66 subjects who reported having felt emotional arousal while saying the sentence were taken into account (30 male and 36 female). Computerised analyses of the subjects' voices were performed by means of Signalyze, a Macintosh platform software (Keller, 1994). The latter provided measurements of a number of vocal parameters related to emotional arousal (Banse and Scherer, 1996; Scherer, 1989). The following vocal parameters were used for statistical analyses: mean F0, F0sd, F0 max/min ratio, voiced energy range. The latter was measured between two mid-point vowel nuclei corresponding to the lowest and the highest peak in the energy envelopes and expressed in pseudo dB units (Zei and Archinard, 1998). The rate of delivery was expressed as the number of syllables uttered per second. Long-term average spectra were also computed.

Results for Intra-Emotion Patterning

Significant differences between male and female subjects were revealed by the ANOVA test. The differences concerned only pitch-related parameters. There was no significant gender-dependent difference either for voiced energy range or for the rate of delivery: both male and female subjects had similar distributions of values regarding the rate of delivery and voiced energy range. Table 23.1 presents the F0 parameters affected by speakers' gender and ANOVA results.

Table 23.1 F0 parameters affected by speakers' gender

Emotions	F0 mean in Hz	ANOVA	F0 max/min ratio	ANOVA	F0 SD	ANOVA
anger	M 128; F 228	$F(1, 64) = 84.6^{***}$	M 2.0; F 1.8	$F(1, 64) = 5.6^{*}$	M 21.2; F 33.8	$F(1, 64) = 11.0^{**}$
joy	M 126; F 236	$F(1, 64) = 116.8^{***}$	M 1.9; F 1.9	$F(1, 64) = .13$	M 22.6; F 36.9	$F(1, 64) = 14.5^{***}$
sadness	M 104; F 201	$F(1, 64) = 267.4^{***}$	M 1.6; F 1.5	$F(1, 64) = .96$	M 10.2; F 19.0	$F(1, 64) = 39.6^{***}$

Note: N = 66. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; M = male; F = female.

As gender is both a sociological variable (related to social category and cultural status) and a physiological variable (related to the anatomy of the vocal tract), we assessed the relation between mean F0 and other vocal parameters. This was done by computing partial correlations between mean F0 and other vocal parameters, with sex of speaker being partialed out. The results show that the subjects with higher F0 also have higher F0 range (expressed as max/min ratio) across all emotions. In anger, the subjects with higher F0 also exhibit higher pitch variability (expressed as F0sd) and faster delivery rate. In sadness the F0 level is negatively correlated with voiced energy range. Table 23.2 presents the results.

Results for Inter-Emotion Patterning

The inter-emotion comparison of vocal data was performed separately for male and female subjects. A paired-samples t-test was applied. The pairs consisted of the same acoustic parameter measured for two emotions. The results presented in Tables 23.2 and 23.4 show significant differences mainly for emotions that differ on the level of physiological activation: anger vs. sadness, and joy vs. sadness. We thus concluded that F0–related parameters, voiced energy range, and the rate of delivery mainly contribute to the differentiation of emotions at the level of physiological arousal.

In order to find vocal indicators of emotional valence, we compared voice quality parameters for anger (a negative emotion with high level of physiological arousal) with those for joy (a positive emotion with high level of physiological arousal). This was inspired by the studies on the measurement of vocal differentiation of hedonic valence in spectral analyses of the voices of astronauts (Popov *et al.*, 1971; Simonov *et al.*, 1980). We thus hypothesised that spectral parameters could significantly differentiate between positive and negative valence of the emotions which have similar levels of physiological activation. To this purpose, long-term average spectra (LTAS) were computed for each voice sample, yielding 128 data points for a range of 40–5 500 Hz.

We used a Bark-based strategy of spectral data analyses, where perceptually equal intervals of pitch are represented as equal distances on the scale. The frequencies covered by 1.5 Bark intervals were the following: 40–161 Hz; 161–297 Hz;

Table 23.2 Partial correlation coefficients between mean F0 and other vocal parameters with speaker's gender partialed out

Mean F0 and emotions	F0 max/min ratio	F0 sd	voiced energy range in pseud dB	Delivery rate
mean F0 in Anger	.43**	.77**	−.03	.39**
mean F0 in Joy	.36**	.66**	−.08	.16
mean F0 in Sadness	.32**	.56**	−.43**	−.13

Note: N = 66. *$p < .05$, **$p < .01$, ***$p < .001$; all significance levels are 2-tailed.

Table 23.3 Acoustic differentiation of emotions in male speakers

Emotions compared	F0 mean in Hz	T-test and P	F0 max/min ratio	T-test and P	F0 SD	T-test and P	Voiced energy range in pseudo d	T-test and P	Delivery rate	T-test and P
sadness	104		1.6		10.2		9.6		3.9	
anger	128	−4.3***	2.0	−6.0***	21.2	−5.7***	14.2	−5.0***	4.6	−2.2*
sadness	104		1.6		10.2		9.6		3.9	
joy	126	−4.6***	1.9	−6.0***	22.7	−7.5***	12.1	−2.5*	4.5	−2.9**
joy	126		1.9		22.7		12.0		4.5	
anger	128	−.4	2.0	−.9	21.2	.8	14.2	−2.8**	4.6	−.2

Note: N = 30. *$p < .05$, **$p < .01$, ***$p < .001$; all significance levels are 2-tailed.

Table 23.4 Acoustic differentiation of emotions in female speakers

Emotions compared	F0 mean in Hz	T-test and P	F0 max/min ratio	T-test and P	F0 SD	T-test and P	voiced energy range in pseudo dB	T-test and P	Delivery rate	T-test and P
Sadness	201		1.5		19.0		10.9		4.2	
Anger	228	−2.7**	1.8	−3.4**	33.8	−4.8***	14.2	−2.9**	5.0	−3.7**
Sadness	201		1.5		19.0		10.9		4.2	
Joy	236	−3.7**	1.9	−5.7***	37.0	−6.1***	12.8	−2.2*	5.0	−3.3**
Joy	236		1.9		37.0		12.8		5.0	
Anger	228	.8	1.8	1.6	33.8	1.0	14.2	−1.0	5.0	−.1

Note: N = 36. *$p < .05$, **$p < .01$, ***$p < .001$; all significance levels are 2-tailed.

297–453 Hz; 453–631 Hz; 631–838 Hz; 838–1 081 Hz; 1 081–1 370 Hz; 1 370–1 720 Hz; 1 720–2 152 Hz; 2 152–2 700 Hz; 2 700–3 400 Hz; 3 400–4 370 Hz; 4 370–5 500 Hz (Hassal and Zaveri, 1979; Pittam and Gallois, 1986; Pittam, 1987). Subsequently mean energy value for each band was computed. We thus obtained 13 spectral energy values per emotion and per subject.

Paired t-tests were applied. The pairs consisted of the same acoustic parameter (the value regarding the same frequency interval) compared across two emotions. The results showed that several frequency bands contributed significantly to the differentiation between anger and joy, thus confirming the hypothesis that the valence dimension of emotions can be reflected in the long term average spectrum.

The results show that in a large portion of the spectrum, energy is higher in anger than in joy. In male subjects it is significantly higher as of 300 Hz up to 3 400 Hz, while in female subjects the spectral energy is higher in anger than in joy in the frequency range from 800–3 400 Hz. Thus our analysis of LTAS curves, based on 1.5 Bark intervals, shows that an overall difference in energy is not the consequence of major differences in the distribution of energy across the spectrum for Anger and Joy. This fact may lend itself to two interpretations: (1) those aspects of voice quality which are measured by spectral distribution are not relevant for the distinction between positive and negative valence of high-arousal emotions or (2) anger and joy also differ on the level of arousal which is reflected in spectral energy (both voiced and voiceless). Table 23.5 presents the details of the results for the Bark-based strategy of the LTAS analysis.

Although we assumed that vocal signalling of emotion can function independently of the semantic and affective information inherent to the text (Banse and Scherer, 1996; Scherer, Ladd, and Silverman, 1984), the generally positive connotations of

Table 23.5 Spectral differentiation between anger and joy utterances in 1.5 Bark frequency intervals.

Frequency bands in Hz	spectral energy in pseudo dB Male subjects	T-test and P	spectral energy in pseudo dB Female subjects	T-test and P
40–161	A 18.6; J 17.6	.69	A 12.2; J 13.8	−1.2
161–297	A 23.5; J 20.8	2.0	A 19.1; J 18.9	.12
297–453	A 26.7; J 22	3.1*	A 21.9; J 20.8	.62
453–631	A 30.9; J 24.3	3.4**	A 24.2; J 21.3	1.5
631–838	A 28.5; J 21.0	4.4**	A 23.6; J 19.3	2.2
838–1 081	A 21.1; J 15.8	3.8**	A 19.4; J 14.7	2.6*
1 081–1 370	A 19.6; J 14.8	3.6**	A 16.9; J 12.6	2.9*
1 370–1 720	A 22.5; J 17.0	3.7**	A 17.5; J 12.9	3.3**
1 720–2 152	A 20.7; J 14.6	3.8**	A 19.7; J 16.1	2.5*
2 152–2 700	A 18.7; J 13.0	3.7**	A 15.2; J 12.4	2.4*
2 700–3 400	A 13.3; J 10.1	2.9*	A 14.7; J 11.3	2.7*
3 400–4 370	A 10.6; J 4.1	2.5	A 8.8; J 3.9	1.7
4 370–5 500	A 1.9; J .60	1.2	A 1.3; J .5	1.9

Note: N = 20 *p < .05, **p < .01, ***p < .001; A = anger; J = joy; All significance levels are 2-tailed.

the words 'accept' and 'deal' sometimes did disturb the subjects' ease of saying the sentence with a tone of anger. Such cases were not taken into account for statistical analyses. However, this fact points to the influence of the semantic content on vocal emotional expression. Most of the subjects reported that emotionally congruent semantic content could considerably help produce appropriate tone of voice. The authors also repeatedly noticed that in the subjects' spontaneous verbal expression, the emotion words were usually said on an emotionally congruent tone.

Conclusion

In spite of remarkable individual differences in vocal tract configurations, it appears that vocal expression of emotions exhibits similar patterning of vocal parameters. The similarities may be partly due to the physiological factors and partly to the contextually driven vocal adaptations governed by stereotypical representations of emotional voice patterns. Future research in this domain may further clarify the influence of cultural and socio-linguistic factors on intra-subject patterning of vocal parameters.

Acknowledgements

The authors thank Jacques Terken, Technische Universiteit Eindhoven, Nederland, for his constructive critical remarks. This article was carried out in the framework of COST 258.

References

Banse, R. and Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*, 614–636.

Hassal, J.H. and Zaveri, K. (1979). *Acoustic Noise Measurements*. Büel and Kjaer.

Keller, E. (1994). *Signal Analysis for Speech and Sound*. InfoSignal.

Mendolia, M. and Kleck, R.E. (1993). Effects of talking about a stressful event on arousal: Does what we talk about make a difference? *Journal of Personality and Social Psychology*, *64*, 283–292.

Pittam, J. (1987). Discrimination of five voice qualities and prediction of perceptual ratings. *Phonetica*, *44*, 38–49.

Pittam, J. and Gallois C. (1986). Predicting impressions of speakers from voice quality acoustic and perceptual measures. *Journal of Language and Social Psychology*, *5*, 233–247.

Popov, V.A., Simonov, P.V. Frolov, M.V. *et al*. (1971). Frequency spectrum of speech as a criterion of the degree and nature of emotional stress. (Dept. of Commerce, JPRS 52698.) *Zh. Vyssh. Nerv. Dieat*., (*Journal of Higher Nervons Activity*) *1*, 104–109.

Scherer, K.R. (1981). Vocal indicators of stress. In J. Darby (ed.), *Speech Evaluation in Psychiatry* (pp. 171–187). Grune and Stratton.

Scherer, K.R. (1989). Vocal correlates of emotional arousal and affective disturbance. *Handbook of Social Psychophysiology* (pp. 165–197). Wiley.

Scherer, K.R. (1992). On social representations of emotional experience: Stereotypes, prototypes, or archetypes? In M.V.H Cranach, W. Doise, and G. Mugny (eds), *Social Representations and the Social Bases of Knowledge* (pp. 30–36). Huber.

Scherer, K.R. (1993). Neuroscience projections to current debates in emotion psychology. *Cognition and Emotion*, *7*, 1–41.

Scherer, K.R. and Kappas, A. (1988). Primate vocal expression of affective state. In D.Todt, P.Goedeking, and D. Symmes (eds), *Primate Vocal Communication* (pp. 171–194). Springer-Verlag.

Scherer, K.R., Ladd, D.R., and Silverman, K.E.A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America*, *76*, 1346–1356.

Scherer, K.R. and Zei, B. (1988). Vocal indicators of affective disorders. *Psychotherapy and Psychosomatics*, *49*, 179–186.

Simonov, P.V., Frolov, M.V., and Ivanov E.A. (1980). Psychophysiological monitoring of operator's emotional stress in aviation and astronautics. *Aviation, Space, and Environmental Medicine*, January 1980, 46–49.

Williams, C.E. and Stevens, K.N. (1972). Emotion and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, *52*, 1238–1250.

Williams, C.E. and Stevens, K.N. (1981). Vocal correlates of emotional states. In J.K. Darby (ed.), *Speech Evaluation in Psychiatry* (pp. 221–240). Grune and Statton.

Zei, B. and Archinard, M. (1998). La variabilité du rythme cardiaque et la différentiation prosodique des émotions, *Actes des XXIIèmes Journées d'Etudes sur la Parole* (pp. 167–170). Martigny.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

24

The Role of Pitch and Tempo in Spanish Emotional Speech

Towards Concatenative Synthesis

Juan Manuel Montero Martínez ,[1] Juana M. Gutiérrez Arriola,[1] Ricardo de Córdoba Herralde,[1], Emilia Victoria Enríquez Carrasco[2] and José Manuel Pardo Muñoz[1]

[1]Grupo De Tecnología del Habla (GTH), ETSI Telecomunicación, Universidad Politécnica de Madrid (UPM), Ciudad Universitaria s/n, 28040 Madrid, Spain

[2] Departamento de Lengua Española y Lingüística General, Universidad Nacional de Educación a Distancia (UNED), Ciudad Universitaria s/n, 28040 Madrid, Spain

juancho@die.upm.es

Introduction

The steady improvement in synthetic speech intelligibility has focused the attention of the research community on the area of naturalness. Mimicking the diversity of natural voices is the aim of many current speech investigations. Emotional voice (i.e., speech uttered under an emotional condition or simulating an emotional condition, or under stress) has been analysed in many papers in the last few years: Montero *et al*. (1999a), Koike *et al*. (1998), Bou-Ghazade and Hansen (1996), Murray and Arnott (1995).

The VAESS project (TIDE TP 1174: Voices Attitudes and Emotions in Synthetic Speech) developed a portable communication device for disabled persons. This communicator used a multilingual formant synthesiser that was specially designed to be capable not only of communicating the intended words, but also of portraying the emotional state of the device user by vocal means. The evaluation of this project was described in Montero *et al*. (1998). The GLOVE voice source used in VAESS allowed controlling Fant's model parameters as described in Karlsson (1994). Although this improved source model could correctly characterise several voices and emotions (and the improvements were clear when synthesising a happy 'brilliant' voice), the 'menacing' cold angry voice had such a unique quality that it was impossible to simulate it in the rule-based VAESS synthesiser. This

led to a synthesis of a hot angry voice, different from the available database examples.

Taking that into account, we considered that a reasonable step towards improving the emotional synthesis was the use of a concatenative synthesiser, as in Rank and Pirker (1998), while taking advantage of the capability of this kind of synthesis to copy the quality of a voice from a database (without an explicit mathematical model).

The VAESS Project: SES database and Evaluation Results

As part of the VAESS project, the Spanish Emotional Speech database (SES) was recorded. It contains two emotional speech recording sessions played by a professional male actor in an acoustically treated studio. Each recorded session includes 30 words, 15 short sentences and three paragraphs, simulating three basic or primary emotions (sadness, happiness and anger), one secondary emotion (surprise) and a neutral speaking style (in the VAESS project the secondary emotion was not used). The text uttered by the actor did not convey any intrinsic emotional content. The recorded database was phonetically labelled in semiautomatic manner. The assessment of the natural voice aimed to judge the appropriateness of the recordings as a model for readily recognisable emotional synthesised speech. Fifteen normal listeners, both men and women of different ages (e.g. between 20 and 50) were selected from several social environments; none of them was used to synthetic speech.

The stimuli contained five emotionally neutral sentences from the database. As three emotions and a neutral voice had to be evaluated (the test did not include surprise examples), 20 different recordings per listener and session were used (only one session per subject was allowed). In each session, the audio recordings of the stimuli were presented to the listener in a random way. Each piece of text was played up to three times.

Table 24.1 shows that the subjects had no difficulty in identifying the emotion that was simulated by the professional actor, and the diagonal numbers (in bold) are clearly above the chance level (20%). A Chi-square test refutes the null hypothesis (with $p < 0.05$), i.e. these results, with a confidence level above 95%, could not have been obtained from a random selection experiment.

Table 24.1 Confusion matrix for natural voice evaluation test (recognition rate in %)

Identified emotion: Synthesised emotion	Neutral	Happy	Sad	Angry	Unidentified
Neutral	**89.3**	1.3	1.3	3.9	3.9
Happy	17.3	**74.6**	1.3	1.3	5.3
Sad	1.3	0.0	**90.3**	1.3	3.9
Angry	0.0	1.3	2.6	**89.3**	6.6

Note: Bold = diagonal number.

Analysing the results on a sentence-by-sentence basis (not emotion-by-emotion), none of them was significantly worse recognised (the identification rate varied from 83.3% to 93.3%).
A similar test evaluating the formant-based synthetic voice developed in the VAESS project is shown in Table 24.2. A Chi-square test also refutes the null hypothesis with $p < 0.05$, but evaluation results with synthesis are significantly worse than those using natural speech.

Copy-Synthesis Experiments

In a new experiment towards improving synthetic voice by means of a concatentive synthesiser, 21 people listened to three copy-synthesis sentences in a random-order forced-choice test (also including a 'non-identifiable' option) as in Heuft *et al.* (1996). In this copy-synthesis experiment, we used a concatenative synthesiser with both diphones (segmental information) and prosody (pitch and tempo) from natural speech. The confusion matrix is shown in Table 24.3.

The copy-synthesis results, although significantly above random-selection level using a Student's test ($p > 0.95$), were significantly below natural recording rates using a Chi-square test. This decrease in the recognition score can be due to several factors: the inclusion of a new emotion in the copy-synthesis test, the use of an automatic process for copying and stylising the prosody (pitch and tempo) linearly, and the distortion introduced by the prosody modification algorithms. It is remarkable that the listeners evaluated cold anger re-synthesised sentences significantly

Table 24.2 Confusion matrix for format-synthesis voice evaluation (recognition rate in %)

Identified emotion: Synthesised emotion	Neutral	Happy	Sad	Angry	Unidentified
Neutral	**58.6**	0.0	29.3	10.6	1.3
Happy	24.0	**46.6**	9.3	2.6	17.3
Sad	9.3	0.0	**82.6**	3.9	3.9
Angry	21.3	21.3	1.3	**42.6**	13.3

Note: Bold = diagonal number.

Table 24.3 Copy-synthesis evaluation test (recognition rate in %)

Identified emotion: Synthesised emotion:	Neutral	Happy	Sad	Surprised	Angry	Unidentified
Neutral	**76.2**	3.2	7.9	1.6	6.3	4.8
Happy	3.2	**61.9**	9.5	11.1	7.9	6.3
Sad	3.2	0.0	**81.0**	4.8	0.0	11.1
Surprised	0.0	7.9	1.6	**90.5**	0.0	0.0
Angry	0.0	0.0	0.0	0.0	**95.2**	4.8

Note: Bold = diagonal number.

above natural recordings (which means that the concatenation distortion made the voice even more menacing).

Table 24.4 shows the evaluation results of an experiment with mixed-emotion copy-synthesis (diphones and prosody are copied from two different emotional recordings; e.g., diphones could be extracted from a neutral sentence and its prosody is modified according to the prosody of a happy recording).

As we can clearly see, in this database cold anger was not prosodically marked, and happiness, although characterised by a prosody (pitch and tempo) that was significantly different from the neutral one, had more recognisable differences from a segmental point of view.

It can be concluded that modelling tempo and pitch of emotional speech are not enough to make a synthetic voice as recognisable as natural speech in the SES database (it does not convey enough emotional information in the parameters that can be easily manipulated in diphone-based concatenative synthesis). Finally, cold anger could be classified as an emotion signalled mainly by segmental means, surprise as a prosodically signalled emotion, while sadness and happiness have important prosodic and segmental components (in sadness tempo and pitch are predominant; happiness is more easy to recognise by means of the characteristics included in the diphone set).

Automatic-Prosody Experiment

Using the prosodic analysis (pitch and tempo) described in Montero *et al.* (1998) from the same database, we created an automatic emotional prosodic module to verify the segmental vs. supra-segmental hypothesis. Combining this synthetic prosody (obtained from paragraph recordings) with optimal-coupling diphones (taken from the short sentence recordings), we carried out an automatic-prosody test. The results are shown in Table 24.5.

The differences between this final experiment and the first copy-synthesis are significant (using a Chi-square test with 4 degrees of freedom and $p > 0.95$), due to the bad recognition rate for surprise. On a one-by-one basis, and using a Student's

Table 24.4 Prosody vs. segmental quality test with mixed-emotions (recognition rate in %)

Identified emotion:		Neutral	Happy	Sad	Surprised	Angry	Unidentified
Diphones	Prosody						
Neutral	Happy	**52.4**	**19.0**	11.9	4.8	0.0	11.9
Happy	Neutral	**4.8**	**52.4**	0.0	9.5	26.2	7.1
Neutral	Sad	**23.8**	0.0	**66.6**	0.0	2.4	7.1
Sad	Neutral	**26.2**	2.4	**45.2**	4.8	0.0	21.4
Neutral	Surprised	**2.4**	16.7	2.4	**76.2**	0.0	2.4
Surprised	Neutral	**19.0**	11.9	21.4	**9.5**	4.8	33.3
Neutral	Angry	**11.9**	19.0	19.0	23.8	**7.1**	19.0
Angry	Neutral	**0.0**	0.0	0.0	2.4	**95.2**	2.4

Note: Bold = diagonal number.

Table 24.5 Automatic prosody experiments (recognition rate in %)

Identified emotion: Synthesised emotion:	Neutral	Happy	Sad	Surprised	Angry	Unidentified
Neutral	**72.9**	0.0	15.7	0.0	0.0	11.4
Happy	12.9	**65.7**	4.3	7.1	1.4	8.6
Sad	8.6	0.0	**84.3**	0.0	0.0	8.6
Surprised	1.4	27.1	1.4	**52.9**	0.0	17.1
Angry	0.0	0.0	0.0	1.4	**95.7**	2.9

Note: Bold = diagonal number.

test, anger, happiness, neutral and sadness results are not significantly different from the copy-synthesis test ($p < 0.05$). An explanation for all these facts is that the prosody in this experiment was trained with the paragraph style, and it had never been evaluated for surprise before (both paragraphs and short sentences were assessed in the VAESS project for sadness, happiness, anger, and neutral styles).

There is an important improvement in happiness recognition rates when using both happy diphones and happy prosody, but the difference is not significant with a 0.95 threshold and a student's distribution.

Conclusion

The results of our experiments show that some of the emotions simulated by the speaker in the database (sadness and surprise) are signalled mainly by pitch and temporal properties and others (happiness and cold anger) mainly by acoustic properties other than pitch and tempo, either related to source characteristics such as spectral balance or to vocal tract characteristics such as lip rounding.

According to the experiments carried out, an improved emotional synthesiser must transmit the emotional information through variations in the prosodic model and by means of an increased number of emotional concatenation units (in order to be able to cover the prosodic variability that characterise some emotions such as surprise).

As emotions cannot be transmitted using only supra-segmental information and as segmental differences between emotions play an important role in their recognisability, it would be interesting to consider that emotional speech synthesis could be a transformation of the neutral voice. By applying transformation techniques (parametric and non-parametric) as in Gutiérrez-Arriola *et al.* (1997), new emotional voices could be developed for a new speaker without recording a new complete emotional database. These transformations should be applied to both voice source and vocal tract. A preliminary emotion-transfer experiment with a glottal source that is modelled as a mixture of a polynomial function and a certain amount of additive noise, has shown that this could be the right solution.

The next step will be the development of a fully automatic emotional diphone concatenation synthesiser. As the range of the pitch variations is larger than for neutral-style speech, the use of several units per diphone must be considered in order to cover this increased range. For more details, see Montero, *et al.* (1999b).

References

Bou-Ghazade, S. and Hansen, J.H.L. (1996). Synthesis of stressed speech from isolated neutral speech using HMM-based models. *Proceedings of International Conference on Spoken Language Processing* (pp. 1860–1863). Philadelphia.

Gutiérrez-Arriola, J., Giménez de los Galanes, F.M., Savoji, M.H., and Pardo, J.M. (1997). Speech synthesis and prosody modification using segmentation and modelling of the excitation signal. *Proceedings of European Conference on Speech Communication and Technology*, Vol. 2 (pp. 1059–1062). Rhodes, Greece.

Heuft, B., Portele, T., and Rauth, M. (1996). Emotions in time domain synthesis. *Proceedings of International Conference on Spoken Language Processing* (pp. 1974–1977). Philadelphia.

Karlsson, I. (1994). Controlling voice quality of synthetic speech. *Proceedings of International Conference on Spoken Language Processing* (pp. 1439–1442). Yokohama.

Koike, K. Suzuki, H., and Saito, H. (1998). Prosodic parameters in emotional speech. *Proceedings of International Conference on Spoken Language Processing*, Vol. 3 (pp. 679–682). Sydney.

Montero, J.M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., and Pardo, J.M. (1999a). Analysis and modelling of emotional speech in Spanish. *Proceedings of the International Congress of Phonetic Sciences*, Vol. 2 (pp. 957–960). San Francisco.

Montero, J.M., Gutiérrez-Arriola, J., Colás, J., Macías-Guarasa, J., Enríquez, E., and Pardo, J.M. (1999b). Development of an emotional speech synthesiser in Spanish. *Proceedings of European Conference on Speech Communication and Technology* (pp. 2099–2102). Budapest.

Montero, J.M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S., and Pardo, J.M. (1998). Emotional speech synthesis: from speech database to TTS. *Proceedings of International Conference on Spoken Language Processing*, Vol. 3 (pp. 923–926). Sydney.

Murray, I.R. and Arnott, J.L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, *16*, 359–368.

Rank, E. and Pirker, H. (1998). Generating emotional speech with a concatenative synthesiser. *Proceedings of International Conference on Spoken Language Processing*, Vol. 3 (pp. 671–674). Sydney.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

25

Voice Quality and the Synthesis of Affect

Ailbhe Ní Chasaide and Christer Gobl
Centre for Language and Communication Studies, Trinity College, Dublin, Ireland
anichsid@tcd.ie

Introduction

Speakers use changes in 'tone of voice' or voice quality to communicate their attitude, moods and emotions. In a related way listeners tend to make inferences about an unknown speaker's personality on the basis of voice quality. Although changes in voice quality can effectively alter the overall meaning of an utterance, these changes serve a paralinguistic function and do not form part of the contrastive code of the language, which has tended to be the primary focus of linguistic research. Furthermore, written representations of language carry no information on tone of voice, and this undoubtedly has also contributed to the neglect of this area. Much of what we do know comes in the form of axioms, or traditional impressionistic comments which link voice qualities to specific affects, such as the following: creaky voice – boredom; breathy voice – intimacy; whispery voice – confidentiality; harsh voice – anger; tense voice – stress (see, for example, Laver, 1980). These examples pertain to speakers of English: although the perceived affective colouring attaching to a particular voice quality may be universal in some cases, for the most part they are thought to be language and culture specific.

Researchers in speech synthesis have recently shown an interest in this aspect of spoken communication. Now that synthesis systems are often highly intelligible, and have a reasonably acceptable intrinsic voice quality, a new goal has become that of making the synthesised voice more expressive and to impart the possibility of personality, in a way that might more closely approximate human speech.

Difficulties in Voice Quality Research

This area of research presents many difficulties. First of all, it is complex: voice quality is not only used for the paralinguistic communication of affect, but varies also as a function of linguistic and extralinguistic factors (see Gobl and Ní

Chasaide, this issue). Unravelling these varying strands which are simultaneously present in any given utterance is not a trivial task. Second, and probably the principal obstacle in tackling this task, is the difficulty in obtaining reliable glottal source data. Appropriate analysis tools are not generally available. Thus, most of the research on voice quality, whether for the normal or the pathological voice, has tended to be auditorily based, employing impressionistic labels, e.g. harsh voice, rough voice, coarse voice, etc. This approach has obvious pitfalls. Terms such as these tend to proliferate, and in the absence of analytic data to characterise them, it may be impossible to know precisely what they mean and to what degree they may overlap. For example: is harsh voice the same as rough voice, and if not, how do they differ? Different researchers are likely to use different terms, and it is difficult to ensure consistency of usage. The work of Laver (1980) has been very important in attempting to standardise usage within a descriptive framework, underpinned where possible by physiological and acoustic description. See also the work by Hammarberg (1986) on pathological voice qualities.

Most empirical work on the expression of moods and emotions has concentrated on the more measurable aspects, F0 and amplitude dynamics, with considerable attention also to temporal variation (see for example the comprehensive analyses reviewed in Scherer, 1986 and in Kappas, *et al.* 1991). Despite its acknowledged importance, there has been little empirical research on the role of voice quality. Most studies have involved analyses of actors' simulations of emotions. This obviously entails a risk that stereotypical and exaggerated samples are being obtained. On the other hand obtaining a corpus of spontaneously produced affective speech is not only difficult, but will lack the control of variables that makes for detailed comparison. At the ISCA 2000 Workshop on Speech and Emotion, there was considerable discussion of how suitable corpora might be obtained. It was also emphasised that for speech technology applications such as synthesis, the small number of emotional states typically studied (e.g., anger, joy, sadness, fear) are less relevant than the milder moods, states and attitudes (e.g., stressed, bored, polite, intimate, etc.) for which very little is known.

In the remainder of this chapter we will present some exploratory work in this area. We do not attempt to analyse emotionally coloured speech samples. Rather, the approach taken is to generate samples with different voice qualities, and to use these to see whether listeners attach affective meaning to individual qualities. This work arises from a general interest in the voice source, and in how it is used in spoken communication. Therefore, to begin with, we illustrate attempts to provide acoustic descriptions for a selection of the voice qualities defined by Laver (1980). By re-synthesising these qualities, we can both fine-tune our analytic descriptions and generate test materials to explore how particular qualities may cue affective states and attitudes. Results of some pilot experiments aimed at this latter question are then discussed.

Acoustic Profiles of Particular Voice Qualities

Analysis has been carried out for a selected number of voice qualities, within the framework of Laver (1980). These analyses were based on recordings of sentences

and passages spoken with the following voice qualities: modal voice, breathy voice, whispery voice, creaky voice, tense voice and lax voice. The subject was a male phonetician, well versed in the Laver system, and the passages were produced without any intended emotional content.

The analytic method is described in the accompanying chapter (Gobl and Ní Chasaide Chapter 27, this issue) and can be summarised as follows. First of all, interactive inverse filtering is used to cancel out the filtering effect of the vocal tract. The output of the inverse filter is an estimate of the differentiated glottal source signal. A four-parameter model of differentiated glottal flow (the LF model, Fant, Liljencrants and Lin, 1985) is then matched to this signal by interactive manipulation of the model parameters. To capture the important features of the source signal, parameters are measured from the modelled waveform: EE, RA, RK and RG, which are described in Gobl and Ní Chasaide (this issue). For a more detailed account of these techniques and of the glottal parameters measured, see also Gobl and Ní Chasaide (1999a).

Space would not permit a description of individual voice qualities here. Figure 25.1, however, illustrates schematic source spectra for four voice qualities. These

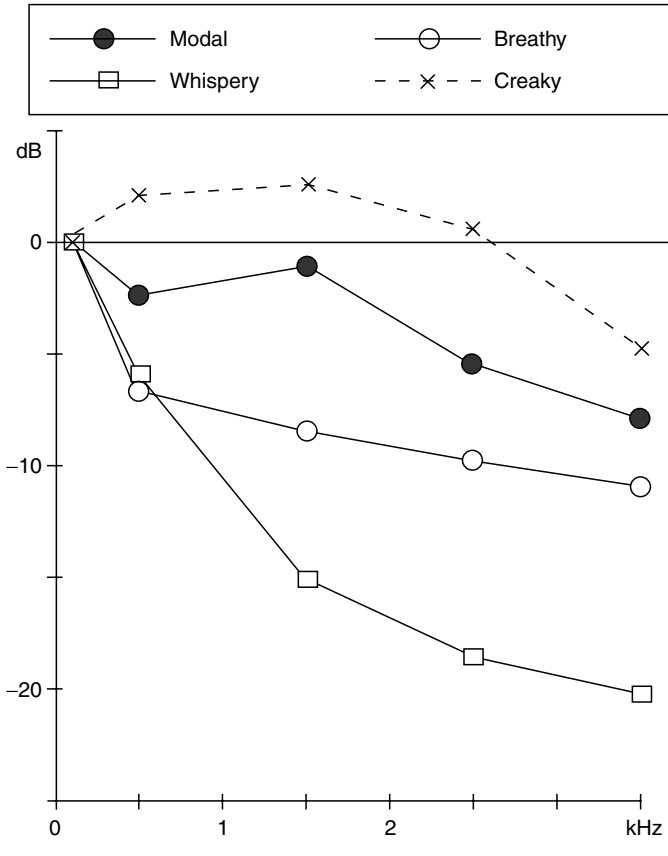


Figure 25.1 Schematic source spectra taken from the midpoint of a stressed vowel, showing the deviation from a -12 dB/octave spectral slope

were based on measurements of source spectra, obtained for the midpoint in a stressed vowel. Some elaboration on individual voice qualities can be found in Gobl (1989) and Gobl and Ní Chasaide (1992). It should be pointed out that the differences in voice qualities cannot be expressed in terms of single global spectral transformations. They involve rather complex context-dependent transformations. This can readily be appreciated if one bears in mind that a voluntary shift in voice quality necessarily interacts with the speaker's intrinsic voice quality and with the types of glottal source modulations described in Gobl and Ní Chasaide (this issue), which relate to the segmental and the suprasegmental content of utterances.

Re-synthesis of Voice Qualities

In order to resynthesise these voice qualities, we have employed the modified LF model implementation of KLSYN88a (Sensimetrics Corporation, Boston, MA; for a description, see Klatt and Klatt, 1990). As indicated above, in our source analyses we have worked mainly with the parameters EE, RA, RK and RG. Although the control parameters of the source model in KLSYN88a are different, they can be derived from our analysis parameters. The following source parameters were varied: F0, AV (amplitude of voicing, derived from EE), TL (spectral tilt, derived from RA and F0), OQ (open quotient, derived from RG and RK), SQ (speed quotient, derived from RK). Aspiration noise (AH) is not quantifiable with the analysis techniques employed, and subsequently, in our resynthesis we have needed to experiment with this parameter, being guided in the first instance by our own auditory judgements. A further parameter that was manipulated in our resynthesis was DI (diplophonia), which is a device for achieving creakiness. This parameter alters every second pulse by shifting the pulse towards the preceding pulse, as well as reducing the amplitude. The extent of the shift (as a percentage of the period) as well as the amount of amplitude reduction is determined by the DI value.

Re-synthesis offers the possibility of exploring the perceptual correlates of changes to source parameters, individually or in combination. One such study, reported in Gobl and Ní Chasaide (1999b), examined the source parameter settings for breathy voice perception. A somewhat surprising finding concerned the relative importance of the TL (spectral tilt) and AH (aspiration noise) parameters. An earlier study (Klatt and Klatt, 1990) had concluded that spectral tilt was not a strong cue to breathiness whereas aspiration noise was deemed to play a major role. Results of our study suggest rather that spectral balance properties are of crucial importance. TL, which determines the relative amplitude of the higher frequencies, emerges as the major perceptual cue. The parameters OQ, SQ and BW, which on their own have little effect, are perceptually quite important when combined. Together, these last determine the spectral prominence of the very lowest frequencies. AH emerged in this study as playing a relatively minor role.

On the basis of re-synthesised voice qualities one can explore the affective colouring that different qualities evoke for the listener. In and experiment reported in Gobl and Ní Chasaide (2000) the Swedish utterance *ja adjö* [ˈjaaˈjø] was synthesised with the following voice qualities: modal voice, tense voice, breathy voice,

creaky voice, whispery voice, lax-creaky voice and harsh voice. Unlike the first five, source values for the last two voice qualities were not directly based on prior analytic data. In the case of harsh voice, we attempted to approximate as closely as is permitted by KLSYN88a the description of Laver (1980). Lax-creaky voice represents a departure from the Laver system. Creaky voice in Laver's description involves considerable glottal tension, and this is what would be inferred from the results of our acoustic analyses. Note, for example, the relatively flat source spectrum in the creaky voice utterance in Figure 25.1 above, a feature one would expect to find for tense voice. Intuitively, we felt that there is another type of creaky voice one frequently hears, one which auditorily sounds like a creaky version of lax voice. In our experiments we therefore included such an exemplar.

Synthesis of the modal utterance was based on a prior pulse-by-pulse analysis of a natural recording, and the other voice qualities were created from it by manipulations of the synthesis parameters described above. Because of space constraints, it is not possible to describe here the ranges of values used for each parameter, and the particular modifications for the individual voice qualities. However, the reader is referred to the description provided in Gobl and Ní Chasaide (2000). Two things should be noted here. First of all, the modifications from modal voice were not simply global changes, but included dynamic changes of the type alluded to in Gobl and Ní Chasaide (this issue) such as onset/offset and stress related differences. Second, F0 manipulations were included only to the extent that they were deemed an integral aspect of a particular voice quality. Thus, for tense voice, F0 was increased by 5 Hz and for the creaky and lax-creaky voice qualities, F0 was lowered by 20 to 30 Hz. The large changes in F0 which are described in the literature as correlates of particular emotions were intentionally not introduced initially.

Perceived Affective Colouring of Particular Voice Qualities

A series of short perception tests elicited listener's responses in terms of pairs of opposite affective attributes: *relaxed*/*stressed*, *content*/*angry*, *friendly*/*hostile*, *sad*/*happy*, *bored*/*interested*, *intimate/formal*, *timid*/*confident* and *afraid*/*unafraid*. For each pair of attributes the different stimuli were rated on a seven-point scale, ranging from -3 to $+3$. The midpoint 0, indicated that neither of the pair of attributes was detected, whereas extent of any deviation from zero showed the degree to which one or other of the two attributes was deemed present. For each pair of attributes, listeners' responses were averaged for the seven individual test stimuli. In Figure 25.2, the maximum strength with which any of the attributes was detected for all the voice qualities is shown in absolute terms as deviations from 0 (= no perceived affect) to 3 (i.e. $+3$ or -3 = maximally perceived).

Listeners' responses do suggest that voice quality variations alone can alter the affective colouring of an utterance. The most strongly detected attributes were *stressed*, *relaxed*, *angry*, *bored*, *formal*, *confident*, *hostile*, *intimate* and *content*. The least strongly detected were attributes such as *happy*, *unafraid*, *friendly* and *sad*. By and large, these latter attributes differ from the former in that they represent emotions rather than milder condition such as speaker state and attitudes. The striking exception, of course, is *angry*.

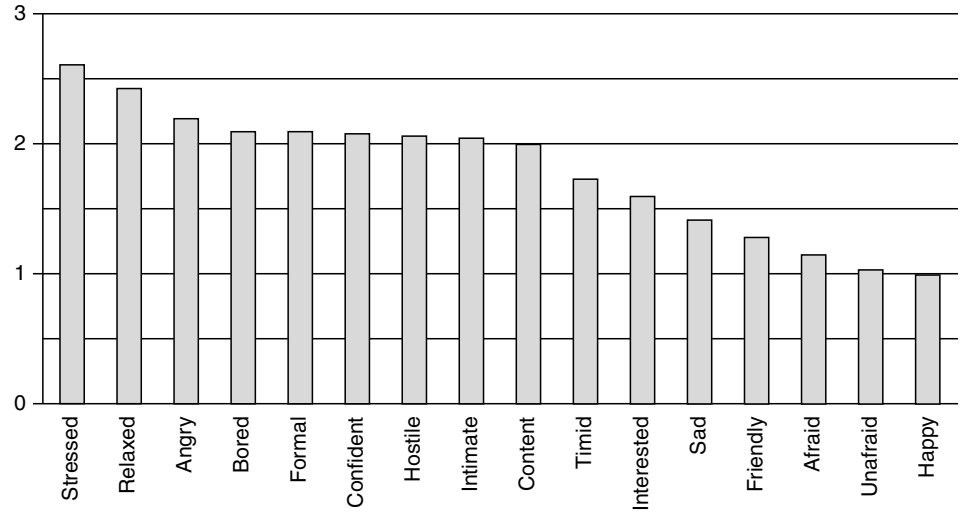


Figure 25.2 Maximum ratings for perceived strength (shown on y-axis) of affective attributes for any voice quality, shown as deviations from 0 (= no perceived affect) to 3 (maximally perceived)

In Figure 25.3, ratings for the affective attributes associated with the different voice qualities can be seen. Here again, 0 equals no perceived affect and +/−3 indicate a maximal deviation from neutral. Note that the positive or negative sign is in itself arbitrary. Although traditional observations have tended to link individual voice qualities to specific attributes (e.g., creaky voice and boredom) it is clear from this figure that there is no one-to-one mapping from quality to attribute. Rather a voice quality tends to be associated with a constellation of attributes: for example, tense voice gets high ratings for *stressed*, *angry*, *hostile*, *formal* and *confident*. Some of these attributes are clearly related, some less obviously so. Although the traditional observations are borne out to a reasonable extent, these results suggest some refinements. Breathy voice, traditionally regarded as the voice quality associated with intimacy, is less strongly associated with it than is lax-creaky voice. Furthermore, creaky voice scored less highly for *bored* (with which it is traditionally linked) than did the lax-creaky quality, which incidentally was rated very highly also for the attributes *relaxed* and *content*.

Voice Quality and F0 to Communicate Affect

As mentioned earlier, most experimental work to date on the expression of emotion in speech has focused particularly on F0 dynamics. These large F0 excursions, described in the literature for specific emotions, were not included in the initial series of tests. In a follow-up study (Bennett, 2000) these same basic stimuli have been used, with and without large F0 differences. The F0 contours used were modelled on those presented in Mozziconacci (1995) for the emotions *joy*, *boredom*, *anger*, *sadness*, *fear*, *indignation*, and which were based on descriptive analyses. The

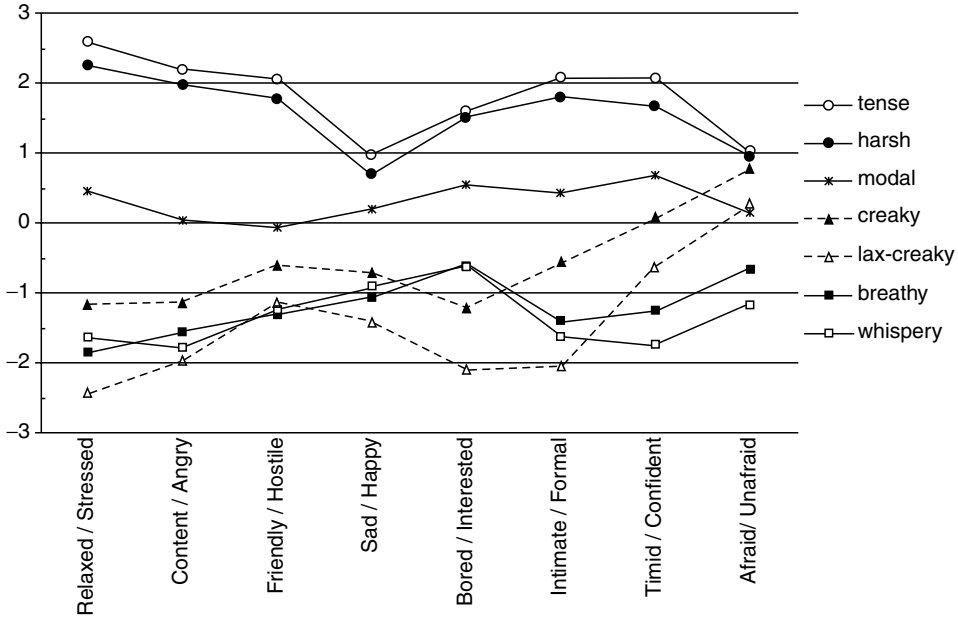


Figure 25.3 Relative ratings of perceived strength (shown on y-axis) for pairs of opposite affective attributes across all voice qualities. 0 = no perceived affect, +/ − 3 = maximally perceived

aim in this instance was to explore the extent to which voice quality modification might enhance the detection of affective states beyond what can be elicited through F0 manipulations alone. The fundamental frequency contours provided in Mozziconacci (1995) are illustrated in Figure 25.4.

The F0 of the modal stimulus in the earlier experiment (Gobl and Ní Chasaide, 2000) was used as the 'neutral' reference here. Mozziconacci's non-neutral contours were adapted to the F0 contour of this reference, by relative scaling of the F0 values. From the neutral reference utterance, six stimuli were generated by simply changing the F0 contour, corresponding to Mozziconacci's non-neutral contours. From these six, another six stimuli were generated which differed in terms of voice quality. Voice qualities from the first experiment were paired to F0 contours associated with particular emotions as follows: the F0 contour for *joy* was paired with tense voice quality, *boredom* with lax-creaky voice, *anger* with tense voice, *sadness* with breathy voice, *fear* with whispery voice and *indignation* with harsh voice. The choice of voice quality to pair with a particular F0 contour was made partially on the basis of the earlier experiment, partially from suggestions in the literature and partially from intuition. It should be pointed out that source parameter values are not necessarily the same across large differences in F0. However, in this experiment no further adjustments were made to the source parameters.

The perception tests were carried out in essentially the same way as in the first experiment, but with the exclusion of the attributes *friendly/hostile* and *timid/confident*, and with the addition of the attribute *indignant*, which featured as one of the

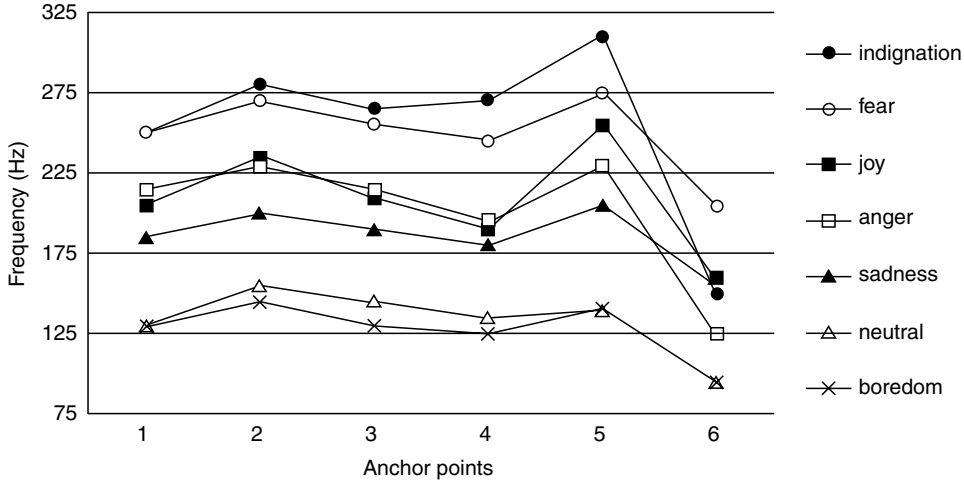


Figure 25.4 Fundamental frequency contours corresponding to different emotions, from Mozziconacci (1995)

attributes in Mozziconacci's study. Note that as there is no obvious opposite counterpart to *indignant*, ratings for this attribute were obtained on a four-point scale (0 to 3). Figure 25.5 displays the data in a way similar to Figure 25.2 and shows the maximum ratings for stimuli involving F0 manipulations only as white columns and for stimuli involving F0 + voice quality manipulations as black columns.

The stimuli which included voice quality manipulations were more potent in signalling the affective attributes, with the exception of *unafraid*. For a large number of attributes the difference in results achieved for the two types of stimuli is striking. The reason for the poor performance in the *unafraid* case is likely to be that the combination of whispery voice and the F0 contour resulted in an unnatural sounding stimulus.

Figure 25.6 shows results for the subset of affective states that featured in Mozziconacci's experiment, but only for those stimuli which were expected to evoke these states. Thus, for example, for stimuli that were intended to elicit *anger*, this figure plots how high these scored on the 'content/angry' scale. Similarly, for stimuli that were intended to elicit *boredom*, the figure shows how high they scored on the 'bored/interested' scale. Note that *joy* is here equated with *happy* and *fear* is equated with *afraid*. Responses for the neutral stimulus (the modal stimulus with a neutral F0 contour, which should in principle have no affective colouring) are also shown for comparison, as grey columns. As in Figure 25.5, the white columns pertain to stimuli with F0 manipulation alone and the black columns to stimuli with F0 + voice quality manipulation.

Results indicate again for all attributes, excepting fear, that the highest detection rates are achieved by the stimuli which include voice quality manipulation. In fact, the stimuli which involved manipulation of F0 alone achieve rather poor results. One of the most interesting things about Figure 25.6 concerns what it does not

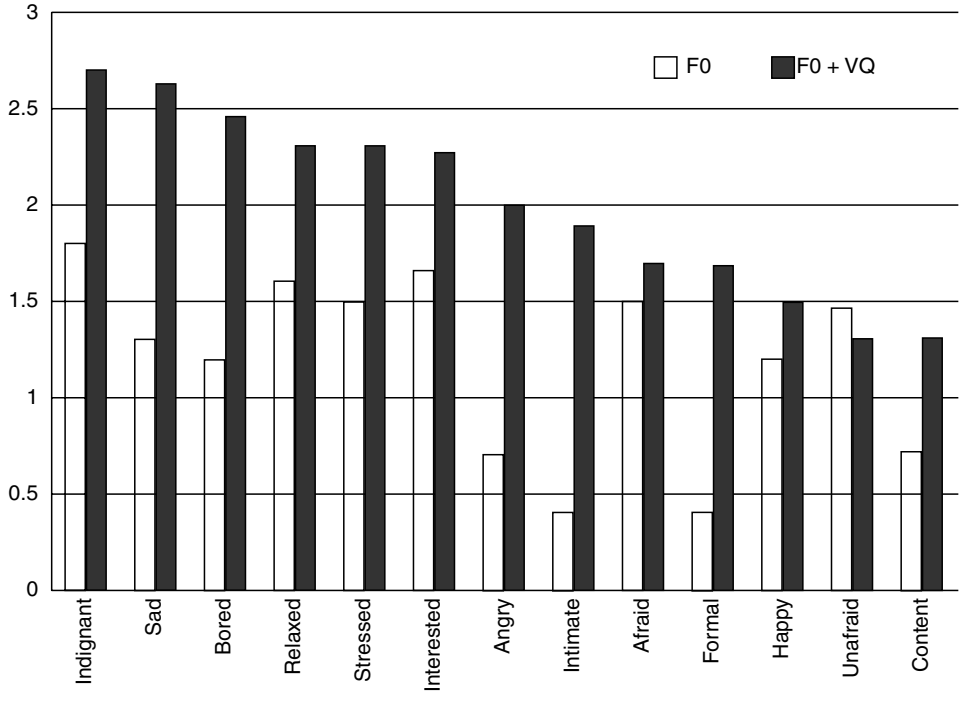


Figure 25.5 Maximum ratings for perceived strength (shown on y-axis) of affective attributes for stimuli where F0 alone (white) and F0 + voice quality (black) were manipulated. 0 = no perceived affect, 3 = maximally perceived

show. The highest rate of detection of a particular attribute was not always yielded by the stimulus which was intended/expected to achieve it. For example, the stimulus perceived as the most sad was not the expected one, which had breathy voice (frequently mentioned in connection with sad speech) with the '*sad*' F0 contour, but rather the lax-creaky stimulus with the '*bored*' F0 contour. As Mozziconacci's '*bored*' F0 contour differed only marginally from the neutral (see Figure 4) it seems likely that voice quality is the main determinant in this case. These mismatches can be useful in drawing attention to linkages one might not have expected.

Conclusion

These examples serve to illustrate the importance of voice quality variation to the global communication of meaning, but they undoubtedly also highlight how early a stage we are at in being able to generate the type of expressive speech that must surely be the aim for speech synthesis. This work represents only a start. In the future we hope to explore how F0, voice quality, amplitude and other features interact in the signalling of attitude and affect. In the case where a particular voice quality seems to be strongly associated with a given affect (e.g., tense voice and anger) it would be interesting to explore whether gradient, stepwise increases in

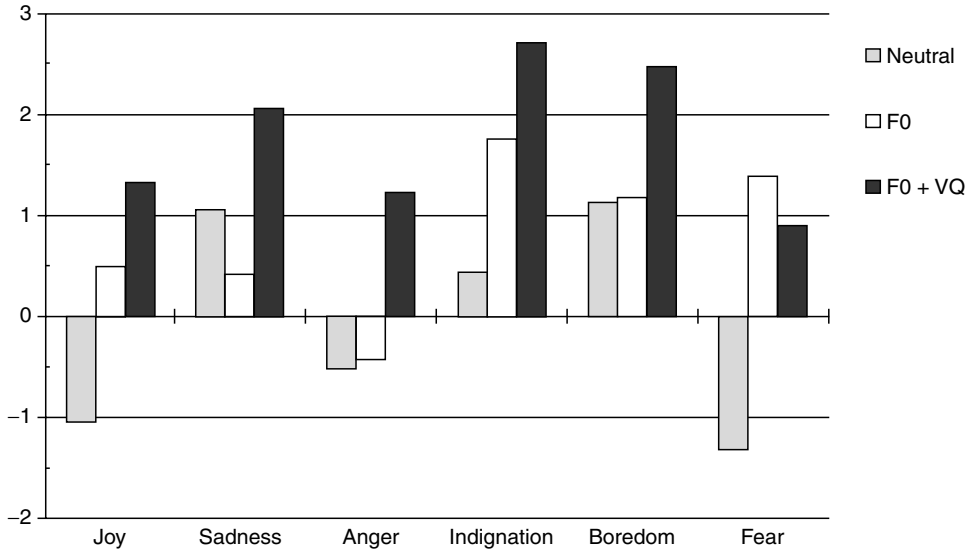


Figure 25.6 Ratings for perceived strength (shown on y-axis) of affective attributes for stimuli designed to evoke these states. Stimulus-type: manipulation of F0 alone (white), F0 + voice quality (black) and neutral (grey). 0 = no perceived affect, 3 = maximally perceived. Negative values indicate that the attribute was not perceived, and show rather the detection (and strength) of the opposite attribute

parameter settings yield correspondingly increasing degrees of anger. Similarly, it would be interesting to examine further the relationship between different types of creaky voice and boredom and other affects such as sadness.

A limitation we have encountered in our work so far concerns the fact that we have used different systems for analysis and synthesis. From a theoretical point of view they are essentially the same, but at a practical level differences lead to uncertainties in the synthesised output (see Mahshie and Gobl, 1999). Ideally, what is needed is a synthesis system that is directly based on the analysis system, and this is one goal we hope to work towards.

The question arises as to how these kinds of voice quality changes might be implemented in synthesis systems. In formant synthesis, provided there is a good source model, most effects should be achievable. In concatenative synthesis two possibilities present themselves. First of all, there is the possibility of frequency domain manipulation of the speech output signal to mimic source effects. A second possibility would be to record numerous corpora with a variety of emotive colourings – a rather daunting prospect.

In order to improve this aspect of speech synthesis, a better understanding is needed in two distinct areas: (1) we need more information on the rules that govern the transformation between individual voice qualities. It seems likely that these transformations do not simply involve global rules, but rather complex, context sensitive ones. Some of the illustrations in Gobl and Ní Chasaide (this issue) point in this direction. (2) We need to develop an understanding of the complex mappings between voice quality, F0 and other features and listeners' perception of

affect and attitude. The illustrations discussed in this chapter provide pointers as to where we might look for answers, not the answers themselves. In the first instance it makes sense to explore the question using semantically neutral utterances. However, when more is known about the mappings in (2), one would also be in a position to consider how these interact with the linguistic content of the message and the pragmatic context in which it is spoken.

These constitute a rather long-term research agenda. Nevertheless, any progress in these directions may bring about incremental improvements in synthesis, and help to deliver more evocative, colourful speech and breathe some personality into the machines.

Acknowledgements

The authors are grateful to COST 258 for the forum it has provided to discuss this research and its implications for more natural synthetic speech.

References

Bennett, E. (2000). *Affective Colouring of Voice Quality and F0 Variation*. MPhil. dissertation, Trinity College, Dublin.

Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), *4*, 1–13.

Gobl, C. (1989). A preliminary study of acoustic voice quality correlates. *STL-QPSR* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), *4*, 9–21.

Gobl, C. and Ní Chasaide, A. (1992). Acoustic characteristics of voice quality. *Speech Communication*, *11*, 481–490.

Gobl, C. and Ní Chasaide, A. (1999a). Techniques for analysing the voice source. In W.J. Hardcastle and N. Hewlett (eds), *Coarticulation: Theory, Data and Techniques* (pp. 300–321). Cambridge University Press.

Gobl, C. and Ní Chasaide, A. (1999b). Perceptual correlates of source parameters in breathy voice. *Proceedings of the XIVth International Congress of Phonetic Sciences* (pp. 2437–2440). San Francisco.

Gobl, C. and Ní Chasaide, A. (2000). Testing affective correlates of voice quality through analysis and resynthesis. In R. Cowie, E. Douglas-Cowie and M. Schröder (eds), *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research* (pp. 178–183). Belfast, Northern Ireland.

Hammarberg, B. (1986). Perceptual and acoustic analysis of dysphonia. *Studies in Logopedics and Phoniatrics 1*, Doctoral thesis, Huddinge University Hospital, Stockholm, Sweden.

Kappas, A., Hess, U., and Scherer, K.R. (1991). Voice and emotion. In R.S. Feldman and B. Rimé (eds), *Fundamentals of Nonverbal Behavior* (pp. 200–238). Cambridge University Press.

Klatt, D.H. and Klatt, L.C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87*, 820–857.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press.

Mahshie, J. and Gobl, C. (1999). Effects of varying LF parameters on KLSYN88 synthesis. *Proceedings of the XIVth International Congress of Phonetic Sciences* (pp. 1009–1012). San Francisco.

Mozziconacci, S. (1995). Pitch variations and emotions in speech. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Vol. 1 (pp. 178–181). Stockholm.

Scherer, K.R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*, 143–165.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

26

Prosodic Parameters of a 'Fun' Speaking Style

Kjell Gustafson and David House
Centre for Speech Technology, Department of Speech, Music and Hearing, KTH
Drottning Kristinas väg 31, 100 44 Stockholm, Sweden
kjellg | davidh @speech.kth.se
http://www.speech.kth.se/

Introduction

There is currently considerable interest in examining different speaking styles for speech synthesis (Abe, 1997; Carlson *et al*., 1992). In many new applications, naturalness and emotional variability have become increasingly important aspects. A relatively new area of study is the use of synthetic voices in applications directed specifically towards children. This raises the question as to what characteristics these voices should exhibit from a phonetic point of view.

It has been shown that there are prosodic differences between child-directed natural speech (CDS) and adult-directed natural speech (ADS). These differences often lie in increased duration and larger fundamental frequency excursions in stressed syllables of focused words when the speech is intended for children (Snow and Ferguson, 1977; Kitamura and Burnham, 1998; Sundberg, 1998). Although many studies have focused on speech directed to infants and on the implications for language acquisition, these prosodic differences have also been observed when parents read aloud to older children (Bredvad-Jensen, 1995). It could be useful to apply similar variation to speech synthesis for children, especially in the context of a fun and interesting educational programme.

The purpose of this chapter is to discuss the problem of how to arrive at prosodic parameters for voices and speaking styles that are suitable in full-scale text-to-speech systems for child-directed speech synthesis. Our point of departure is the classic prosodic parameters of F0 and duration. But intimately linked with these is the issue of the voice quality of voices used in applications directed to children.

Background and Goals

A central goal is to investigate how children react to prosodic variation which differs from default prosodic rules designed for text-to-speech applications directed to adults. The objective is to produce fun and appealing voices by determining what limits should be placed on the manipulation of duration and F0 in terms of both acoustic parameters and prosodic categories.

Another goal must be to arrive at a coherent voice and speaking style. A problem of current technology is that both of the generally available basic synthesis techniques have serious limitations when it comes to creating voices appropriate to the character that they are meant to represent. Concatenative synthesis relies on the existence of databases of recorded speech which necessarily reflect the sex and age of the speaker. In order to create a voice suitable for a child character (to be used, for example, as an animated character in a computer game), it is necessary to record a database of a child's speech. Unfortunately, databases of children's voices are uncommon, and creating them is time-consuming and expensive. Formant synthesis, on the other hand, offers the possibility of shaping the voice according to the requirements. However, knowledge of how to parameterise formant synthesis to reflect speaker characteristics in terms of physical make-up, target language and emotional and attitudinal state, is still at a fairly primitive stage. As a consequence, few convincing examples have been created. It is important to stress the close link between linguistic (e.g. prosodic) and paralinguistic factors (e.g. voice quality) when a coherent and convincing voice is to be achieved. The under-researched area of voice characteristics, consequently, is one that needs more attention if convincing voices are to be achieved, both in applications aimed at adults and, not least, when the target is to produce voices that will appeal to children. It is important that the voices are 'consistent', i.e. that segmental and prosodic characteristics match the character and the situation being portrayed, and likewise that the voice quality reflects both the person and the situation that are the target of the application.

Testing Prosodic Parameters

In a previous study, presented at Eurospeech '99 (House *et al.*, 1999), prosodic parameters were varied in samples of both formant and concatenative synthesis. An animated character (an astronaut originally created for an educational computer game by Levande Böcker i Norden AB) was adapted to serve as an interactive test environment for speech synthesis. The astronaut in a spacesuit inside a spaceship was placed in a graphic frame in the centre of the computer screen. Eight text fields in a vertical list on the right side of the frame were linked to sound files. By clicking on a text field the subjects could activate the sound file which also activated the visual animation. The animation began and ended in synchrony with the sound file. The test environment is illustrated in Figure 26.1.

Three sentences, appropriate for an astronaut, were synthesised using a developmental version of the Infovox 230 formant-based Swedish male voice and the

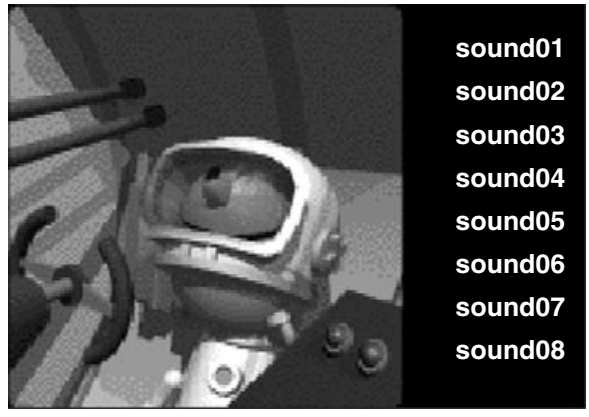


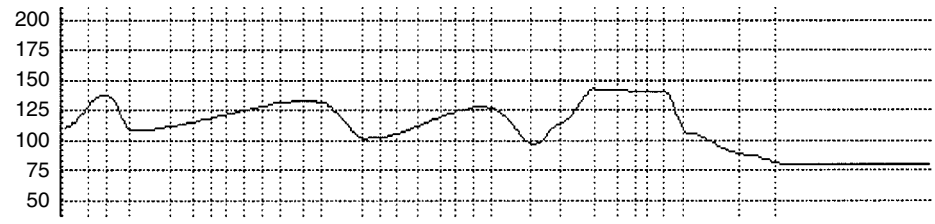
Figure 26.1 Illustration of the test environment

Infovox 330 concatenated diphone Swedish female voice. Four prosodically different versions of each sentence and each voice were synthesised: (1) a default version; (2) a version with a doubling of duration in the focused words; (3) a version with a doubling of the maximum F0 values in the focused words; and (4) a combination of 2 and 3. There were thus a total of eight versions of each sentence and 24 stimuli in all. The sentences are listed below with the focused words indicated in capitals.

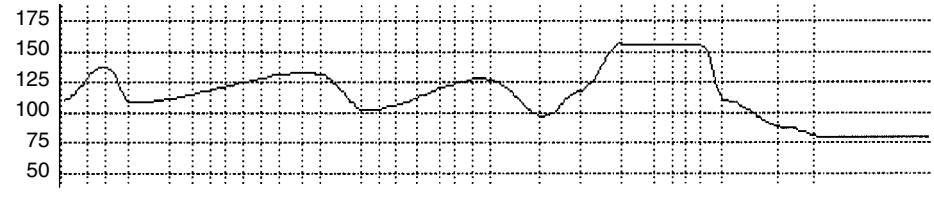
(1) Vill du följa MED mig till MARS? (Do you want to come WITH me to MARS?)
(2) Idag ska jag flyga till en ANNAN planet. (Today I'm going to fly to a DIFFERENT planet.)
(3) Det tar mer än TVÅ DAGAR att åka till månen. (It takes more than TWO DAYS to get to the moon.)

Figure 26.2 shows parameter plots for the formant synthesis version of sentence 2. As can be seen from the diagrams, the manipulation was localised to the focused word(s).

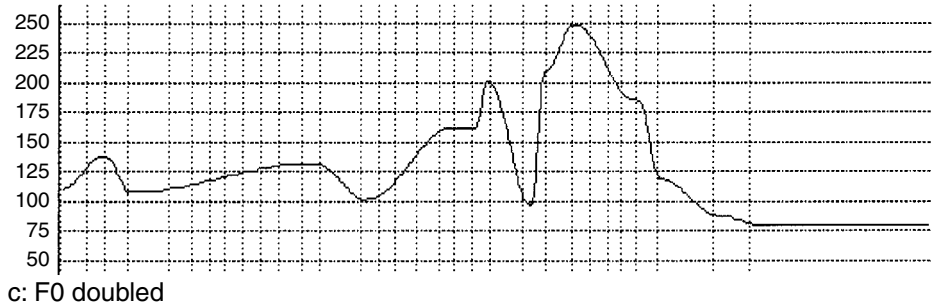
Although the experimental rules were designed to generate a doubling of both F0 maxima and duration in various combinations, there is a slight deviation from this ideal in the actual realisations. This is due to the fact that there are complex rules governing how declination slope and segment durations vary with the length of the utterance, and this interaction affects the values specified in the experiments. However, as it was not the intention in this experiment to test exact F0 and duration values, but rather to test default F0 and duration against rather extreme values of the same parameters, these small deviations from the ideal were not judged to be of consequence for the results.



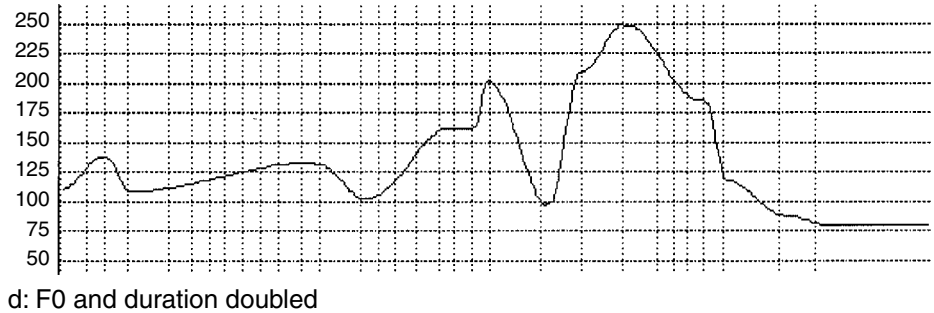
a: default F0 and duration



b: duration doubled



c: F0 doubled



d: F0 and duration doubled

Figure 26.2 Parameter plots for sentence 2 (formant synthesis version)

Results

Children and an adult control group were asked to compare these samples and to evaluate which were the most fun and which were the most natural. Although the study comprised a limited number of subjects (eight children, four for a scaling

task and four for a ranking task as described below, and a control group of four adults), it is clear that the children responded to prosodic differences in the synthesis examples in a fairly consistent manner, preferring large manipulations in F0 and duration when a fun voice is intended. Even for naturalness, the children often preferred larger excursions in F0 than are present in the default versions of the synthesis which is intended largely for adult users. Differences between the children and the adult listeners were according to expectation, where children preferred greater prosodic variation, especially in duration for the fun category. Figure 26.3 shows the mean scores of the children's votes for naturalness and fun in the scaling task, where they were asked to give a score for each of the prosodic types (from 1 to 5, where 5 was best). Figure 26.4 shows the corresponding ratings for the adult control group.

These figures give the combined score for the three test sentences and the two types of synthesis (formant and concatenative). One thing that emerges from this is that the children gave all the different versions an approximately equal fun rating, but considered the versions with prolonged duration as less natural. The adults, on the other hand, show almost identical results to the children as far as naturalness is concerned, but give a lower fun rating too for the versions involving prolonged duration.

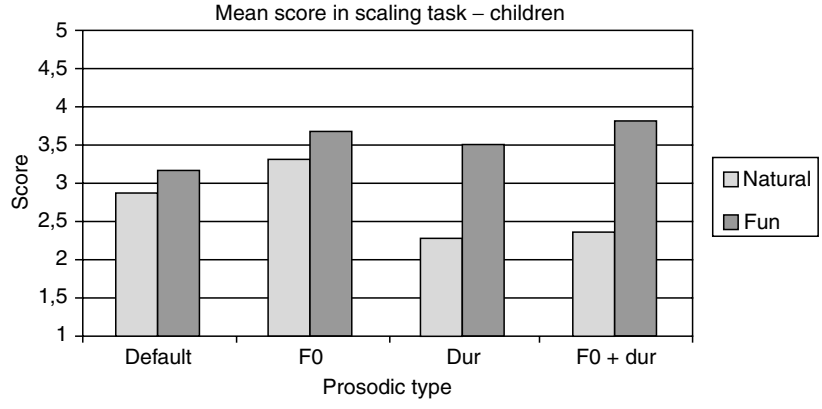


Figure 26.3 Comparison between fun and naturalness scaling – children

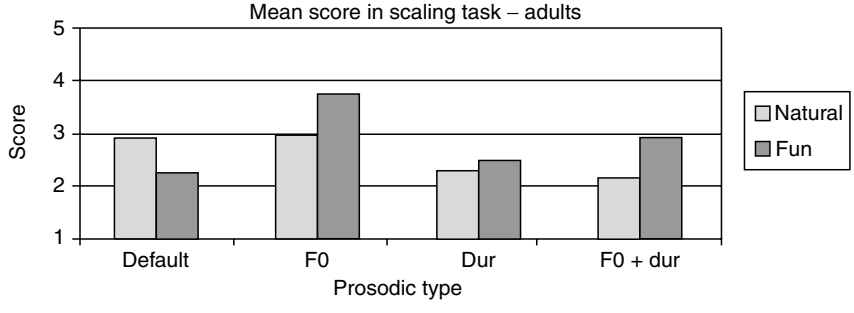


Figure 26.4 Comparison between fun and naturalness scaling – adults

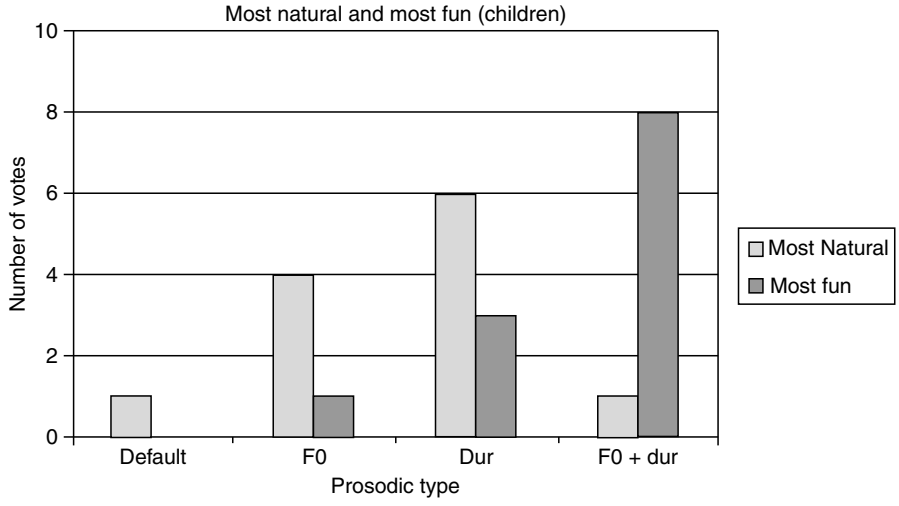


Figure 26.5 Children's ranking test: votes by four children for different realisations of each of three sentences

Figure 26.5 gives a summary for all three sentences of the results in the ranking task, where the children were asked to identify which of the four prosodically different versions was the most fun and which was the most natural. The children that performed this task clearly preferred more 'extreme' prosody, both when it comes to naturalness and especially when the target is a fun voice. The results of the two tasks cannot be compared directly, as they were quite different in nature, but it is interesting to note that the versions involving a combination of exaggerated duration and F0 got the highest score in both tasks. In a web-based follow-up study with 78 girls and 56 boys currently being processed, the preference for more extreme F0 values for a fun voice is very clear.

An additional result from the earlier study was that the children preferred the formant synthesis over the diphone-based synthesis. In the context of this experiment the children may have had a tendency to react to formant synthesis as more appropriate for the animated character portraying an astronaut while the adults may have judged the synthesis quality from a wider perspective. An additional aspect is the concordance between voice and perceived physical size of the animated character. For a large character, such as a lion, children might prefer an extremely low F0 with little variation for a fun voice. The astronaut, however, can be perceived as a small character more suitable to high F0 and larger variation.

Another result of importance is the fact that the children responded positively to changes involving the focused words only. Manipulations involving non-focused words were not tested, as this was judged to produce highly unnatural and less intelligible synthesis. Manipulations in the current synthesis involved raising both peaks (maximum F0 values) of the focal accent 2 words. This is a departure from the default rules (Bruce and Granström, 1993) but is consistent with production and perception data presented in Fant and Kruckenberg (1998). This strategy may be preferred when greater degrees of emphasis are intended.

How to Determine the Parameter Values for Fun Voices

Having established that more extreme values of both the F0 and duration parameters contribute to the perception of the astronaut's voice as being fun, a further natural investigation will be to try to establish the ideal values for these parameters. This experimentation should focus on the interaction between the two parameters. One experimental set-up that suggests itself is to get subjects to determine preferred values by manipulating visual parameters on a computer screen, for instance objects located in the x–y plane.

Further questions that present themselves are: To what extent should the manipulations be restricted to the focused words, especially since this strategy puts greater demands on the synthesis system to correctly indicate focus? Should the stretching of the duration uniformly affect all syllables of the focused word(s), or should it be applied differentially according to segmental and prosodic category? When the F0 contour is complex as is the case in Swedish accent 2 words, how should the different F0 peaks be related (should all peaks be affected equally, or should one be increased more than the other or others by, for instance, a fixed factor? Should the F0 valleys be unaffected (as in our experiment) or should increased 'liveliness' manifest itself in a deepening of the valley between two peaks? Also, in the experiment referred to above, the post-focal lowering of F0 was identical in the various test conditions. However, this is a parameter that can be expected to be perceptually important both for naturalness and for a fun voice. An additional question concerns sentence variability. In longer sentences or in longer texts, extreme variations in F0 or duration may not produce the same results as in single, isolated test sentences. In longer text passages focus relationships are also often more complicated than in isolated sentences.

As has been stressed above, the prosody of high-quality synthesis is intimately linked with the characteristics of voice quality. If one's choice is to use concatenative synthesis, the problem reduces itself to finding a speaker with a voice that best matches the target application. The result may be a voice with a good voice quality, but one which is not ideally suited to the application. High-quality formant synthesis, on the other hand, offers the possibility of manipulating voice source characteristics in a multi-dimensional continuum between child-like and adult male-like speech. But this solution relies on choosing the right setting of a multitude of parameters.

An important question then becomes how to navigate successfully in such a multi-dimensional space. A tool for this purpose, SLIDEBAR, capable of manipulating up to ten different parameters, was developed as part of the VAESS (Voices, Attitudes, and Emotions in Speech Synthesis) project (Bertenstam *et al.*, 1997). One of the aims of this project was to develop (formant) synthesis of several emotions (happy, angry, sad, as well as 'neutral') and of both an adult male and female and a child's voice, for a number of European languages. The tool functions in a Windows environment, where the experimenter uses 'slidebars' to arrive at the appropriate parameter values. Although none of the voices developed as part of the project were meant specifically to be 'fun', the voices that were designed were arrived at by the use of the slidebar interface, manipulating both prosodic and voice quality parameters simultaneously.

In future experimentation, the following are prosodic dimensions that one would like to manipulate simultaneously. These are some of the parameters that were found to be relevant in the modelling of convincing prosody in the context of a man-machine dialogue system (the Waxholm project) for Swedish (Bruce *et al.*, 1995):

- the height of the F0 peak of the syllable with primary stress;
- the height of the F0 peak of the syllable with secondary stress;
- the F0 range in the pre-focal domain;
- the F0 slope following the stressed syllable;
- the durational relations between stressed and unstressed syllables;
- the durational relations between vowels and consonants in the different kinds of syllables;
- the tempo of the pre-focal domain.

In addition to such strictly prosodic parameters, voice quality characteristics, such as breathy and tense voice, are likely to be highly relevant to the creation of a convincing 'fun' voice. Further investigations are also needed to establish how voice quality charateristics interact with the prosodic parameters.

Conclusion

Greater prosodic variation combined with appropriate voice characteristics will be an important consideration when using speech synthesis as part of an educational computer program and when designing spoken dialogue systems for children (Potamianos and Narayanan, 1998). If children are to enjoy using a text-to-speech application in an educational context, more prosodic variation needs to be incorporated in the prosodic rule structure. On the basis of our experiments referred to above and our experiences with the Waxholm and VAESS projects, one hypothesis for a 'fun' voice would be a realisation that uses a wide F0 range in the domain of the focused word, a reduced F0 range in the pre-focal domain, a faster tempo in the pre-focal domain, and a slightly slower tempo in the focal domain.

The interactive dimension of synthesis can also be exploited, making it possible for children to write their own character lines and have the characters speak these lines. To this end, children can be allowed some control over prosodic parameters with a variety of animated characters. Further experiments in which children can create voices to match various animated characters could prove highly useful in designing text-to-speech synthesis systems for children.

Acknowledgements

The research reported here was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. We are grateful for having had the opportunity to expand this research

within the framework of COST 258. We wish to thank Linda Bell and Linn Johansson for collaboration on the earlier paper and David Skoglund for assistance in creating the interactive test environment. We would also like to thank Björn Granström, Mark Huckvale and Jacques Terken for comments on earlier versions of this chapter.

References

Abe, M. (1997). Speaking styles: Statistical analysis and synthesis by a text-to-speech system. In J.P.H. van Santen, R. Sprout, J.P. Olive, and J. Hirschberg (eds), *Progress in Speech Synthesis* (pp. 495–510). Springer-Verlag.

Bertenstam, J., Granström, B., Gustafson, K., Hunnicutt, S., Karlsson, I., Meurlinger, C., Nord, L., and Rosengren, E. (1997). The VAESS communicator: A portable communication aid with new voice types and emotions. *Proceedings Fonetik '97* (*Reports from the Department of Phonetics, Umeå University, 4*), 57–60.

Bredvad-Jensen, A-C. (1995). Prosodic variation in parental speech in Swedish. *Proceedings of ICPhS-95* (pp. 389–399). Stockholm.

Bruce, G. and Granström, B. (1993). Prosodic modelling in Swedish speech synthesis. *Speech Communication, 13*, 63–73.

Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D., and Touati, P. (1995). Towards an enhanced prosodic model adapted to dialogue applications. In P. Dalsgaard *et al*. (eds), *Proceedings of ESCA Workshop on Spoken Dialogue Systems, May-June 1995* (pp. 201–204). Vigsø, Denmark.

Carlson, R., Granström, B., and Nord, L. (1992). Experiments with emotive speech – acted utterances and synthesized replicas. *Proceedings of the International Conference on Spoken Language Processing. ICSLP–92* (vol. 1, pp. 671–674). Banff, Alberta, Canada.

Fant, G. and Kruckenberg, A. (1998). Prominence and accentuation. Acoustical correlates. *Proceedings FONETIK 98* (pp. 142–145). Department of Linguistics, Stockholm University.

House, D., Bell, L., Gustafson, K., and Johansson, L. (1999). Child-directed speech synthesis: Evaluation of prosodic variation for an educational computer program. *Proceedings of Eurospeech 99* (pp. 1843–1846). Budapest.

Kitamura, C. and Burnham, D. (1998). Acoustic and affective qualities of IDS in English. *Proceedings of ICSLP 98* (pp. 441–444). Sydney.

Potamianos, A. and Narayanan, S. (1998). Spoken dialog systems for children. *Proceedings of ICASSP 98* (pp. 197–201). Seattle.

Snow, C.E. and Ferguson, C.A. (eds) (1977). *Talking to Children: Language Input and Acquisition*. Cambridge University Press.

Sundberg, U. (1998). *Mother Tongue – Phonetic Aspects of Infant-Directed Speech. Perilus XXI*. Department of Linguistics, Stockholm University.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

27

Dynamics of the Glottal Source Signal

Implications for Naturalness in Speech Synthesis

Christer Gobl and Ailbhe Ní Chasaide
Centre for Language and Communication Studies, Trinity College, Dublin, Ireland
cegobl@tcd.ie

Introduction

The glottal source signal varies throughout the course of spoken utterances. Furthermore, individuals differ in terms of their basic source characteristics. Glottal source variation serves many linguistic, paralinguistic and extralinguistic functions in spoken communication, but our understanding of the source is relatively primitive compared to other aspects of speech production, e.g., variation in the shaping of the supraglottal tract. In this chapter, we outline and illustrate the main types of glottal source variation that characterise human speech, and discuss the extent to which these are captured or absent in current synthesis systems. As the illustrations presented here are based on an analysis methodology not widely used, this methodology is described briefly in the first section, along with the glottal source parameters which are the basis of the illustrations.

Describing Variation in the Glottal Source Signal

According to the acoustic theory of speech production (Fant, 1960), speech can be described in terms of source and filter. The acoustic source during phonation is generally measured as the volume velocity (airflow) through the glottis. The periodic nature of the vocal fold vibration results in a quasi-periodic waveform, which is typically referred to as the *voice source* or the *glottal source*. This waveform constitutes the input signal to the acoustic filter, the vocal tract. Oscillations are introduced to the output lip volume velocity signal (the oral airflow) at frequencies corresponding to the resonances of the vocal tract. The output waveform is the

convolution of the glottal waveform and the impulse response of the vocal tract filter. The radiated sound pressure is approximately proportional to the differentiated lip volume velocity.

So if the speech signal is the result of a sound source modified by the filtering effect of the vocal tract, one should in principle be able to obtain the source signal through the cancellation of the vocal tract filtering effect. Insofar as the vocal tract transfer function can be approximated by an all-pole model, the task is to find accurate estimates of the formant frequencies and bandwidths. These formant estimates are then used to generate the inverse filter, which can subsequently be used to filter the speech (pressure) signal. If the effect of lip radiation is not cancelled, the resulting signal is the differentiated glottal flow, the time-derivative of the true glottal flow. In our voice source analyses, we have almost exclusively worked with the differentiated glottal flow signal.

Although the vocal tract transfer function can be estimated using fully automatic techniques, we have avoided using these as they are too prone to error, often leading to unreliable estimates. Therefore the vocal tract parameter values are estimated manually using an interactive technique. The analysis is carried out on a pulse-by-pulse basis, i.e. all formant data are re-estimated for every glottal cycle. The user adjusts the formant frequencies and bandwidths, and can visually evaluate the effect of the filtering, both in the time domain and the frequency domain. In this way, the operator can optimise the filter settings and hence the accuracy of the voice source estimate.

Once the inverse filtering has been carried out and an estimate of the source signal has been obtained, a voice source model is matched to the estimated signal. The model we use is the LF model (Fant *et al.*, 1985), which is a four-parameter model of differentiated glottal flow. The model is matched by marking certain timepoints and a single amplitude point in the glottal waveform. The analysis is carried out manually for each individual pulse, and the accuracy of the match can be visually assessed both in the time and the frequency domains. For a more detailed account of the inverse filtering and matching techniques and software used, see Gobl and Ní Chasaide (1999a) and Ní Chasaide *et al.* (1992).

On the basis of the modelled LF waveform, we obtain measures of salient voice source parameters. The parameters that we have mainly worked with are EE, RA, RG and RK (and OQ, derived from RG and RK). EE is the excitation strength, measured as the (absolute) amplitude of the differentiated glottal flow at the maximum discontinuity of the pulse. It is determined by the speed of closure of the vocal folds and the airflow through them. A change in EE results in a corresponding amplitude change in all frequency components of the source with the exception of the very lowest components, particularly the first harmonic. The amplitude of these lowest components is determined more by the pulse shape, and therefore they vary less with changes in EE. The RA measure relates to the amount of residual airflow of the return phase, i.e. during the period after the main excitation, prior to maximum glottal closure. RA is calculated as the return time, TA, relative to the fundamental period, i.e. $RA = TA/T_0$, where TA is a measure that corresponds to the duration of the return phase. The acoustic consequence of this return phase is manifest in the spectral slope, and an increase in RA results in a greater attenuation of the higher frequency components. RG is a measure of the 'glottal

frequency' (Fant, 1979), as determined by the opening branch of the glottal pulse, normalised to the fundamental frequency. RK is a measure of glottal pulse skew, defined by the duration of the closing branch of the glottal pulse relative to the duration of the opening branch. OQ is the open quotient, i.e. the proportion of the pulse for which the glottis is open. The relationship between RK, RG and OQ is the following: $OQ = (1 + RK)/2RG$. Thus, OQ is positively correlated with RK and negatively correlated with RG. It is mainly the low frequency components of the source spectrum that are affected by changes in RK, RG and OQ. The most notable acoustic effect is perhaps the typically close correspondence between OQ and the amplitude of the first harmonic: note however that the degree of correspondence varies depending on the values of RG and RK.

The source analyses can be supplemented by measurements from spectral sections (and/or average spectra) of the speech output. The amplitude levels of the first harmonic and the first four formants may permit inferences on source effects such as spectral tilt. Though useful, this type of measurement must be treated with caution, see discussion in Ní Chasaide and Gobl (1997).

Single Speaker Variation

The term neutral voice is used to denote a voice quality which does not audibly include non-modal types of phonation, such as creakiness, breathiness, etc. The terms neutral and modal, however, are sometimes misunderstood and taken to mean that voice source parameters are more or less constant. This is far from the true picture: for any utterance, spoken with neutral/modal voice there is considerable modulation of voice source parameters. Figure 27.1 illustrates this modulation for the source parameters EE, RG, RK, RA and OQ in the course of the Swedish utterance *Inte i detta århundrade* (Not in this century). Synthesis systems, when they do directly model the voice source signal, do not faithfully reproduce this natural modulation, characteristic of human speech, and one can only assume that this contributes in no small way to the perceived unnaturalness of synthetic speech.

This source modulation appears to be governed by two factors. First, some of the variation seems to be linked to the segmental and suprasegmental patterns of a language: this we might term linguistic variation. Beyond that, speakers use changes in voice quality to communicate their attitude to the interlocutor and to the message, as well as their moods and emotions, i.e. for paralinguistic communication.

Linguistic factors

In considering the first, linguistic, type of variation, it can be useful to differentiate between segment-related variation and that which is part of the suprasegmental expression of utterances. Consonants and vowels may be contrasted on the basis of voice quality, and such contrasts are commonly found in South-East Asian, South African and Native American languages. It is less widely appreciated that in languages where voice quality is not deemed to have a phonologically contrastive function, there are nevertheless many segment-dependent variations in the source.

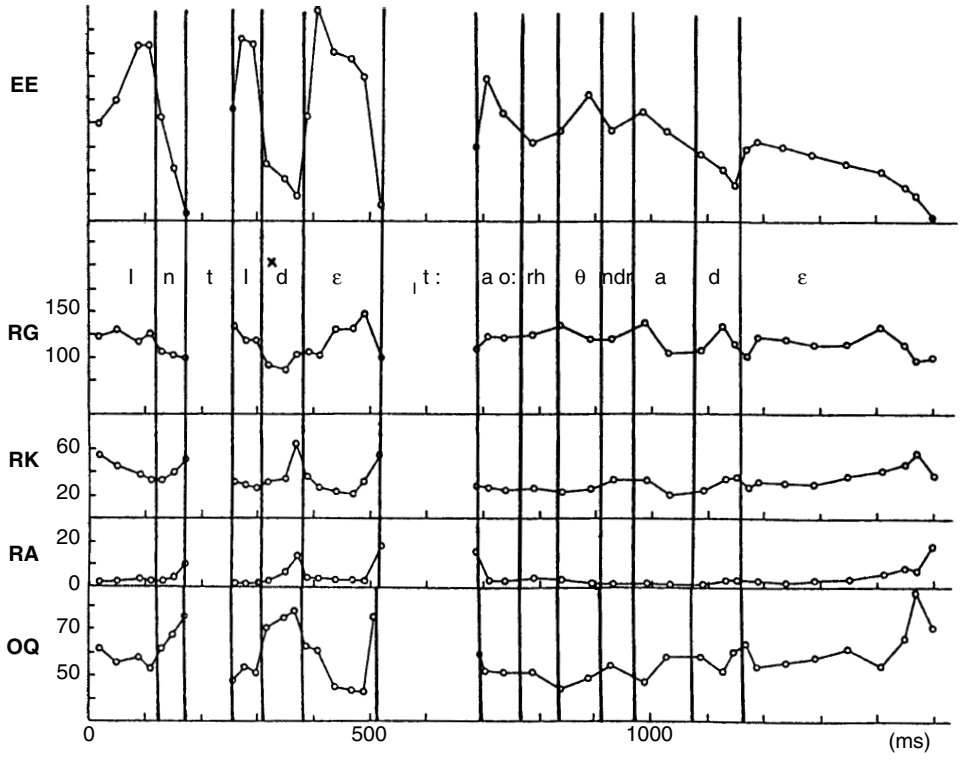


Figure 27.1 Source data for EE, RG, RK, RA and OQ, for the Swedish utterance *Inte i detta århundrade*

Figure 27.2 illustrates the source values for EE, RA and RK during four different voiced consonants / l b m v / and for 100 ms of the preceding vowel in Italian and French (note the consonants of Italian here are geminates). Differences of a finer kind can also be observed for different classes of vowels. For a fuller description and discussion, see Gobl *et al.* (1995) and Ní Chasaide *et al.* (1994).

These segment-related differences probably reflect to a large extent the downstream effects of the aerodynamic conditions that pertain when the vocal tract is occluded in different ways and to varying degrees. Insofar as these differences arise from speech production constraints, they are likely to be universal, intrinsic characteristics of consonants and vowels.

Striking differences in the glottal source parameters may also appear as a function of how consonants and vowels combine. In a cross-language study of vowels preceded and/or followed by stops (voiced or voiceless) striking differences emerged in the voice source parameters of the vowel. Figure 27.3 shows source parameters EE and RA for a number of languages, where they are preceded by / p / and followed by / p(:) b(:) /. The traces have been aligned to oral closure in the post-vocalic stop (= 0 ms). Note the differences between the offsets of the French data and those of the Swedish: these differences are most likely to arise from differences in the timing in the glottal abduction gesture for voiceless stops in the two languages. Compare also the onsets following / p / in the Swedish and German data: these differences may

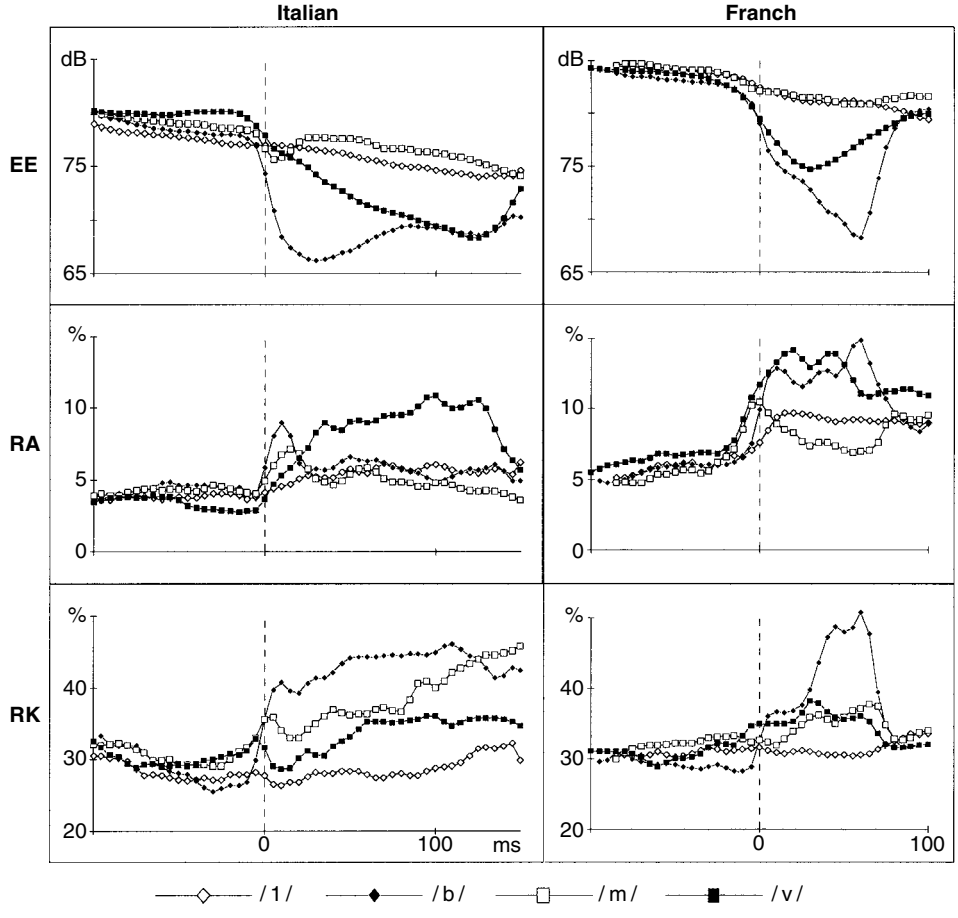


Figure 27.2 Source data for EE, RA and RK during the consonants /l(:) m(:) v(:) b(:) / and for 100 ms of the preceding vowel, for an Italian and a French speaker. Values are aligned to oral closure or onset of constriction for the consonant (= 0 ms)

relate rather to the tension settings in the vocal folds (for a fuller discussion, see Gobl and Ní Chasaide, 1999b). Clearly, the differences here are likely to form part of the language/dialect specific code.

Not all such coarticulatory effects are language dependent. Fricatives (voiceless and voiced) appear to make a large difference to the source characteristics of a preceding vowel, an influence similar to that of the Swedish stops, illustrated above. However, unlike the case of the stops, where the presence and extent of influence appear to be language/dialect dependent, the influence of the fricatives appears to be the same across these same languages. The most likely explanation for the fact that fricatives are different from stops lies in the production constraints that pertain to the former. Early glottal abduction may be a universal requirement if the dual requirements of devoicing and supraglottal frication are to be adequately met (see also discussion, Gobl and Ní Chasaide, 1999b).

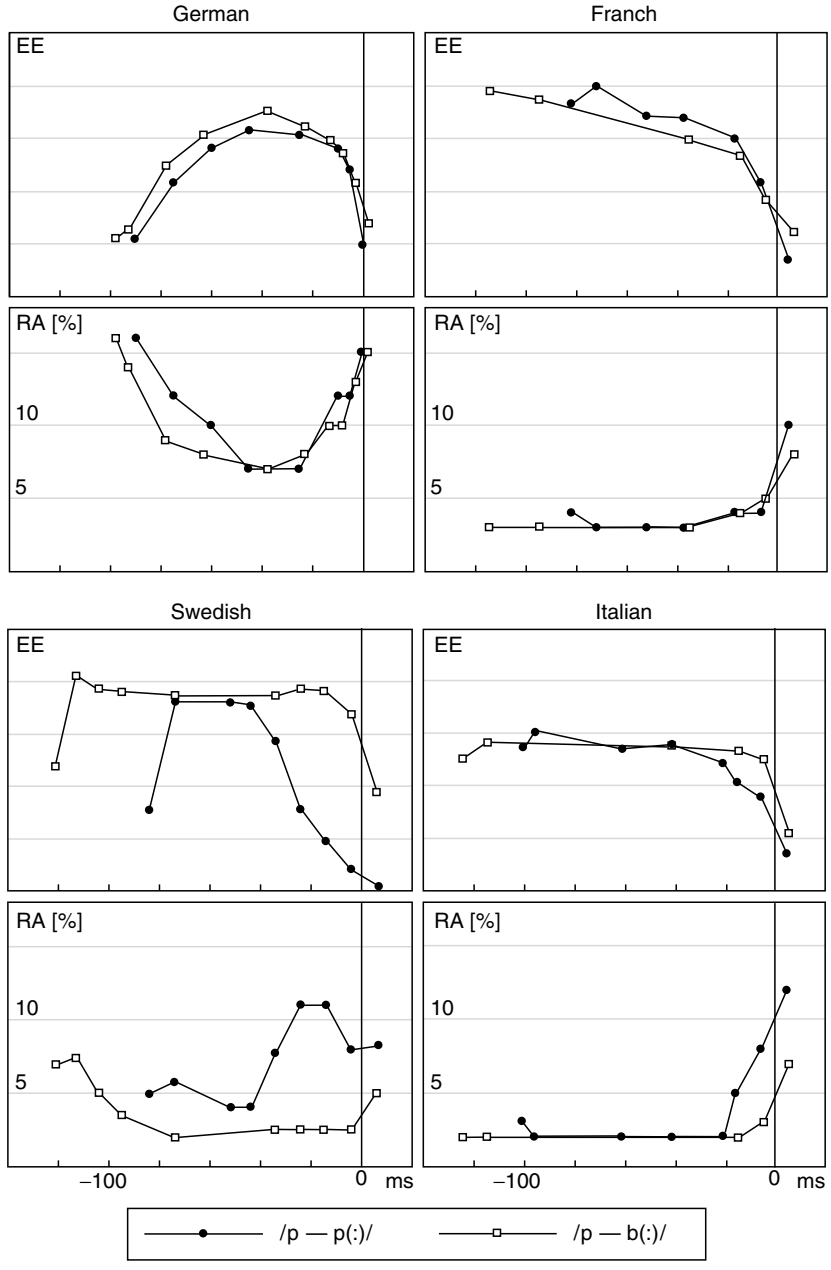


Figure 27.3 Vowel source data for EE and RA, superimposed for the /p—p(:)/ and /p—b(:)/ contexts, for German, French, Swedish and Italian speakers. Traces are aligned to oral closure (= 0 ms)

For the purpose of this discussion, we would simply want to point out that there are both universal and language specific coarticulatory phenomena of this kind. These segmentally determined effects are generally not modelled in formant based

synthesis. On the other hand, in concatenative synthesis these effects should in principle be incorporated. However, insofar as the magnitude of the effects may depend on the position, accent or stress (see below) these may not be fully captured by such systems.

Much of the source variation that can be observed relates to the suprasegmental level. Despite the extensive research on intonation, stress and tone, the work has concentrated almost entirely on F0 (and sometimes amplitude) variation. However, other source parameters are also implicated. Over the course of a single utterance, as in Figure 27.1, one can observe modulation that is very reminiscent of F0 modulation. Note, for example, a declination in EE (excitation strength). The termination of utterances is typically marked by changes in glottal pulse shape that indicate a gradual increase in breathiness (a rising RA, RK and OQ). Onsets of utterances tend to exhibit similar tendencies, but to a lesser extent. A shift into creaky voice may also be used as a phrase boundary marker in Swedish (Fant and Kruckenberg, 1989). The same voice quality may fulfil the same function in the RP accent of English: Laver (1980) points out that such a voice quality with a low falling intonation signals that a speaker's contribution is completed.

Not surprisingly, the location of stressed syllables in an utterance has a large influence on characteristics of the glottal pulse shape, not only in the stressed syllable itself but also in the utterance as a whole. Gobl (1988) describes the variation in source characteristics that occur when a word is in focal position in an utterance, as compared to prefocal or postfocal. The most striking effect appears to be that the dynamic contrast between the vowel nucleus and syllable margins is enhanced in the focally stressed syllable: the stressed vowel tends to exhibit a stronger excitation, less glottal skew and less dynamic leakage, whereas the opposite pertains to the syllable margin consonants. Pierrehumbert (1989) also illustrates source differences between high and low tones (in pitch accents) and points out that an adequate phonetic realisation of intonation in synthetic speech will require a better understanding of the interaction of F0 and other voice source variables.

In tone languages, the phonetic literature suggests that many tonal contrasts involve complex source variations, which include pitch and voice quality. This comes out clearly in the discussions as to whether certain distinctions should be treated phonologically as tonal or voice quality contrasts. For a discussion, see Ní Chasaide and Gobl (1997). Clearly, both are implicated, and therefore an implementation in synthesis which ignores one dimension would be incomplete.

Some source variation is intrinsically linked in any case to variation in F0, see Fant (1997). To the extent that the glottal source features covary with F0, it should be an easy matter to incorporate these in synthesis. However, whatever the general tendencies to covariation, voice quality can be (for most of a speaker's pitch range) independently controlled, and this is a possibility which is exploited in language.

Paralinguistic factors

Beyond glottal source variation that is an integral part of the linguistic message, speakers exploit voice quality changes (along with F0, timing and other features) as a way of communicating their attitudes, their state of mind, their moods and

emotions. Understanding and modelling this type of variation are likely to be of considerable importance if synthetic speech is to come near to having expressive nuances of the human performance. This aspect of source variation is not dealt with here as it is the subject matter of a separate chapter (Ní Chasaide and Gobl, Chapter 25, this volume).

Cross-Speaker Variation

Synthesis systems also need to incorporate different voices, and obviously, glottal source characteristics are crucial here. Most synthesis systems offer at least the possibility of selecting between a male, a female and a child's voice. The latter two do not present a particular problem in concatenative synthesis: the method essentially captures the voice quality of the recorded subject. In the case of formant synthesis it is probably fair to say that the female and child's voices fall short of the standard attained for the male voice. This partly reflects the fact that the male voice has been more extensively studied and is easier to analyse. Another reason why male voices sound better in formant-based synthesis may be that cruder source modelling is likely to be less detrimental in the case of the male voice. The male voice typically conforms better to the common (oversimplified) description of the voice source as having a constant spectral slope of -12 dB/octave, and thus the traditional modelling of the source as a low-pass filtered pulse train is more suitable for the male voice. Furthermore, source-filter interaction may play a more important role in the female and child's voice, and some of these interaction effects may be difficult to simulate in the typical formant synthesis configuration.

Physiologically determined differences between the male and female vocal apparatus will, of course, affect both vocal tract and source parameters. Vocal tract differences are relatively well understood, but there is relatively little data on the differences between male and female source characteristics, apart from the well-known F0 differences (females having F0 values approximately one octave higher). Nevertheless, experimental results to date suggest that the main differences in the source concern characteristics for females that point towards an overall breathier voice quality.

RA is normally higher for female voices. Not only is the return time longer in relative terms (relative to the fundamental period) but generally also in absolute terms. As a consequence, the spectral slope is typically steeper, with weaker higher harmonics. Most studies also report a longer open quotient, which would suggest a stronger first harmonic, something which would further emphasise the lower frequency components of the source relative to the higher ones (see, for instance, Price, 1989 and Holmberg *et al.*, 1988). Some studies also suggest a more symmetrical glottal pulse (higher RK) and a slightly lower RG (relative glottal frequency). However, results for these latter two parameters are less consistent, which could partly be due to the fact that it is often difficult to measure these accurately. It has also often been suggested that female voices have higher levels of aspiration noise, although there is little quantitative data on this. Note, however, the comments in Klatt (1987) and Klatt and Klatt (1990), who report a greater tendency for noise excitation of the third formant region in females compared to males.

It should be pointed out here that even within the basic formant synthesis configuration, it is possible to generate very high quality copy synthesis of the child and female voices (for example, Klatt and Klatt, 1990). It is more difficult to derive these latter voices from the male voice using transformation rules, as the differences are complex and involve both source and filter features. In the audio example included, a synthesised utterance of a male speaker is transformed in a stepwise manner into a female sounding voice. The synthesis presented in this example was carried out by the first author, originally as part of work on the female voice reported in Fant *et al.* (1987). The source manipulations effected in this illustration were based on personal experience in analysing male and female voices, and reflect the type of gender (and age) related source differences encountered in the course of studies such as Gobl (1988), Gobl and Ní Chasaide (1988) and Gobl and Karlsson (1991).

The reader should note that this example is intended as an illustration of what can be achieved with very simple global manipulations, and should not be taken as a formula for male to female voice transformation. The transformation here is a cumulative process and each step is presented separately and repeated twice. The source and filter parameters that were changed are listed below and the order is as follows:

- copy synthesis of Swedish utterance *ja adjö* (yes goodbye), male voice
- doubling of F0
- reduction in the number of formants
- increase in formant bandwidths
- 15% increase in F1, F2, F3
- OQ increase
- RA increase
- original copy synthesis

There are of course other relevant parameters, not included here, that one could have manipulated, e.g., aspiration noise. Dynamic parameter transformations and features such as period-to-period variation are also likely to be important.

Beyond the gross categorical differences of male/female/child, there are many small, subtle differences in the glottal source which enable us to differentiate between two similar speakers, for example, two men of similar physique and same accent. These source differences are likely to involve differences in the intrinsic baseline voice quality of the particular speaker. Very little research has focused directly on this issue, but studies where groups of otherwise similar informants were used (e.g., Gobl and Ní Chasaide, 1988; Holmberg *et al.*, 1988; Price, 1989; Klatt and Klatt, 1990) suggest that the types of variation encountered is similar to the variation that a single speaker may use for paralinguistic signalling, and which is discussed in Ní Chasaide and Gobl (see Chapter 25, this volume).

Synthesis systems of the future will hopefully allow for a much richer choice of voices. Ideally one would envisage systems where the prospective user might be able to tailor the voice to meet individual requirements. For many of the currently common applications of speech synthesis systems, these subtler differences might appear irrelevant. Yet one does not have to look far to see how important this facility would be for certain groups of users, and undoubtedly, enhancements of this type

would greatly extend the acceptability and range of applications of synthesis systems. For example, one important current application concerns aids for the vocally handicapped. In classrooms where vocally handicapped children communicate through synthesised speech, it is a very real drawback that there is normally only a single child voice available. In the case of adult users who have lost their voice, dissatisfaction with the voice on offer frequently leads to a refusal to use these devices.

The idea of tailored, personalised voices is not technically impossible, but involves different tasks, depending on the synthesis system employed. In principle, concatenative systems can achieve this by recording numerous corpora, although this might not be the most attractive solution. Formant-based synthesis, on the other hand, offers direct control of voice source parameters, but do we know enough about how these parameters might be controlled?

Conclusion

All the functions of glottal source variation discussed here are important in synthesis, but the relative importance depends to some extent on the domain of application. The task of incorporating them in synthesis presents different kinds of problems depending on the method used. The basic methodology used in concatenative synthesis is such that it captures certain types of source variations quite well, e.g., basic voice types (male/female/child) and intersegmental coarticulatory effects. Other types of source variation, e.g., suprasegmental, paralinguistic and subtle, fine-grained cross-speaker differences are not intrinsically captured, and finding a means of incorporating these will present a considerable challenge.

In formant synthesis, as one has direct control over the glottal source, it should in principle be possible to incorporate all types of source variation discussed here. At the level of analysis there are many source parameters one can describe, and the task of effectively controlling these in synthesis might appear daunting. One possible way to proceed in the first instance would be to harness the considerable covariation that tends to occur among parameters such as EE, RA, RK and RG (see, for example, Gobl, 1988). On the basis of such covariation, Fant (1997) has suggested global pulse shape parameters, which might provide a simpler way of controlling the source. It must be said, however, that the difficulty of incorporating source variation in formant-based synthesis concerns not only the implementation but also our basic knowledge as to what the rules are for the human speaker.

Acknowledgements

The authors are grateful to COST 258 for the forum it has provided to discuss this research and its implications for more natural synthetic speech.

References

Fant, G. (1960). *The Acoustic Theory of Speech Production*. Mouton (2nd edition 1970).

Fant, G. (1979). Vocal source analysis – a progress report. *STL-QPSR* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), *3–4*, 31–54.

Fant, G. (1997). The voice source in connected speech. *Speech Communication*, *22*, 125–139.
Fant, G., Gobl, C., Karlsson, I., and Lin, Q. (1987). The female voice – experiments and overview. *Journal of the Acoustical Society of America*, *82* S90(A).
Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), *2*, 1–83.
Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), *4*, 1–13.
Gobl, C. (1988). Voice source dynamics in connected speech. *STL-QPSR* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), *1*, 123–159.
Gobl, C. and Karlsson, I. (1991). Male and female voice source dynamics. In J. Gauffin and B. Hammarberg (eds), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms* (pp. 121–128). Singular Publishing Group.
Gobl, C. and Ní Chasaide, A. (1988). The effects of adjacent voiced/voiceless consonants on the vowel voice source: a cross language study. *STL-QPSR* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), *2–3*, 23–59.
Gobl, C. and Ní Chasaide, A. (1999a). Techniques for analysing the voice source. In W.J. Hardcastle and N. Hewlett (eds) *Coarticulation: Theory, Data and Techniques* (pp. 300–321). Cambridge University Press.
Gobl, C. and Ní Chasaide, A. (1999b). Voice source variation in the vowel as a function of consonantal context. In W.J. Hardcastle and N. Hewlett (eds), *Coarticulation: Theory, Data and Techniques* (pp. 122–143). Cambridge University Press.
Gobl, C., Ní Chasaide, A., and Monahan, P. (1995). Intrinsic voice source characteristics of selected consonants. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, 1, 74–77.
Holmberg, E.B., Hillman, R.E., and Perkell, J.S. (1988). Glottal air flow and pressure measurements for loudness variation by male and female speakers. *Journal of the Acoustical Society of America*, *84*, 511–529.
Klatt, D.H. (1987). Acoustic correlates of breathiness: first harmonic amplitude, turbulence noise and tracheal coupling. *Journal of the Acoustical Society of America*, *82*, S91(A).
Klatt, D.H., and Klatt, L.C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, *87*, 820–857.
Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press.
Ní Chasaide, A. and Gobl, C. (1993). Contextual variation of the vowel voice source as a function of adjacent consonants. *Language and Speech*, *36*, 303–330.
Ní Chasaide, A. and Gobl, C. (1997). Voice source variation. In W.J. Hardcastle and J. Laver (eds), *The Handbook of Phonetic Sciences* (pp. 427–461). Blackwell.
Ní Chasaide, A., Gobl, C., and Monahan, P. (1992). A technique for analysing voice quality in pathological and normal speech. *Journal of Clinical Speech and Language Studies*, *2*, 1–16.
Ní Chasaide, A., Gobl, C., and Monahan, P. (1994). Dynamic variation of the voice source: intrinsic characteristics of selected vowels and consonants. *Proceedings of the Speech Maps Workshop, Esprit/Basic Research Action no. 6975*, Vol. 2. Grenoble, Institut de la Communication Parlée.
Pierrehumbert, J.B. (1989). A preliminary study of the consequences of intonation for the voice source. *STL-QPSR* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), *4*, 23–36.
Price, P.J. (1989). Male and female voice source characteristics: Inverse filtering results. *Speech Communication*, *8*, 261–277.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

28

A Nonlinear Rhythmic Component in Various Styles of Speech

Brigitte Zellner Keller and Eric Keller
Laboratoire d'analyse informatique de la parole (LAIP)
Université de Lausanne, CH-1015 Lausanne, Switzerland
Brigitte.ZellnerKeller@imm.unil.ch, Eric.Keller@imm.unil.ch

Introduction

A key objective for our laboratory is the construction of a dynamic model of the temporal organisation of speech and the testing of this model with a speech synthesiser. Our hypothesis is that the better we understand how speech is organised in the time dimension, the more fluent and natural synthetic speech will sound (Zellner Keller, 1998; Zellner Keller and Keller, in press). In view of this, our prosodic model is based on the prediction of temporal structures from which we derive durations and on which we base intonational structures.

It will be shown here that ideas and data on the temporal structure of speech fit quite well into a complex nonlinear dynamic model (Zellner Keller and Keller, in press). Nonlinear dynamic models are appropriate to the temporal organisation of speech, since this is a domain characterised not only by serial effects contributing to the dynamics of speech, but also by small events that may produce nonlinearly disproportionate effects (e.g. a silent pause within a syllable that produces a strong disruption in the speech flow). Nonlinear dynamic modelling constitutes a novel approach in this domain, since *serial interactions* are not systematically incorporated into contemporary predictive models of timing for speech synthesis, and *nonlinear effects* are not generally taken into account by the linear predictive models in current use.

After a discussion of the underlying assumptions of models currently used for the prediction of speech timing in speech synthesis, it will be shown how our 'BioPsychoSocial' model of speech timing fits into a view of speech timing as a dynamic nonlinear system. On this basis, a new rhythmic component will be proposed and discussed with the aim of modelling various speech styles.

Prediction of Timing in Current Speech Synthesisers

While linguistic approaches are rare in recent predictive models of speech timing, quantitative approaches have undoubtedly been favoured by recent developments in computational and statistical methods. Quantitative approaches are generally based on databases of empirical data (i.e. speech unit durations), organised in such a manner that statistical analysis can be performed. Typically, the goal is to find an optimal statistical method for computing durations of speech units. Four types of statistical methods have been widely investigated in this area. The first two methods allow nonlinear transformations of the relations between input and output, and the second two are purely linear modelling techniques.

Artificial neural networks (*ANN*), as proposed for example by Campbell (1992) or Riedi (1998) are implemented in various European speech synthesis systems (SSS) (cf. Monaghan, this volume). In this approach, durations are computed on the basis of various input parameters to the network, such as the number of phonemes, the position in the tone group, the type of foot, the position of word or phrase stress, etc. ANNs find their optimal output (i.e. the duration of a given speech unit) by means of a number of summation and threshold functions.

Classification and regression trees (*CARTs*) as proposed by Riley (1992) for durational modelling are binary decision trees derived from data by using a recursive partitioning algorithm. This hierarchical arrangement progresses from one decision (or branch) to another, until the last node is reached. The algorithm computes the segmental durations according to a series of contextual factors (manner of articulation, adjacent segments, stress, etc.). A common weakness of CARTs in this application is the relative sparsity of data for final output nodes, due to the large number of phonemes in many languages, their unequal frequency of occurrence in most data sets, and the excessive number of relevant interactions between adjoining sounds. This is referred to as the 'sparsity problem' (van Santen and Shih, 2000).

The *sum-of-products model*, proposed by van Santen (1992) and Klabbers (2000) is a type of additive decomposition where phonemes that are affected similarly by a set of factors are grouped together. For each subclass of segments, a separate sum-of-products model is computed according to phonological knowledge. In other words, this kind of model gives the duration for a given phoneme-context combination.

A *hierarchical arrangement of the General Linear Model*, proposed by Keller and Zellner (1995) attemps to predict a dependent variable (i.e. the duration of a sound class) in terms of a hierarchical structure of independent variables involving segmental, syllabic and phrasal levels. In an initial phase, the model incorporates segmental information concerning type of phoneme and proximal phonemic context. Subsequently, the model adds information on whether the syllable occurs in a function or a content word, on whether the syllable contains a schwa and on where in the word the syllable is located. In the final phase, the model adds information on phrase-level parameters such as phrase-final lengthening. As in the sum-of-products model, the sparsity problem was countered by a systematic grouping of phonemes (see also Zellner (1998) and Siebenhaar *et al.* (Chapter 16. this volume) for details of the grouping procedure).

Apart from theoretical arguments for choosing one statistical method over another, it is noticeable that the performances of all these models are reasonably good since correlation coefficients between predicted and observed durations are high (0.85–0.9) and the RMSE (Root Mean Square Error) is around 23 ms (Klabbers, 2000). The level of precision in timing prediction is thus statistically high. However, the perceived timing in SSS built with such models is still unnatural in many places. In this chapter, it is suggested that part of this lack of rhythmic naturalness derives from a number of questionable assumptions made in statistical predictive models of speech timing.

Theoretical Implications of Current Statistical Approaches

A *linear relation* between variables is often assumed for the prediction of speech timing. This means that small causes are supposed to produce relatively small effects, while large causes are supposed to produce proportionally larger effects. However, experience with predictive systems shows that small errors in the prediction of durations may at times produce serious perceptual errors, while the same degree of predictive error produces only a small aberration in a different context. Similarly, a short pause may produce a dramatic effect if it occurs in a location where pauses are never found in human speech, but the same pause duration is totally acceptable if it occurs in places where pauses are common (e.g. before function words). Nonlinearity in temporal structure is thus a well-known and well-documented empirical fact, and this property must be modelled properly.

The *underestimation of variability* is also a common issue. Knowledge of the initial conditions of events (e.g. conditions affecting the duration of a phone) is generally assumed to render the future instances of the same event predictable (i.e. the duration of the same phone in similar conditions is assumed to be about the same). However, it is a well-documented fact that complex human gestures such as speech gestures can never be repeated in exactly the same manner, even under laboratory conditions. A major reason for this uncertainty derives from numerous unknown and variable factors affecting the manner in which a speaker produces an utterance (for example, the speaker's pre-existing muscular and emotional state, his living and moving environment, etc.). Many unexplained errors in the prediction of speech timing may well reflect our ignorance of complex interactions between the complete set of parameters affecting the event. An appropriate quantitative approach should explore ways of modelling this kind of uncertainty.

Interactions are the next source of modelling difficulty. Most statistical approaches model only 'simple' interactions (e.g. the durational effect of a prosodic boundary is modified by fast or slow speech rate). What about complex, multiple interactions? For example, speaking under stress may well affect physiological, psychological and social parameters which in turn act on durations in a complex fashion. Similarly, close inspection of some of the thousands of interactions found in our own statistical prediction model has revealed some very strong interactive effects between specific sound classes and specific combinations of predictor values (e.g. place in the word and in the phrase). Because of the 'sparsity problem', reliable and detailed information about these interactions is difficult to come by, and modelling

such complex interactions is difficult. Nevertheless, their potential contribution to the current deficiencies in temporal modelling should not be ignored.

Another assumption concernes the *stability* of the system. It is generally assumed that event structures are stable. However, speech rate is not stable over time, and there is no evidence that relations between all variables remain stable as speech rate changes. It is in fact more likely that various compensation effects occur as speech rate changes. Detailed information on this source of variation is not currently available.

A final assumption underlying many current timing models is that of *causal relation*, in that timing events are often explained in terms of a limited number of causal relations. For example, the duration of the phone *x* is supposed to be *caused* by a number of factors such as position in the syllable, position in the prosodic group, type of segments, etc. However, it is well known in statistics that variables may be related to each other without a causal link because the true cause is to be found elsewhere, in a third variable or even in a set of several other factors. Although the net predictive result may be the same, the bias of a supposed causal relation between statistical elements may reduce the chances of explaining speech timing in meaningful scientific terms. This should be kept in mind in the search for further explanatory parameters in speech timing.

In summary, it seems reasonable to question the common assumption that speech timing is a static homogeneous system. Since factors and interactions of factors are likely to change over time, and since speech timing phenomena show important nonlinear components, it is imperative to begin investigating the dynamics and the 'regulating mechanisms' of the speech timing system in terms of a nonlinear dynamic model. For example, the mechanism could be described in terms of a set of constraints or attractors, as will be shown in the following section.

New Directions: The BioPsychoSocial Speech Timing Model

The BioPsychoSocial Speech Timing Model (Zellner, 1996, 1998) is based on the assumptions that speech timing is a complex multidimensional system involving nonlinearities, complex interactions and dynamic change (changes in the system over time). The aim of this model is to make explicit the numerous factors which contribute to a given state of the system (e.g. a particular rhythm for a particular style of speech).

The BioPsychoSocial Speech Timing Model is based on three levels of constraints that govern speech activity in the time domain (Zellner, 1996, 1998; Zellner-Keller and Keller, forthcoming):

1. Bio-psychological: e.g. respiration, neuro-muscular commands, psycho-rhythmic tendencies.
2. Social: e.g. linguistic and socio-linguistic constraints.
3. Pragmatic: e.g. type of situation, feelings, type of cognitive tasks.

These three sets of constraints and underlying processes have different temporal effects. The bio-psychological level is the 'base level' on which the two others will

superimpose their own constraints. The time domain resulting from these constraints represents the sphere within which speech timing occurs. According to the speaker's state (e.g. when speaking under psychological stress), each level may influence the others in the time domain (e.g. if the base level is reduced because of stress, this reduction in the time domain will project onto the other levels, which in turn will reduce the temporal range of durations).

During speech, this three-tiered set of constraints must satisfy both serial and parallel constraints by means of a multi-articulator system acting in both serial and parallel fashions (glottal, velar, lingual and labial components). Speech gestures produced by this system must be coordinated and concatenated in such a manner that they merge in the temporal dimension to form a stream of identifiable acoustic segments. Although many serial dependencies are documented in the phonetic literature, serial constraints between successive segments have not been extensively investigated for synthesis-oriented modelling of speech timing. In the following section we propose some gains in naturalness that can be obtained by modelling such constraints.

The Management of Temporal Constraints: The Serial Constraint

One type of limited serial dependency, which has often been incorporated into the prediction of speech timing for SSS, is the effect of the *identity* of the preceding and following sounds. This reflects well-known phonetic interactions between adjoining sounds, such as the fact that, in many languages, vowels preceding voiced consonants tend to be somewhat longer than similar vowels preceding unvoiced consonants (e.g. 'bead' vs. 'beat'). Other predictive parameters can also be reinterpreted as partial serial dependencies: e.g. the lengthening of syllable duration due to proximity to the end of a phrase or a sentence.

There is some suggestion in the phonetic literature that the serial dimension may be of interest for timing (Gay, 1981; Miller, 1981; Port *et al.*, 1995; Zellner-Keller and Keller, forthcoming), although it is clearly one of the statistically less important contributors to syllable timing, and has thus have been neglected in timing models. For example, a simple serial constraint described in the literature is the syllabic alternation pattern for French (Duez and Nishinuma, 1985; Nishinuma and Duez, 1988). This pattern suggests that such a serial dependency might produce negative correlations ('anticorrelations') between rhythmic units. It also suggests that serial dependencies of the type $X_{k+1}|X_1 \ldots X_k$ could be investigated using autocorrelational techniques, a well-established method to explore these issues in temporal series (Williams, 1997).

Two theoretically interesting possibilities can be investigated using such techniques. Serial dependency in speech timing can be seen as occurring either on the linguistic or on the absolute time line. Posed in terms of linguistic time units (e.g. syllables), the research question is familiar, since it investigates the rhythmic relationship between syllables that are either adjacent or that are separated by two or more syllables. This question can also be formulated in terms of absolute time, by addressing the rhythmic relations between elements at various distances from each other in absolute time. Posed in this fashion, the focus of the question is directed more at the cognitive or motor processing dimension, since it raises issues of neural motor control and gestural interdependencies within a sequence of articulations.

Modelling a Rhythmic Serial Component

This question was examined in some detail in a previous study (Keller *et al.*, 2000). In short, a significant temporal serial dependency (an anticorrelation) was identified for French, and to a lesser extent for English. That study documents a weak, but statistically significant, serial position effect for both languages in that we identified a durational anticorrelation component that manifested itself reliably within 500 ms, or at a distance of one or two syllables (Figures 28.1–28.5). Also, there is some suggestion of further anticorrelational behaviour at larger syllable lags. It may be that these speech timing events are subject to a time window roughly 500 ms in duration. This interval may relate to various delays in neurophysiological and/or articulatory functioning. It may even reflect a general human rhythmic tendency (Port *et al.*, 1995).

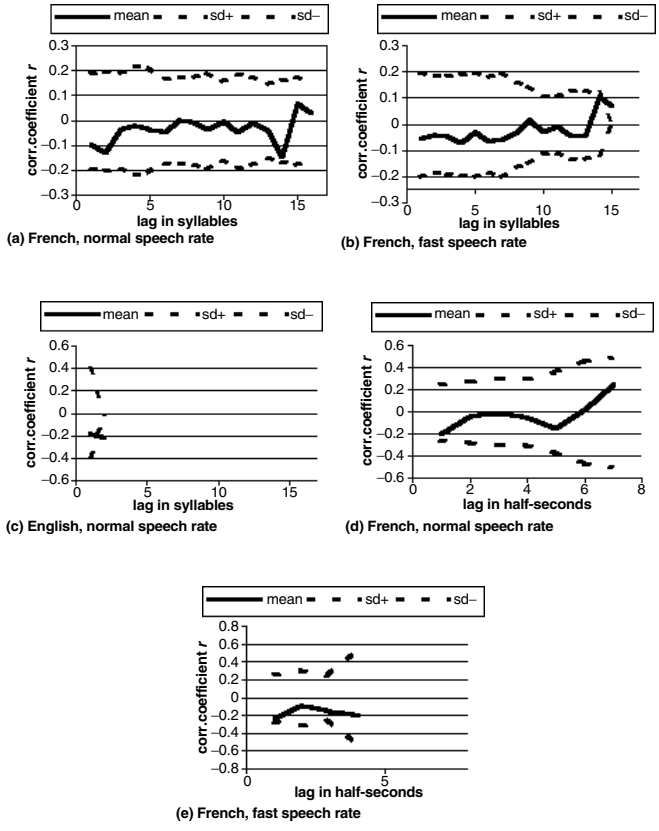


Figure 28.1–28.5 Autocorrelation results for various syllable and half-second lags. Figures 28.1 to 28.3 show the results for the analysis of the linguistic time line, and Figures 28.4 and 28.5 show results for the analysis of the absolute time line. Autocorrelations were calculated between syllabic durations seperated by various lags, and lags were calculated either in terms of syllables or in terms of half-seconds. In all cases and for both languages, negative autocorrelations were found at low lags (lag 1 and lag 2). Results calculated in real time (half-seconds) were particularly compelling

This anticorrelational effect was applied to synthetic speech and implemented as a 'smoothing' parameter in our speech synthesis system (available at www.unil.ch/ imm/docs/LAIP/LAIPTTS.html). As judged informally by researchers in our laboratory, strong anticorrelational values lend the speech output an elastic 'swingy' effect, while weak values produce an output that sounds more controlled and more regimented. The reading of a news report appeared to be enhanced by the addition of an anticorrelational effect, while a train announcement with strong anticorrelational values sounded inappropriately swingy. The first setting may thus be appropriate for a pleasant reading of continuous text, and the latter may be more appropriate for stylised forms of speech such as train announcements.

Conclusion

As has been stated frequently, speech rhythm is a very complex phenomenon that involves an extensive set of predictive parameters. Many of these parameters are still not adequately represented in current timing models. Since speech timing is a complex multidimensional system involving nonlinearities, complex interactions and dynamic changes, it is suggested here that a specific serial component in speech timing should be incorporated into speech timing models. A significant anticorrelational parameter was identified in a previous study, and was incorporated into our speech synthesis system where it appears to 'smooth' speech timing in ways that seem typical of human reading performance. This effect may well be a useful control parameter for synthetic speech.

Acknowledgements

Grateful acknowledgement is made to the Office Fédéral de l'Education (Berne, Switzerland) for supporting this research through its funding in association with Swiss participation in COST 258, and to the Canton de Vaud and the University of Lausanne for funding research leaves for the two authors, hosted in Spring 2000 at the University of York (UK).

References

Campbell, W.N. (1992). Syllable-based segmental duration. In G. Bailly *et al.* (eds), *Talking Machines. Theories, Models, and Designs* (pp. 211–224). Elsevier.

Duez, D. and Nishinuma, Y. (1985). Le rythme en français. *Travaux de l'Institut de Phonétique d'Aix*, *10*, 151–169.

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica*, *38*, 148–158.

Keller, E. and Zellner, B. (1995). A statistical timing model for French. *13th International Congress of the Phonetic Sciences*, *3*, 302–305. Stockholm.

Keller, E. and Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics*, *17*, 53–75. University of York. (Available from http://www.unil.ch/imm/docs/LAIP/Zellnerdoc.html)

Keller E. Zellner Keller, B., and Local, J. (2000). A serial prediction component for speech timing. In Sendlmeir, W. (ed). *Speech and Signals. Aspects of Speech Synthesis and Automatic Speech Recognition*. (pp. 40–49). Forum Phoneticum, 69. Frankfurt am Main: Hector.

Klabbers, E. (2000). *Segmental and Prosodic Improvements to Speech Generation*. PhD thesis, Eindhoven University of Technology (TUE).

Miller, J.L. (1981). Some effects of speaking rate on phonetic perception. *Phonetica*, *38*, 159–180.

Nishinuma, Y. and Duez, D. (1988). Étude perceptive de l'organisation temporelle de l'énoncé en français. *Travaux de l'Institut de Phonétique d'Aix*, *11*, 181–201.

Port, R., Cummins, F., and Gasser, M. (1995). A dynamic approach to rhythm in language: Toward a temporal phonology. In B. Luka and B. Need (eds), *Proceedings of the Chicago Linguistics Society*, *1996* (pp. 375–397). Department of Linguistics, University of Chicago.

Riedi, M. (1998). *Controlling Segmental Duration in Speech Synthesis Systems*. PhD thesis. ETH. Zürich.

Riley, M. (1992). Tree-based modelling of segmental durations. In G. Bailly *et al*. (eds), *Talking Machines: Theories, Models, and Designs* (pp. 265–273). Elsevier.

van Santen, J.P.H. (1992). Deriving text-to-speech durations form natural speech. In G. Bailly *et al*. (eds), *Talking Machines: Theories, Models and Designs* (pp. 265–275). Elsevier.

van Santen, J.P.H. and Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *JASA*, *107*, 1012–1026.

Williams, G.P. (1997). *Chaos Theory Tamed*. Taylor and Francis.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée*, *1*, 7–23.

Zellner, B. (1998). Caractérisation et prédiction du débit de parole en français: Une étude de cas. Unpublished PhD thesis, Faculté des Lettres, Université de Lausanne. (Available from http://www.unil.ch/imm/docs/LAIP/Zellnerdoc.html).

Zellner Keller, B. and Keller, E. (in press). The chaotic nature of speech rhythm: Hints for fluency in the language acquisition process. In Ph. Delcloque and V.M. Holland (eds), *Speech Technology in Language Learning*: Recognition, Synthesis, Visualisation, Talking Heads and Integration. Swets and Zeitlinger.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Part IV

Issues in Segmentation and Mark-up

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

29

Issues in Segmentation and Mark-up

Mark Huckvale
Phonetics and Linguistics, University College London
Gower Street, London, WC1E 6BT, UK
mark@phonetics.ucl.ac.uk

The chapters in this section discuss meta-level descriptions of language data in speech synthesis systems. In the conversion of text to a synthetic speech signal the text and the signal are explicit: we see the text going in and we hear the speech coming out. But a synthetic speech signal is also an interpretation of the text, and thus contains implicit knowledge of the mapping between this interpretation and the stored language data on which the synthesis is based: text conventions, grammar rules, pronunciations and phonetic realisations. To know how to synthesise a convincing version of an utterance requires a linguistic analysis and the means to realise the components of that analysis. Meta-level descriptions are used in synthesis to constrain and define linguistic analyses and to allow stored data to be indexed, retrieved and exploited.

As the technology of speech synthesis has matured, the meta-level descriptions have increased in sophistication. Synthesis systems are now rarely asked to read plain text; instead they are given e-mail messages or web pages or records from databases. These materials are already formatted in machine-readable forms, and the information systems that supply them know more about the text than can be inferred from the text itself. For example, e-mail systems know the meanings of the mail headers, or web browsers know about the formatting of web pages, or database systems know the meaning of record fields. When we read about standards for 'mark-up' of text for synthesis, we should see these as the first attempt to encode this meta-level knowledge in a form that the synthesis system can use. Similarly, the linguistic analysis that takes place inside a synthesis system is also meta-level description: the grammatical, prosodic and phonological structure of the message adds to the text. Importantly this derived information allows us to access machine pronunciation dictionaries or extract signal 'units' from corpora of labelled recordings. What information we choose to put in those meta-level descriptions constrains how the system operates: how is pronunciation choice affected by the

context in which a word appears? Does the position of a syllable in the prosodic structure affect which units are selected from the database? There are also many practical concerns: how does the choice of phonological description affect the cost of producing a labelled corpus? How does the choice of phonological inventory affect the precision of automatic labelling? What are the perceptual consequences of a trade-off between pitch accuracy and temporal accuracy in unit selection?

The five chapters that follow focus on two main issues: how should we go about marking up text for input to synthesis systems, and how can we produce labelled corpora of speech signals cheaply and effectively? Chapter 30 by Huckvale describes the increasing influence of the mark-up standard XML within synthesis, and demonstrates how it has been applied to mark up databases, input text, and dialogue systems as well as for linguistic description of both phonological structure and information structure. The conclusions are that standards development forces us to address significant linguistic issues in the meta-level description of text. Chapter 31 by Monaghan discusses how text should be marked up for input to synthesis systems: what are the fundamental issues and how are these being addressed by the current set of proposed standards? He concludes that current schemes are still falling into the trap of marking up the form of the text, rather than marking up the function of the text. It should be up to synthesis systems to decide how to say the text, and up to the supplier of the text to indicate what the text means. Chapter 32 by Hirst presents a universal tool for characterising F0 contours which automatically generates a mark-up of the intonation of a spoken phrase. Such a tool is a prerequisite for the scientific study of intonation and the generation of models of intonation in any language. Chapter 33 by Horák explores the possibility of using one synthesis system to 'bootstrap' a second generation system. He shows that by aligning synthetic speech with new recordings, it is possible to generate a new labelled database. Work such as this will reduce the cost of designing new synthetic voices in the future. Chapter 34 by Warakagoda and Natvig explores the possibility of using speech recognition technology for the labelling of a corpus for synthesis. They expose the cultural and signal processing differences between the synthesis and recognition camps.

Commercialisation of speech synthesis will rely on producing speech which is expressive of the meaning of the spoken message, which reflects the information structure implied by the text. Commercialisation will also mean more voices, made to order more quickly and more cheaply. The chapters in this section show how improvements in mark-up and segmentation can help in both cases.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

30

The Use and Potential of Extensible Mark-up (XML) in Speech Generation

Mark Huckvale
Phonetics and Linguistics, University College London
Gower Street, London, WC1E 6BT, UK
mark@phonetics.ucl.ac.uk

Introduction

The Extensible Mark-up Language (XML) is a simple dialect of Standard Generalised Mark-up Language (SGML) designed to facilitate the communication and processing of textual data on the Web in more advanced ways than is possible with the existing Hypertext Mark-up Language (HTML). XML goes beyond HTML in that it attempts to describe the *content* of documents rather than their *form*. It does this by allowing authors to design mark-up that is specific to a particular application, to publish the specification for that mark-up, and to ensure that documents created for that application conform to that mark-up. Information may then be published in an open and standard form that can be readily processed by many different computer applications.

XML is a standard proposed by the World Wide Web Consortium (W3C). W3C sees XML as a means of encouraging 'vendor-neutral data exchange, media-independent publishing, collaborative authoring, the processing of documents by intelligent agents and other metadata applications' (W3C, 2000).

XML is a dialect of SGML specifically designed for computer processing. XML documents can include a formal syntactic description of their mark-up, called a Document Type Definition (DTD), which allows a degree of content validation. However, the essential structure of an XML document can be extracted even if no DTD is provided. XML mark-up is hierarchical and recursive, so that complex data structures can be encoded. Parsers for XML are fairly easy to write, and there are a number of publicly available parsers and toolkits. An important aspect of XML is that it is designed to support Unicode representations of text so that all European and Asian languages as well as phonetic characters may be encoded.

Here is an example of an XML document:

```
<?xml version ='1.0'?>
<!DOCTYPE LEXICON[
<! ELEMENT LEXICON (ENTRY) * >
<! ELEMENT ENTRY (HW, POSSEQ, PRONSEQ) >
<! ELEMENT HW (# PCDATA) >
<! ELEMENT POSSEQ (POS) * >
<! ELEMENT POS (# PCDATA) >
<! ELEMENT PRONSEQ (PRON) * >
<! ELEMENT PRON (#PCDATA) >
<! ATTLIST  ENTRY
       ID       ID       #REQUIRED>
<! ATTLIST  POS
       PRN      CDATA    #REQUIRED>
<! ATTLIST  PRON
       ID       ID       #REQUIRED>
] >
<LEXICON>
<ENTRY ID="READ">
  <HW> read </HW>
  <POSSEQ>
   <POS PRN="#ID(READ-1)">
      V (past) </POS>
   <POS PRN="#ID(READ-2)">
      V (pres) </POS>
   <POS PRN="#ID(READ-2)">
      N (com, sing) </POS>
  </POSSEQ>
  <PRONSEQ>
   <PRON ID="READ-1">'red </PRON>
   <PRON ID="READ-2">'rid </PRON>
  </PRONSEQ>
</ENTRY>
...
</LEXICON>
```

In this example the heading '<?xml ... ? >' identifies an XML document in which the section from ' <! DOCTYPE LEXICON [' to '] >' is the DTD for the data marked up between the <LEXICON> and </LEXICON> tags. This example shows how some of the complexity in a lexicon might be encoded. Each entry in the lexicon is bracketed by <ENTRY>; within this are a headword <HW>, a number of parts of speech <POSSEQ>, and a number of pronunciations <PRONSEQ>. Each part of speech section <POS> gives a grammatical class for one meaning of the word. The <POS> tag has an *attribute* PRN, which identifies the ID attribute of the relevant pronunciation <PRN>. The DTD provides a formal specification of the tags, their nesting, their attributes and their content.

XML is important for development work in speech synthesis at almost every level. XML is currently being used for marking up corpora, for marking up text to

be input to text-to-speech systems, for marking up simple dialogue applications. But these are only the beginning of the possibilities: XML could also be used to open up the internals of synthesis-by-rule systems. This would give access to their working data structures and create open architectures allowing the development of truly distributed and extensible systems. Joint efforts in the standardisation of mark-up, particularly at the higher linguistic levels, will usefully force us to address significant linguistic issues about how language is used to communicate.

The following sections of this chapter describe some of the current uses of XML in speech generation and research, how XML has been used in the ProSynth project (ProSynth, 2000) to create an open synthesis architecture, and how XML has been used in the SOLE project (Sole, 2000) to encode textual information essential for effective prosody generation.

Current Use of XML in Speech Generation

Mark-up for Spoken Language Corpora

The majority of spoken language corpora available today are distributed in the form of binary files containing audio and text files containing orthographic transcription with no specific or standardised markup. This reflects the concentration of effort in speech recognition on the mapping between the signal and the word sequence. It is significant that missing from such data is a description of the speaker, the environment, the goals of the communication or its information content. Speech recognition systems cannot, on the whole, exploit prior information about such parameters in decoding the word sequence. On the other hand, speech synthesis systems must explicitly model speaker and environment characteristics, and adapt to different communication goals and content.

Two recent initiatives at improving the level of description of spoken corpora are the American Discourse Resource Initiative (DRI, 2000) and the Multi-level Annotation Tools Engineering project (MATE, 2000). The latter project aims to propose a standard for the annotation of spoken dialogue covering levels of prosody, syntax, co-reference, dialogue acts and other communicative aspects, with an emphasis on interactions between levels. In this regard they have been working on a multi-level XML description (Isard *et al.*, 1998) and a software workbench for annotation.

In the multi-level framework, the lowest-level XML files label contiguous stretches of audio signals with units that represent phones or words, supported by units representing pauses, breath noises, lip-smacks, etc. The next level XML files group these into dialogue moves by each speaker. Tags in this second level link to one or more units in the lowest level file. Further levels can then be constructed, referring down to the dialogue moves, which might encode particular dialogue strategies. Such a multi-level structure allows correlations to be drawn between the highest-level goals of the discourse and the moves, words and even the prosody used to achieve them.

Mark-up of Text for Input to TTS

SABLE is an XML-based markup scheme for text-to-speech synthesis, developed to address the need for a common text-to-speech (TTS) control paradigm (Sable, 2000). SABLE provides a standard means for marking up text to be input to a TTS system to identify particular characteristics of the text, or of the required speaker, or the required realisation. SABLE is intended to supersede a number of earlier control languages, such as Microsoft SAPI, Apple Speech Manager, or the Java Speech Mark-up Language (JSML).

SABLE provides mark-up tags for Speaker Directives: for example, emphasis, break, pitch, rate, volume, pronunciation, language, or speaker type. It provides tags for text description: for example, to identify times, dates, telephone numbers or other common formats; or to identify rows and columns in a table. It can also be extended for specific TTS engines and may be used to aid in synchronisation with other media.

Here is a simple example of SABLE:

```
<DIV TYPE="paragraph">
   New e-mail from
   <EMPH> Tom Jones </EMPH> regarding
   <PITCH BASE="high" RANGE="large">
      <RATE SPEED="−20%">
         latest album </RATE>
   </PITCH>.
</DIV>
<AUDIO SRC="beep.aiff"/>
```

In this example, the subject of an e-mail is emphasised by setting a higher base pitch, a larger pitch range and a slower rate. Information necessary to specify such a requirement would come from the e-mail reader application which has privileged access to the structure of the source data. The message is terminated by an audible beep.

Mark-up of Speech-Driven Applications

VoiceXML is an XML-based language for building network based conversational applications (VoiceXML, 2000). Such applications interact with users by voice output–input in a manner analogous to how a web browser interacts with a user using screen and keyboard. VoiceXML is supported by a voice-driven browser that exploits the recognition and synthesis technology of IBM ViaVoice products. VoiceXML is not designed for general purpose dialogue systems, but can be used to build conversational applications that involve menu choices, form filling and TTS.

To construct a VoiceXML application, pages of marked text are processed by the voice browser which speaks prompts and accepts verbal responses restricted by menus or validated form fields. At the heart of VoiceXML are the tags <VXML>: which groups VoiceXML elements like an HTML page; <MENU>: which presents a set of choices and target links; <FORM>: which groups fields of information

required from user; and <PROMPT>: which identifies a chunk of text to be spoken. Output text can be marked up with JSML, and input responses can be constrained by a simple grammar.

Here is a simple example of VoiceXML:

```
<menu>
<prompt>Welcome home. Say one of: <enumerate/></prompt>
<choice next= "http://www.sports.example/vxml/start.vxml">
Sports </choice>
<choice next="http://www.weather.example/intro.vxml">
Weather </choice>
<choice next="http://www.stargazer.example/voice/astronews.vxml">
Stargazer astrophysics news </choice>
<noinput>Please say one of <enumerate/></noinput>
</menu>
```

In this example of a <MENU> dialog, the welcome message in the <PROMPT> tag, is followed by the list of choices to the user. If the user repeats back one of the prompts, the relevant page is loaded according to the NEXT attribute of the <CHOICE> tag. This dialog might proceed as follows:

C: Welcome home. Say one of: sports; weather; Stargazer astrophysics news.
H: Astrology.
C: I did not understand what you said.
C: Welcome home. Say one of: sports; weather; Stargazer astrophysics news.
H: Sports.
C: (*proceeds to* http://www.sports.example/vxml/start.vxml)

Potential for XML in Speech Generation

The emerging standards for mark-up described above, the MATE project for corpora, the SABLE system for TTS and the VoiceXML system for applications, are important to the development of speech synthesis systems, but they do not address a number of significant issues. This section draws examples from the recent research projects to demonstrate how XML could help address the problems of proprietary synthesis architectures, knowledge representation, and inexpressive delivery.

Opening up Synthesis Architectures

An important contribution to current research and development activities in speech synthesis has been made by open source initiatives such as Festival (Festival, 2000), and public domain resources such as MBROLA (MBROLA, 2000). However, even these systems retain proprietary data formats for working data structures, and use knowledge representation schemes closely tied to those structures. This means that phoneticians and linguists willing and able to contribute to better synthesis systems are presented with complex and arbitrary interfaces which require considerable investment to conquer.

An alternative is to provide open, non-proprietary textual representations of data structures at every level and stage of processing. In this way additional or alternative components may be easily added even if they are encoded in different computer languages and run on different machines. In the ProSynth project (Ogden *et al.*, 2000), XML is used to encode the external data structures at all levels and stages. Synthesis is a pipeline of processes that perform utterance composition and phonetic interpretation. These processes are constructed to take XML-marked input, to modify structures and attributes, and to generate XML-marked output. As well as the representation of the utterance undergoing interpretation, XML is also used to mark up the input text and the pronunciation lexicon. For output, the XML format is converted to proprietary formats for MBROLA, HLSyn (see Heid and Hawkins, 1998) or for prosody-manipulated natural speech.

Here is a fragment of working data structure from ProSynth:

```
<AG ACCENT="H* L" TYPE="NUCLEAR">
 <FOOT DUR="1" FPITCH="100" IPITCH="140" LON="50" POF="30" PON="23"
STRENGTH="STRONG">
  <SYL DUR="1" FPOS="1" RFPOS="2" RWPOS="2" STRENGTH="STRONG"
WEIGHT="HEAVY" WPOS="1" WREF="WORD3">
   <ONSET DUR="1" STRENGTH="STRONG">
    <CNS AMBI="N" CNSCMP="N" CNSGRV="N" CNT="Y" DUR="1"
INHDUR="0.125" MINDUR="0.08" NAS="N" RHO="N" SON="N" STR="Y"
VOCGRV="Y" VOCHEIGHT="OPEN" VOCRND="N" VOI="N">s</CNS>
   </ONSET>
   <RHYME CHECKED="Y" DUR="1" STRENGTH="STRONG" VOI="N"
WEIGHT="HEAVY">
    <NUC CHECKED="Y" DUR="0.4896" LONG="Y" STRENGTH="STRONG" VOI="N"
WEIGHT="HEAVY">
      <VOC DUR="1" GRV="Y" HEIGHT="OPEN" INHDUR="0.11"
MINDUR="0.035" RND="N">A</VOC>
      <VOC DUR="1" GRV="Y" HEIGHT="OPEN" INHDUR="0.11"
MINDUR="0.035" RND="N">A</VOC>
    </NUC>
    <CODA DUR="1" VOI="N">
     <CNS AMBI="N" CNSCMP="N" CNSGRV="Y" CNT="N" DUR="0.85"
INHDUR="0.11" MINDUR="0.05" NAS="Y" RHO="N" SON="Y" STR="N"
VOCGRV="Y" VOCHEIGHT="OPEN" VOCRND="N" VOI="Y">m</CNS>
     <CNS AMBI="Y" CNSCMP="N" CNSGRV="Y" CNT="N" DUR="0.85"
INHDUR="0.08" MINDUR="0.06" NAS="N" RHO="N" SON="N" STR="N"
VOCGRV="Y" VOCHEIGHT="OPEN" VOCRND="N" VOI="N">p</CNS>
    </CODA>
   </RHYME>
  </SYL>
```

This extract is the syllable 'samp' from the phrase 'it's a sample'. The phone transcription /sAAmp/ is marked by CNS (consonant) and VOC (vocalic) nodes. These are included in ONSET, NUC (nucleus) and CODA nodes, which in turn form RHYME and SYL (syllable) constituents. The SYL nodes occur under FOOT

nodes, and the FOOT under AG (accent group) nodes. Phonetic interpretation has set some attributes on the nodes to define the durations and fundamental frequency contour.

Declarative Knowledge Representation

A continuing difficulty in the creation of open architectures for speech synthesis is the interdependency of rules for transforming text to a realised phonetic transcription. Context-sensitive rewrite rule formalisms are a particular problem: the output of one rule typically feeds many others in ways that make it difficult to know the effect of a change. Often a new rule or a change to the ordering of rules can break the system.

It is generally accepted that the weaknesses of rewrite rules can be overcome by a declarative formalism. With a declarative knowledge representation, a structure is enhanced and enriched rather than modified by matching rules. Changes to the structure are always performed in a reversible way, so that rule ordering is not an issue. In ProSynth, the context for phonetic interpretation is established by the metrical hierarchy extending within and above the syllable. Thus the realisation of a phone can depend on where in a syllable it occurs, where the syllable occurs in a foot, and where the foot occurs in an accent group or intonation phrase. Thus context is established hierarchically rather than left and right. Knowledge for phonetic interpretation is expressed as declarative rules which modify attributes stored in the working data structure which is externally represented as XML.

The language formalism for knowledge representation used in ProSynth is called ProXML. Phonetic interpretation knowledge stored in ProXML is interpreted to translate one stream of XML into another in the synthesis pipeline. The ProXML language draws on elements of Cascading Style Sheets as well as the 'C' programming language (see Huckvale, 1999 for more information).

Here is a simple example of ProXML:

```
/* Klatt Rule 9: Postvocalic context of vowels */
NUC{
   node coda = ../RHYME/CODA;
   if (coda==nil)
    :DUR *= 1.2;
   else  {
    node cns = coda/CNS;
    if ((cns:VOI=="Y") &&
        (cns:CNT=="Y") &&
        (cns:SON=="N"))
     :DUR *= 1.6;
    else if ((cns:VOI=="Y") &&
             (cns:CNT=="N") &&
             (cns:SON=="N"))
     :DUR *= 1.2;
```

```
      else if ((cns:VOI=="Y") &&
               (cns:NAS=="Y") &&
               (cns:SON=="Y"))
        :DUR *= 0.85;
      else if ((cns:VOI=="N") &&
               (cns:CNT=="N") &&
               (cns:SON=="N"))
        :DUR *= 0.7;
    }
}
```

This example, based on Klatt duration rule 9 (Klatt, 1979), operates on all NUC (vowel nucleus) nodes. The relative duration of a vowel nucleus, DUR, is calculated from properties of the rhyme: in particular whether the coda is empty, has a voiced fricative, a voiced stop, a nasal or a voiceless stop. The statement ':DUR *= 0.7' means adjust the current value of the DUR attribute (of the NUC node) by the factor 0.7.

Modelling Expressive Prosody

Despite recent improvements in signal generation methods, it is still the case that synthetic speech sounds monotonous and generally inexpressive. Most systems deliberately aim to produce neutral readings of plain text; they do not try to interpret the text nor construct a spoken phrase to have some desired result. This lack of expressiveness is due to the poverty of the underlying linguistic representation: text analysis and understanding systems are simply not capable of delivering high-quality interpretations directly from unmarked input. However, for many applications, such as information services, the text itself is generated by the computer system, and its meaning is available alongside information about the state of the dialogue with the user.

The problem then becomes how to mark up the appropriate information structure and discourse function of the text in such a way that the speech generation system can deliver appropriate and expressive prosody. Note that neither the SABLE system nor the MATE project address this problem directly. As can be seen from the example, SABLE is typically used to simply indicate emphasis, or to modify prosody parameters directly. Mark up in MATE is a standard for actual human discourse, not for input to synthesis systems.

In the SOLE project, descriptions of museum objects are automatically generated and spoken by a TTS system. The application thus has knowledge of the meaning and function of the text. To obtain effective prosody for such descriptions, XML mark-up is used to identify rhetorical structure, noun-phrase type, and topic/comment structure, on top of standard punctuation (Hitzeman *et al.*, 1999).

Here is a simple example of text marked up for rhetorical relations:

```
<rhet-elem type="contrast">
  <nucleus> The
```

```
      <rhet-emph type="object">
        god </rhet-emph>
      was
      <rhet-emph type="property">
        gilded </rhet-emph>;
    </nucleus>
    <nucleus> the
      <rhet-emph type="object">
        demon </rhet-emph>
      was
      <rhet-emph type="property">
         stained in black ink and polished to a high sheen
      </rhet-emph>.
    </nucleus>
  </rhet-elem>
```

In this example, a contrast is drawn between the gilding of the god and the staining of the demon. The rhetorical structure is one of contrast, and contains elements of rhetorical emphasis appropriate for objects and properties.

It is clear that much further work is required in this area, in particular to decide on which aspects of information structure or discourse function have effects on prosody. Mark-up for dialogue would also have to take into account the modelled state of the listener; it would indicate which information was given, new or contradictory. Such mark-up might also express the degree of 'certainty' of the information, it might convey 'urgency' or 'deliberation'; even 'irritation' or 'conspiracy'.

Conclusion

This is an exciting time for synthesis: open architectures and open sources, large corpora, powerful computer systems, quality public-domain resources. But the availability of these has not replaced the need for detailed phonetic and linguistic analysis of the interpretation and realisation of linguistic structures. Progress will require the efforts of a multidisciplinary team distributed across many sites. XML provides standards, open architectures, declarative knowledge formalisms, computational flexibility and computational efficiency to support future speech generation systems. Rather than being a regressive activity, standards development forces us to address significant issues in the classification and representation of linguistic events in spoken discourse.

Acknowledgements

The author wishes to thank the ProSynth project team in York, Cambridge and UCL. Thanks also go to COST 258 for providing a forum for discussion about mark-up. The ProSynth project is supported by the UK Engineering and Physical Sciences Research Council.

References

Discourse Resource Initiative. Retrieved 11 October 2000 from the World Wide Web: http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html

Festival. The Festival Speech Synthesis System. Retrieved 11 October 2000 from the World Wide Web: http://www.cstr.ed.ac.uk/projects/festival/

Heid, S. and Hawkins, S. (1998). PROCSY: A hybrid approach to high-quality formant synthesis using HLSyn. *Proceedings of 3rd ESCA/COCOSDA International Workshop on Speech Synthesis* (pp. 219–224). Jenolan Caves, Australia.

Hitzeman, J., Black, A., Mellish, C., Oberlander, J., Poesio, M., and Taylor, P. (1999). An annotation scheme for concept-to-speech synthesis. *Proceedings of European Workshop on Natural Language Generation* (pp. 59–66). Toulouse, France.

Huckvale, M.A. (1999). Representation and proceedings of linguistic structures for an all-prosodic synthesis system using XML. *Proceedings of EuroSpeech-99* (pp. 1847–1850). Budapest, Hungary.

Isard, A., McKelvie, D., and Thompson, H. (1998). Towards a minimal standard for dialogue transcripts: A new SGML architecture for the HCRC map task corpus. *Proceedings of International Conference on Spoken Language Processing* (pp. 1599–1602). Sydney, Australia.

Klatt, D. (1979). Synthesis by rule of segmental durations in English sentences In B. Lindblom and S. Ohman (eds), *Frontiers of Speech Communication Research* (pp. 287–299). Academic Press.

Mate. Multilevel Annotation, Tools Engineering. Retrieved 11 October 2000 from the World Wide Web: http://mate.nis.sdu.dk/

MBROLA. The MBROLA Project. Retrieved 11 October 2000 from the World Wide Web: http://tcts.fpms.ac.be/synthesis/mbrola.html

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovicova, J., and Heid, S. (2000). ProSynth: An integrated prosodic approach to device-independent natural-sounding speech synthesis. *Computer Speech and Language*, *14*, 177–210.

ProSynth. An integrated prosodic approach to device-independent, natural-sounding speech synthesis. Retrieved October 11, 2000 from the World Wide Web: http://www.phon.ucl.ac.uk/project/prosynth.html

Sable. A Synthesis Markup Language. Retrieved 11 October 2000 from the World Wide Web: http://www.bell-labs.com/project/tts/sable.html

Sole. The Spoken Output Labelling Explorer Project. Retrieved 11 October 2000 from the World Wide Web: http://www.cstr.ed.ac.uk/projects/sole.html

VoiceXML. Retrieved 11 October 2000 from the World Wide Web: http://www.alphaworks.ibm.com/tech/voicexml

W3C. Extensible Mark-up Language (XML). Retrieved 11 October 2000 from the World Wide Web: http://www.w3c.org/XML

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

31

Mark-up for Speech Synthesis

A Review and Some Suggestions

Alex Monaghan
Aculab Plc, Lakeside, Bramley Road
Mount Farm, Milton Keynes MK1 1PT, UK
Alex.Monaghan@aculab.com

Introduction

This chapter reviews the reasons for using mark-up in speech synthesis, and examines existing proposals for mark-up. Some of the problems with current approaches are discussed, some solutions are suggested, and alternative approaches are proposed.

For most major European languages, and many others, there now exist synthesis systems which take plain text input and produce reasonable quality output (intelligible, not too mechanical and not too monotonous). The main deficit in these systems, and the main obstacle to user acceptance, is the lack of appropriate prosody (Sonntag, 1999; Sonntag *et al.*, 1999; Sluijter *et al.*, 1998). In the general case, prosody (pausing, F0, duration and amplitude, amongst other things) is only partially predictable from unrestricted plain text (Monaghan, 1991; 1992; 1993). Interestingly, the phonetic details of synthetic prosody do not appear to make much difference: the choice of straight-line or cubic spline interpolation for modelling F0, or a duration of 250 ms or 300 ms for a major pause, is relatively unimportant. What matters is the marking of *structure* and *salience*: pauses and emphasis must be placed correctly, and the hierarchy of phrasing and prominence must be adequately conveyed. These are difficult tasks, and there is widespread acceptance in the speech technology community that the generation of entirely appropriate prosody for unrestricted plain text will have to wait for advances in natural language processing and linguistic science.

At the same time, there is an increasing amount and range of non-plain-text material which could be used as input by speech synthesis applications. This material includes formatted documents (such as this one), e-mail messages, web pages,

and the output of automatic systems for database query (DBQ) or natural language generation (NLG). In all these cases, the material provides information which is not easily extracted from plain text, and which could be used to improve the naturalness and comprehensibility of synthetic speech. The encoding, or mark-up, of this information generally indicates the structure of the material and the relative importance of various items, which is exactly the sort of information that speech synthesis systems require to generate appropriate prosody. Mark-up therefore provides the possibility of deducing appropriate prosody for non-plain-text material, and of adding prosodic and other information explicitly for particular applications. Its use should allow speech synthesisers to achieve a level of naturalness and expressiveness which has not been possible from plain text input.

In order for synthesis systems to make use of this additional information, they must either process the mark-up directly or translate it into a more convenient representation. Some systems already have an internal mark-up language which allows users to annotate the input text (e.g. DECtalk, Festival, INFOVOX), and many applications process a specific mark-up language to optimise the output speech. There are also several general-purpose mark-up standards which are relevant to speech synthesis, and which may be useful for a broader range of applications. At the time of going to press, speech synthesis mark-up proposals are still emerging. In the next few years, mark-up in this area will become standardised and we may see new applications and new users of speech synthesis as a result.

If speech synthesis systems are to make effective use of mark-up, there are three basic questions which should be answered:

- Why is the mark-up being used?
- How is the mark-up being generated?
- What is the set of markers?

These questions are discussed in the remainder of this chapter with reference to various applications and existing mark-up proposals. If they can be answered, for a particular system or application, then the naturalness and acceptability of the synthetic speech may be dramatically increased. A certain amount of scene-setting is required before we can address the main issues, so we will briefly outline some major applications of speech synthesis and the importance of prosodic mark-up in such applications.

Telephony Applications

The major medium-term applications of speech synthesis are mostly based on telephony. These include remote access to stored information over the telephone, simple automatic services such as directory enquiries and home banking, and full-scale interactive dialogue systems for booking tickets, completing forms, or providing specialist helpdesk facilities. Such applications generally require the synthesis of large or complex pieces of text (typically one or more paragraphs). For present purposes we can identify four different classes of input to telephony applications:

1. Formatted text
2. Known text types
3. Automatically generated text
4. Canned text.

Formatted text

Most machine-readable text today is formatted in some way. Documents are prepared and exchanged using various word-processing (WP) formats (WORD, LaTeX, RTF, etc.). Data are represented in spreadsheets and a range of database formats. Specific formats exist for address lists, appointments, and other commonly used data. Many large companies have proprietary formats for their data and documents, and of course there are universal formatting languages developed for the Internet.

Speech synthesis from formatted text presents a paradox. On the one hand, conventional plain-text synthesisers cannot produce acceptable spoken renditions of such text, because they read out the formatting codes as e.g. '*backslash subsection asterisk left brace introduction right brace*' which renders much of the text incomprehensible. On the other hand, if the synthesiser were able to recognise this as a section heading command, it could use that information to improve the naturalness and comprehensibility of its output (by, say, slowing down the speech rate for the heading and putting appropriate pauses around it). As word processing and data formats become ever more widespread, there will be an increasing amount of material which is paradoxically inaccessible via plain-text synthesisers but produces very high quality output from synthesisers which can process the mark-up codes.

Known Text Types

In many applications, the type of information which the text contains is known in advance. In an e-mail reader, for instance, we know that there will be a header and a body, and possibly a footer; we know that the header will contain information about the provenance of the message and probably also its subject; we know that the body may contain unrestricted text, but that certain conventions (such as special abbreviations, smiley faces, and attached files) apply to message bodies; and we know that the footer contains information about the sender such as a name, address and telephone number. Depending on the user's preferences, a speech synthesiser can suppress most of the information in the header (possibly only reading out the date, sender, and subject information) and can respond appropriately to smileys, attachments and other domain-specific items in the body. This level of sophistication is possible because the system knows the characteristics of the input, and because the different types of information are clearly marked in the document.

E-mail messages are actually plain text documents, but their contents are so predictable that they can be interpreted as though they were formatted. Many other types of information follow a similarly predictable pattern: address lists (name, street address, town, post code, telephone number), invoices (item, quantity, unit price, total), and more complex documents such as web pages and online forms. Some of these have explicit mark-up codes. In others the formatting is

implicit in the punctuation, line breaks and key words. Either way, synthesis systems could take advantage of the predictable content and structure of these text types to produce more appropriate spoken output.

Automatically Generated Text

This type of text is still relatively rare, but its usage is growing rapidly. Commoner examples of automatically generated text include web pages generated by search engines or DBQ systems, error messages generated by all types of software, and the output of NLG systems such as chatbots[1] or dialogue applications (Luz, 1999). The use of autonomous agents and spoken dialogue systems is predicted to increase dramatically in the next few years, and other applications such as automatic translation and the generation of official documents may become more widespread.

The crucial factor in all these examples is that the system which generates the text possesses a large amount of high-level knowledge about the text. This knowledge often includes the function, meaning and context of the text, since it was generated in response to a particular command and was intended to convey particular information. Take a software error message such as '*The requested URL /~alex/fg was not found on this server.*': the system knows that this message was generated in response to an HTTP command, that the URL was the one provided by the user, and that this error message does not require any further action by the user. The system also knows what the server name is, how fast the HTTP connection was, and various other things including the fact that this error will not usually cause the system to crash. Much of this information is not relevant for speech synthesis, but we could imagine a synthesiser which used different voice qualities depending on the seriousness of the error message, or different voices depending on the type of software which generated the message.

Applications such as automatic translation and spoken dialogue systems generally perform exactly the kind of deep linguistic analysis which would allow a synthesiser to generate optimal pronunciation and prosody for a given text. In some cases there is actually no need to generate text at all: the internal representation used by the generation system can be processed directly by a synthesis system. Automatically generated input therefore offers the best hope of natural-sounding synthetic speech in the short to medium term, and could be viewed as one extreme of the mark-up continuum, with plain text at the other extreme.

Canned Text

Many telephony applications of speech synthesis involve restricted domains or relatively static text: examples include dial-up weather forecasts (a 200-word text which only changes three times a day) and speech synthesis of teletext pages. In these applications, the small amount of text and the infrequent updates mean that manual or semi-automatic mark-up is possible. The addition of mark-up to soft-

[1] For example, Alice, http://206.184.206.210/alice_page.htm

ware menus and dialogue boxes, or to call-centre prompts, could greatly improve the quality of synthesiser output and thus the level of customer satisfaction.

Adding mark-up manually, perhaps using a customised editing tool, would be relatively inexpensive for small amounts of text and could solve many of the grosser errors in synthesis from plain text, such as the difficulty of deciding whether '5/6' should be read as 'five sixths' or as a date, or the impossibility of predicting how a particular company name or memorable telephone number should be rendered.

Prosodic Mark-up

Mark-up can be used to achieve many things in synthetic speech: to change voice or language variety, to specify pronunciation for unusual words, to insert non-speech sounds into the output, and even to synchronise the speech with images or other software processes. Useful though these all are, the main area in which mark-up can improve synthetic speech is the control of prosody. The lack of appropriate prosody is generally seen as the single biggest problem with current speech synthesis systems. In a recent evaluation of e-mail readers for mobile phones, published in June 2000 (http://img.cmpnet.com/commweb 2000/whites/umtestingreport.pdf), CT Labs tested four synthesis systems which had been adapted to process email messages (Monaghan (a), Webpage). All four were placed at the mid-point of a five-point scale. Approximately two-thirds of the glaring errors produced by these systems were errors in prosody: inappropriate pausing, emphasis or F0 contours.

Adding prosodic mark-up to a text involves two types of information: information about the structure of the text, and information about the importance or salience of items within that structure. These notions of structure and salience are central to prosody.

Prosodic structure conveys the boundaries between topics, between paragraphs within a topic, and between sentences within a paragraph. These boundaries are generally realised by different durations of pausing, as well as by boundary tones, speech rate changes, and changes in pitch register. Within a single sentence or utterance, smaller phrases may be realised by shorter pauses or by pitch changes.

Prosodic salience conveys the relative importance of items within a unit of structure. It is generally realised by pitch excursions and increases in duration on the salient items, but may also involve changes in amplitude and articulatory effort. It depends on pragmatic, semantic, syntactic and other factors, particularly the notion of focus (Monaghan, 1993).

Although prosody is a difficult problem for plain-text synthesis, it is largely a solved problem once we allow annotated input. Annotating the structure and marking the salient items of a text require training, but they can be done reliably and consistently. Indeed, the formatting of text using a WP package, or using HTML for web pages, is nothing other than marking structure and salience. The use of paragraph and section breaks, indenting, centring and bullet points shows the structure of a document, while devices such as capitalisation, bolding, italics, underlining and different font sizes indicate the relative salience of items.

Why mark-up?

This is the simplest of our three questions to answer: because it's there! Many documents already contain mark-up: to process them without treating the mark-up specially would produce disastrous results (Monaghan (b), Webpage), and to remove the mark-up would be both awkward and illogical since we would be discarding useful information. A synthesis system which can process mark-up intelligently is able to produce optimal output from formatted documents, web pages, DBQ and NLG systems, and many more input types than a system which only handles plain text.

Even if the document you are processing does not already contain explicit mark-up, for many text types it is quite simple to insert mark-up automatically. The current Aculab TTS system[2] processes e-mail headers to extract date, sender, subject and other information, allowing it to ignore information which is irrelevant or distracting for the user. Similar techniques could improve the quality of speech synthesis for telephone directory entries, online forms, stock market indices, and any other sufficiently predictable format.

If we wish to synthesise speech from the output of an automatic system, such as a spoken dialogue system or a machine translation package, it simply doesn't make sense to pass through an intermediate stage of plain text just because the synthesiser cannot handle other formats. The information on structure, salience and other linguistic factors which is available in the internal representations of dialogue or translation systems is the answer to the prayers of synthesiser developers. Spoken language translation projects such as Verbmobil (Wahlster, 2000) rely on this information to drive their speech synthesisers. The ability to use such rich information sources will distinguish between the rather dull synthesis of today and the expressive speech-based interfaces of tomorrow.

To put it bluntly, mark-up adds information to the text. Such information can be used by a synthesis system to improve the prosody and other aspects of the spoken output. This is, after all, the main motivation for producing formatted documents and web pages: the formatting adds information, it makes the structure of the document more obvious and draws attention to the salient items.

How mark-up?

This question is not too difficult to answer after the discussions above. For formatted documents and automatically generated text, the mark-up has already been inserted: all the synthesis system has to do is interpret it. Interpretation of mark-up is not a trivial task, but we have shown elsewhere that it can be done and that it can dramatically improve the quality of synthetic speech (Monaghan, 1994; Fitzpatrick and Monaghan, 1998; Monaghan, 1998; Fitzpatrick, 1999; Monaghan (c), Webpage).

For known text types, certain key words or character sequences can be identified and replaced by mark-up codes. In the Aculab TTS email pre-processor, text strings such as 'Subject:' in the message header are recognised and these prompt

[2] Downloadable from http://www.aculab.com

the system to process the rest of the line in a special way. Similarly, predictable information in the message body (indentation, Internet addresses, smileys, separators, attached files, etc.) can be automatically identified by unique text strings and processed appropriately. Comparable techniques could be applied to telephone directory entries (Monaghan (d), Webpage), invoices, spreadsheets, and any text type where the format is known in advance.

The manual annotation of text is quite time-consuming, but is still feasible for small amounts of text which are not frequently updated. Obviously, the amount of text and the frequency of updates should be in inverse proportion. Good candidates for manual annotation would include weather forecasts, news bulletins, teletext information, special offers, and short web page updates or 'Message of the Day' text. The actual process of annotating might be based on mark-up templates for a particular application, or simply on trial and error. At least one authoring tool for speech synthesis mark-up already exists (Wouters, *et al.*, 1999), and there will no doubt be more, so it may be possible to speed up the annotation process considerably.

What mark-up?

This is the tough one. What we would all like is an ideal mark-up language for speech synthesis which incorporates all the features required by system developers, application developers, researchers and general users. This ideal mark-up would have at least the following characteristics:

- the possibility of specifying exactly how something should sound (including pronunciation, emphasis, pitch contour, duration, amplitude, voice quality, articulatory effort, etc.): this gives low-level control of the synthesiser;
- intuitive, meaningful, easy-to-use categories (spell out a string of characters; choose pronunciation in French, English, German or any other language; specify emotions such as sad, happy or angry, etc.): this gives high-level control of the synthesiser;
- device-independence, so it has the same effect on any synthesiser.

Of course, no such mark-up language exists or is likely to in the near future. The problems of reconciling high-level control and low-level control are considerable, and the goal of device-independence is currently unattainable because of the proliferation of architectures, methodologies and underlying theories in existing speech synthesis systems (Monaghan, Chapter 9, this volume). What does exist is a number of mark-up languages which have had varying degrees of success. Some of these are specific to speech synthesis, but others are not.

The W3C Proposal

At the time of going to press, a proposal has been submitted to the World Wide Web Consortium (W3C) for a speech synthesis mark-up language.[3] This proposal

[3] http://www.w3.org/TR/speech-synthesis

is intended to supersede several previous proposed standards for speech synthesis mark-up, including Sun Microsystems JSML, Microsoft Speech API and Apple Speech Manager, as well as the independent standards SSML and SABLE (Sproat *et al.*, 1998). The W3C proposal has taken elements from all of these previous proposals and combined them in an attempt to produce a universal standard. While the W3C proposal is (currently) only a draft proposal, and may change considerably, it is the most detailed proposal to date for speech synthesis mark-up. The detailed examples below are therefore based on the W3C draft (October 2000 version), although the general approach and consequent problems apply to most of the previous standards mentioned above. The main objectives of these proposals are:

- to provide consistent output across different synthesisers and platforms;
- to support a wide range of applications of speech synthesis;
- to handle a large number of languages both within and across documents;
- to allow both manual annotation and automatic generation;
- to be implementable with current technology, and compatible with other standards.

These objectives are addressed by a set of mark-up tags which are compatible with most international mark-up standards. The tag set provides markers for structural items (paragraphs and sentences), pronunciation (in the IPA phonetic alphabet), a range of special cases such as numbers and addresses, changes in the voice or the language to be used, synchronisation with audio files or other processes, and – most importantly – a bewildering array of prosodic features. This is typical of recent speech synthesis mark-up schemes in both the types of mark-up which they provide and the problems which they present for users and implementers of the scheme. Here we will concentrate on prosodic aspects.

The prosodic tags provided in recent proposals include the following:

- emphasis – level of prosodic prominence (4 values)
- break – level of prosodic break between words (4 values), with optional specification of pause duration
- prosody – six different prosodic attributes: three for F0, two for duration, and one for amplitude
- pitch – the baseline value for F0
- contour – the pitch targets (specified by time and frequency values)
- range – the limits of F0 variation
- rate – speech rate in words per minute
- duration – the total duration of a portion of text
- volume – amplitude on a scale of 0–100

Most of the prosody attributes take absolute values (in the appropriate units, such as Hertz or seconds), relative values (plus or minus, and percentages), and qualitative descriptions (highest/lowest, fastest/slowest, medium/default). Users can decide which of these to use, or even combine them all. Such schemes thus offer both high-level and low-level control.

Problems with current proposals

Not surprisingly, in their efforts to provide an ideal mark-up language such proposals run foul of the problems mentioned above, particularly the confusion of high-level and low-level control and the impossibility of device-independent mark-up. The following paragraphs examine the major difficulties from a prosodic viewpoint.

The emphasis and break tags are essentially high-level markers, and the prosody attributes are a mixture of high- and low-level markers, but there is nothing to prevent the co-occurrence of high- and low-level markers on the same item. For example, in most European languages emphasis would normally be realised by F0 excursions, but an emphasised item might also have a prosody contour marker specifying a totally flat F0 contour. In such a case, should the high-level marker or the low-level marker take precedence?

The potential for contradictory mark-up is a general problem with current proposals. Other examples of potentially contradictory mark-up include the rate and duration tags, the contour and pitch/range tags, and the interactions between the emphasis tag and all the prosody attributes. Although they attempt to fulfil the very different needs of, say, a researcher who is fine-tuning a duration algorithm and an application developer who simply wants a piece of text such as '*really slowly*' to be read out really slowly, no current proposal gives adequate indications of the appropriate usage of the tags. Moreover, they give absolutely no indication of how such contradictions should be resolved. What is probably required here is a clear distinction between a high-level interface (akin to a software API) and a low-level interface for use by researchers and system developers.

The issue of device independence is related to the distinction between high-level and low-level tags. Most synthesis systems have a model of emphasis, prosodic breaks and (less often) speech rate, since these are high-level prosodic phenomena which occur in all languages and are important for high quality output. (We will leave aside the number of levels of emphasis or prosodic break, and the inadequacy of 'words per minute' as a measure of speech rate, on the assumption that most systems could construct a reasonable mapping from the proposed values to their own representations.) Notions such as F0 baseline, pitch range and pitch targets, however, are far from universal. The Fujisaki model of F0 (Fujisaki, 2000), one of the most popular in current synthesisers (Mixdorff, Chapter 13, this volume; Monaghan, Chapter 9, this volume), would have great difficulty in realising arbitrary pitch targets, and the notions of range and baseline have no obvious meaning in systems which concatenate stored pitch contours instead of generating a new contour by rule (e.g. Malfrère *et al.*, 1998).

A third major problem with all mark-up proposals for speech synthesis, is the distinction which they make between mark-up and 'non-markup behavior' [*sic*]. The W3C proposal, like most others, assumes that in portions of the input which are not specified by mark-up the synthesis system will simply perform as normal. This may be a reasonable assumption for aspects such as word pronunciation or the choice of a language, but it certainly is not reasonable for prosodic aspects. Let us look at some examples.

The following example of the use of the emphasis tag is given in the W3C proposal (similar examples were given by SABLE and SSML):

That is a <emphasis level = "strong"> huge </emphasis> bank account!

Obviously this is intended to ensure a strong emphasis on *huge*, but what is the 'non-markup behavior' in the rest of the sentence? In British intonation terminology (Crystal, 1969), is *huge* the nucleus of the whole sentence, forcing *bank account* to be deaccented? If not, is *huge* the nucleus of a smaller intonational phrase, and if so, does that intonational phrase extend to the left or the right? If *huge* is not the nucleus of some domain, and the system correctly places a nuclear accent on *bank*, how should the clash between these two adjacent prominences be resolved? There are no obvious answers to these questions, and none are suggested.

Although the general aim of current mark-up schemes seems to be to complement the text processing abilities of current synthesis systems, by adding additional control and functionality, the break tag is a clear exception. It is defined as follows:

> The *break* element is an empty element that controls the pausing or other prosodic boundaries between words. The use of the *break* element between any pair of words is optional. If the element is not defined, the speech synthesiser is expected to automatically determine a break based on the linguistic context. In practice, the break element is most often used to override the typical automatic behavior of a speech synthesiser.

This definition gives the impression that a break may be specified in the input text in a place where the automatic processing inserts either too weak or too strong a boundary. The implicit assumption is that specifying a break has a purely local effect and that the rest of the output will be unaffected: this is unlikely to be the case. Many current systems generate prosodic breaks based on the number of syllables since the last break and/or the desire for breaks to be evenly spaced (e.g. Keller and Zellner, 1995): inserting a break manually will affect the placement of other boundaries, unless the mark-up is treated as a post-processing step. Even systems which use more abstract methods to determine breaks automatically will need to decide how mark-up influences these methods. Prosodic boundaries also interact with the placement of emphasis: in British intonational terminology again, should the insertion of a break trigger the placement of another nuclear emphasis? Is the difference between 'Shut up and work' and 'Shut up <break/> and work' simply the addition of a pause, or is it actually more like the difference between 'SHUT up and WORK' and 'SHUT UP . . . and WORK' (where capitalisation indicates emphasised words)? Which should it be? The only simple answer seems to be that either all or none of the breaks in an utterance must be specified in current mark-up schemes.

We are forced to conclude that current mark-up schemes achieve neither device-independence nor simple implementation. In addition, it is extremely likely that users' expectations of 'non-markup behavior' will vary, and that they will employ the mark-up in different ways accordingly. Perhaps we should consider alternatives.

Other mark-up languages

There are two classes of mark-up languages which are not specific to speech synthesis but which have great potential for synthesis applications. They are the class of XML-based languages and the mark-up schemes used in WP software. There is some overlap between these two classes, since HTML belongs to both.

XML is 'an emerging standard of textual mark-up which is well suited to efficient computer manipulation' (Huckvale, Chapter 30, this volume). It has been applied to all levels of text mark-up, from the highest levels of dialogue structure (Luz, 1999) to the lowest levels of phonetic control for speech synthesis (Huckvale, Chapter 30, this volume). There is increasing pressure for XML compatibility in telephony applications, and it has several advantages over other tagging schemes: it is extremely flexible, easy to generate and process automatically, and human-readable. XML mark-up has the disadvantage of being rather cumbersome for manual annotation, since its tags are lengthy and complex, but it is an ideal format for data exchange and would be a good choice for a universal synthesis mark-up language.

WP mark-up languages are a small and rapidly shrinking set of data formats including HTML, LaTeX, RTF and Microsoft WORD. Their accessibility ranges from the well-documented, human-readable LaTeX language to the top-secret inner workings of WORD. HTML is the XML-based text formatting language, and thus is a good example of both the XML and the WP styles of mark-up. Synthesis from HTML and LaTeX is relatively straightforward for simple documents, and is possible even for very complex items such as mathematical equations (Fitzpatrick, 1999; Monaghan (e), Webpage). WP mark-up has so far been largely ignored by the speech synthesis community, but it has many advantages. The set of tags is very stable (all packages provide the same core mark-up), the tags encode structure (sections, paragraphs, lists, etc.) and salience (bolding, font size, capitalisation), and their spoken realisation is therefore relatively unproblematic. Formatting can become too complex or ambiguous for speech synthesis in some cases (Monaghan (f), Webpage), but this is quite rare. WP mark-up has evolved over thirty years to a fairly mature state, and is a *de facto* standard which should be taken seriously by speech synthesis developers.

Conclusion

There is little doubt that mark-up is the way ahead for speech synthesis, both in the laboratory and in commercial applications. Mark-up has the potential to solve many of the hard problems for current speech synthesisers. It can provide additional control, and hence quality, in applications where prosodic and other information can be reliably added: examples include automatically generated and manually annotated input. It simplifies the treatment of specific text types, and allows access to the vast body of formatted text: examples include email readers and web browsers. It allows the testing of theories and the building of speech interfaces to advanced telephony applications: examples include the mark-up of focus domains or dialogue moves in spoken dialogue systems, and the possibility of speech output from DBQ and NLG systems.

So far, very little of this potential has been realised. It is important to remember that it is still early days for mark-up in speech synthesis, and that (although systems have been built and proposals have been made) it will be some time before we can reap the full benefits of a universal mark-up language. There are currently no good standards, for very good reasons: different applications have very different requirements, and synthesisers vary greatly in the control they allow over their output. Moreover, as mentioned above, people do not use mark-up consistently. Different users have different assumptions and preferences concerning, say, the use of underlining. The same user may use different mark-up to achieve the same effect on different occasions, or even use the same mark-up to achieve different effects. As an example, the use of capitalisation in the children's classic *The Cat in the Hat* has several different meanings in the space of a few pages: narrow focus, strong emphasis, excitement, trepidation and horror (Monaghan (f), Webpage).

For the moment, speech synthesis mark-up in a particular application is likely to be defined not by any agreed standard but rather by the intersection of the input provided by the application and the output parameters over which a particular synthesiser can offer some control. There is also likely to be a trade-off between the default speech quality of the synthesiser and the flexibility of control: concatenative systems are generally less controllable than systems which build the speech from scratch, and sophisticated prosodic algorithms are more likely than simple ones to be disturbed by an unexpected boundary marker. The increase in academic and commercial interest in mark-up for speech synthesis is a very good thing, and some progress towards a universal mark-up language has been made, but the issues of device-independence, low- and high-level control, and the non-local effects of mark-up, amongst others, are still unresolved. To borrow a few words from Breen (Chapter 37, this volume), 'a great deal of fundamental research is needed . . . before any hard and fast decisions can be made regarding a standard'.

Acknowledgements

This work was originally presented at a meeting of COST 258, a co-operative action funded by the European Commission. It has been revised to incorporate feedback from that meeting, and the author gratefully acknowledges the support of COST 258 and of his colleagues at the meeting.

References

Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge University Press.

Fitzpatrick, D. (1999). *Towards Accessible Technical Documents*. PhD thesis, Dublin City University.

Fitzpatrick, D. and Monaghan, A.I.C. (1998). TechRead: A system for deriving Braille and spoken output from LaTeX documents. *Proceedings of ICCHP '98* (pp. 316–323). IFIP World Computer Congress. Vienna/Budapest.

Fujisaki, H. (2000). The physiological and physical mechanisms for controlling the tonal features of speech in various languages. *Proceedings of Prosody 2000*. Krakow, Poland.

Keller, E. and Zellner, B. (1995). A statistical timing model for French. *13th International Congress of Phonetic Sciences*, Vol. 3 (pp. 302–305). Stockholm.

Luz, S. (1999) State-of-the-art survey of dialogue management tools, DISC deliverable 2.7a. *ESPRIT long-term research concerted action 24823.* Available: http://www.disc2.dk/publications/deliverables/

Malfrère, F., Dutoit, T., and Mertens, P. (1998). Automatic prosody generation using suprasegmental unit selection. *Proceedings of 3rd International Workshop on Speech Synthesis* (pp. 323–328). Jenolan Caves, Australia.

Monaghan, A.I.C. (1991). *Intonation in a Text-to-Speech Conversion System*. PhD thesis, University of Edinburgh.

Monaghan, A.I.C. (1992). Heuristic strategies for higher-level analysis of unrestricted text. In G. Bailly *et al*. (eds), *Talking Machines* (pp. 143–161). Elsevier.

Monaghan, A.I.C. (1993). What determines accentuation? *Journal of Pragmatics*, *19*, 559–584.

Monaghan, A.I.C. (1994). Intonation accent placement in a concept-to-dialogue system. *Proceedings of 2nd International Workshop on Speech Synthesis* (pp. 171–174). New York.

Monaghan, A.I.C. (1998). Des gestes écrits aux gestes parlés. In S. Santi *et al*. (eds), *Oralité et Gestualité* (pp. 185–189). L'Harmattan.

Monaghan, A.I.C. (a). Mark-Up for Speech Synthesis. Email.html. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm.

Monaghan, A.I.C. (b). Mark-Up for Speech Synthesis. Errors.html. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm.

Monaghan, A.I.C. (c). Mark-Up for Speech Synthesis. Html.html. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm.

Monaghan, A.I.C. (d). Mark-Up for Speech Synthesis. Phonebook.html. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm.

Monaghan, A.I.C. (e). Mark-Up for Speech Synthesis. Equations.html. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm.

Monaghan, A.I.C. (f). Mark-Up for Speech Synthesis. Formatting.html. Accompanying Webpage. Sound and multimedia files available at http://www.unil.ch/imm/cost258volume/cost258volume.htm.

Sluijter, A., Bosgoed, E., Kerkhoff, J., Meier, E., Rietveld, T., Sanderman, A., Swerts, M., and Terken, J. (1998). Evaluation of speech synthesis systems for Dutch in telecommunication applications. *Proceedings of 3rd International Workshop on Speech Synthesis* (pp. 213–218). Jenolan Caves, Australia.

Sonntag, G.P. (1999). *Evaluation von Prosodie*. Doctoral dissertation, University of Bann.

Sonntag, G.P., Portele, T., Haas, F. and Köhler, J. (1999). Comparative evaluation of six German TTS systems. *Proceedings of Eurospeech* (pp. 251–254). Budapest.

Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K. and Edgington, M. (1998). SABLE: A standard for TTS markup. *Proceedings of 3rd International Workshop on Speech Synthesis* (pp. 27–30). Jenolan Caves, Australia.

Wahlster, W. (ed.) (2000). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Verlag.

Wouters, J., Rundle, B. and Macon, M.W. (1999). Authoring tools for speech synthesis using the SABLE markup standard. *Proceedings of Eurospeech* (pp. 963–966). Budapest.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

32

Automatic Analysis of Prosody for Multi-lingual Speech Corpora

Daniel Hirst
Université de Provence, Lab. CNRS Parole et Langage
29 av. Schuman, 13621 Aix-en-Provence, France
daniel.hirst@lpl.univ-aix.fr

Introduction

It is generally agreed today that the single most important advance which is required to improve the quality and naturalness of synthetic speech is a move towards better understanding and control of prosody. This is true even for those languages which have been the object of considerable research (e.g. English, French, German, Japanese, etc.) – it is obviously still more true for the vast majority of the world's languages for which such research is either completely non-existent or is only at a fairly preliminary stage. For a survey of studies on the intonation of twenty languages see Hirst and Di Cristo (1998). Even for the most deeply studied languages there is still very little reliable and robust data available on the prosodic characteristics of dialectal and/or stylistic variability.

It seems inevitable that the demand for prosodic analysis of large speech databases from a great variety of languages and dialects as well as from different speech styles will increase exponentially over the next two or three decades, in particular with the increasing availability via the Internet of speech processing tools and data resources.

In this chapter I outline a general approach and describe a set of tools for the automatic analysis of multi-lingual speech corpora based on research carried out in the Laboratoire Parole et Langage in Aix-en-Provence. The tools have recently been evaluated for a number of European and non-European languages (Hirst *et al.*, 1993; Astésano *et al.*, 1997; Courtois *et al.*, 1997; Mora *et al.*, 1997).

The General Approach

The number of factors which contribute to the prosodic characteristics of a particular utterance are quite considerable. These may be universal, language-specific, dialectal, individual, syntactic, phonological, semantic, pragmatic, discursive, attitudinal, emotional . . . and this list is far from complete.

Many approaches to the study of prosody attempt to link such factors directly to the acoustic characteristics of utterances. The approach I outline here is rather different. Following Hirst, Di Cristo and Espesser (2000), I propose to distinguish four distinct levels of representation: the *physical* level, the *phonetic* level, the *surface phonological* level and the *underlying phonological* level. Each level of representation needs to be interpretable in terms of adjacent levels of representation.

The underlying phonological level is conceived of as the interface between the representation of phonological form and syntactic/semantic interpretation. Although it is this underlying phonological level which is ultimately conditioned by the different factors listed above, this level is highly theory-dependent and I shall not attempt to describe it any further here. Instead, my main concern in this chapter is to characterise the phonetic and surface phonological levels of representation for prosody.

I assume, following Trubetzkoy, a fundamental distinction between phonology (the domain of abstract qualitative distinctions) and phonetics/acoustics (the domain of quantitative distinctions). I further assume that phonetics is a level of analysis which mediates between the phonological level and the physical acoustic/physiological level of analysis. For more discussion see Hirst and Di Cristo (1998), Hirst, *et al*. (2000).

The aim of the research programme based on this approach is to develop automatic procedures defining a reversible mapping between acoustic data and phonetic representations on the one hand, and between phonetic representations and surface phonological (or at least 'quasi-phonological') representations on the other hand. This programme aims, consequently, (at least as a first step) not to *predict* the prosodic characteristics of utterances but rather to *reproduce* these characteristics in a robust way from appropriate representations.

The first step, *phonetic representation*, consists in reducing the acoustic data to a small set of quantitative values from which it is possible to reproduce the original data without significant loss of information. The second step, *surface phonological representation*, reduces the quantitative values to qualitative ones, where possible without losing significant information. In the rest of this chapter I describe some specific tools which have been developed in the application of this general research programme.

The prosodic characteristics of speech concern the three dimensions of time, frequency and intensity. Our research up until quite recently has concentrated on the representation of fundamental frequency although we are currently also working on the integration of durational and rhythmic factors into the representation.

Phonetic representation

Duration

A *phonetic* representation of duration is obtained simply by the alignment of a phonological label (phoneme, syllable, word, etc.) with the corresponding acoustic signal. To date such alignments have been carried out manually. This task is very labour-intensive and extremely error-prone. It has been estimated that it generally takes an experienced aligner more than fifteen hours to align phoneme labels for one minute of speech (nearly 1000 times real-time).

Software has been developed to carry out this task (or at least a first approximation) automatically (Dalsgaard *et al.*, 1991; Talkin and Wightman, 1994; Vorsterman, *et al.*, 1996). Such software, which generally uses the technique of Hidden Markov modelling, requires a large hand-labelled training corpus. Recent experiments, however (Di Cristo and Hirst, 1997; Malfrère and Dutoit, 1997), have shown that a fairly accurate alignment of phonemic labels can be obtained without prior training by using a diphone synthesis system such as Mbrola (Dutoit, 1997). Once the corpus to be labelled has been transcribed phonemically, a synthetic version can be generated with a fixed duration for each phoneme and with a constant F0 value. A dynamic time warping algorithm is then used to transfer the phoneme labels from the synthetic speech to the original signal.

Once the labels have been aligned with the speech signal, a second synthetic version can be generated using the duration defined by the aligned labels and the fundamental frequency of the original signal. This second version is then re-aligned with the original signal using the same dynamic time-warping algorithm. This process, which corrects a number of errors in the original alignment (Di Cristo and Hirst, 1997) can be repeated until no further improvement is made.

Fundamental Frequency

A number of different models have been used for modelling or stylising fundamental frequency curves. The MOMEL algorithm (Hirst and Espesser, 1993, Hirst *et al.*, 2000) factors the raw F0 curve into two components: a microprosodic component corresponding to short-term variations of F0 conditioned by the nature of individual phonemes, and a macroprosodic component which corresponds to the longer term variations, independent of the nature of the phonemes. The macroprosodic curves are modelled using a quadratic spline function.

The output of the MOMEL algorithm is a sequence of target points corresponding to the linguistically significant targets as seen in the lower panel of Figure 32.1. These target points can be used for close-copy resynthesis of the original utterance with practically no loss of prosodic information compared with the original F0 curve. It would be quite straightforward to model the microprosodic component as a simple function of the type of phonematic segment, essentially as unvoiced consonant, voiced consonant, sonorant or vowel (see Di Cristo and Hirst, 1986), and to add this back to the synthesised F0 curve, although this is not currently implemented in our system.

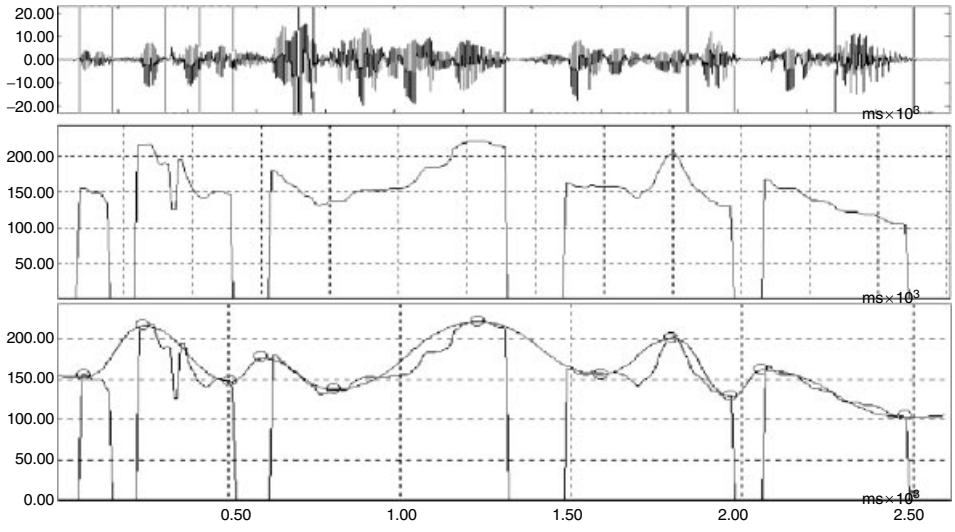


Figure 32.1 Waveform (top), F0 trace (middle) and quadratic spline stylisation (bottom) for the French sentence 'Il faut que je sois à Grenoble Samedi vers quinze heures' (I have to be in Grenoble on Saturday around 3 p.m.). The stylised curve is entirely defined by the target points, represented by the small circles in the bottom figure

Surface Phonological Representation

Duration

The duration of each individual phoneme, as measured from the phonetic representation, can be reduced to one of a finite number of distinctions. The value for each phoneme is calculated using the z-transform of the raw values for mean and standard deviation of the phoneme (Campbell 1992). Currently for French, we assume four phonologically relevant values of duration: normal, shortened, lengthened and extra-lengthened (see Di Cristo *et al*., 1997; Hirst, 1999). A lot more research is needed in this area.

Fundamental Frequency

The target points modelled by the MOMEL algorithm described above could be interpreted in a number of ways. It has been shown, for example (Mixdorff, 1999), that the quadratic spline stylisation provides a good first step for the automatic estimation of parameters for Fujisaki's superpositional model of F0 contours (Fujisaki, 2000) and that this can then be used for the automatic recognition and characterisation of ToBI labels (Mixdorff and Fujisaki, 2000; Mixdorff, Chapter 13, this volume).

A different interpretation is to reduce the target points to quasi-phonological symbols using the INTSINT transcription system described in Hirst and Di Cristo (1998) and Hirst, *et al*. (2000). This system represents target points as values either

globally defined relative to the speaker's pitch range (Top (**T**), Mid (**M**) and Bottom (**B**)) or locally defined relative to the previous target-point. Relative target-points can be classified as Higher (**H**), Same (**S**) or Lower (**L**) with respect to the previous target. A further category consists of smaller pitch changes which are either slightly Upstepped (**U**) or Downstepped (**D**) with respect to the previous target.

Two versions of a text-to-speech system for French have been developed, one stochastic (Courtois *et al.*, 1997) and one rule-based (Di Cristo *et al.*, 1997) implementing these phonetic and surface phonological representations. The software for deriving both the phonetic stylisation as a sequence of target points and the quasi-phonological coding with the INTSINT system is currently being integrated into a general-purpose prosody editor ProZed (Hirst 2000a).

Evaluation and Improvements

Evaluation of the phonetic level of representation of F0 for English, French, German, Italian, Spanish and Swedish has been carried out on the EUROM1 corpus (Chan *et al.*, 1995) within the MULTEXT project (Véronis *et al.* 1994; Astésano *et al.*, 1997). The results show that the algorithm is quite robust. On the English and French recordings (about one and a half hours of speech in total) around 5% of the target points needed manual correction. The majority of these corrections involved systematic errors (in particular before pauses) which improvement of the algorithm should eliminate.

Evaluation of the surface phonological representation has also been undertaken (Campione *et al.*, 1997). Results for the French and Italian versions of the EUROM1 corpus show that while the algorithm described in Hirst *et al.* (2000) seems to preserve most of the original linguistic information, it does not provide a very close copy of the original data and it also contains many undesirable degrees of freedom. A more highly constrained version of the algorithm (Hirst, 2000a, 2000b) assumes that the relationship between the symbolic coding and the actual target points can be defined on a speaker-independent and perhaps even language-independent basis with only two speaker-dependent variables corresponding to the speaker's *key* (approximately his mean fundamental frequency) and his overall pitch *range*. For the passages analysed the values of key and range were optimised within the parameter space: key $=$ mean $\pm$ 20 Hz, range $\in$ [0.5–2.5] octaves. The mean optimal range parameter resulting from this analysis was not significantly different from 1.0 octave.

It remains to be seen how far this result is due to the nature of the EUROM1 corpus which was analysed (40 passages consisting each of 5 semantically connected sentences) and whether it can be generalised to other speech styles and other (particularly non-European) languages.

Figure 32.2 shows the sequence of target points derived from the raw F0 curve compared with those generated from the INTSINT coding. The actual MOMEL targets, INTSINT coding and derived targets for this 5-sentence passage are available in the accompanying example file (HI00E01.txt, Webpage).

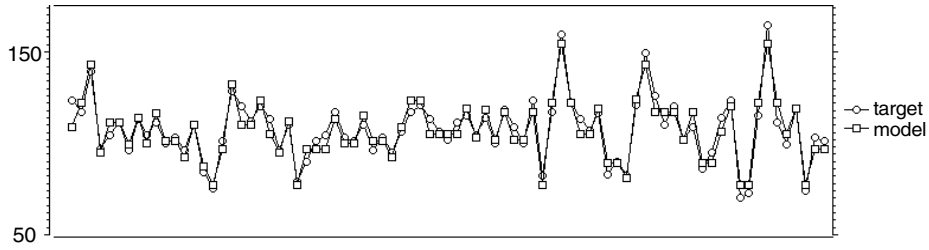


Figure 32.2 Target points output from the MOMEL algorithm and those generated by the optimised INTSINT coding algorithm using the two parameters key = 109 Hz, range = 1.0 octave for passage fao0 of the EUROM1 (English) corpus

Perspectives

The ProZed software described in this chapter will be made freely available for non-commercial research and will be interfaced with other currently available non-commercial speech processing software such as Praat (Boersma and Weenink 1995–2000) and Mbrola (Dutoit, 1997). Information on these and other developments will be made regularly available on the web page and mailing list of SProSIG, the Special Interest Group on Speech Prosody recently created within the framework of the International Speech Communication Association *ISCA*.[1] It is hoped that this will encourage the development of comparable speech databases, knowledge bases and research paradigms for a large number of languages and dialects and that this in turn will lead to a significant increase in our knowledge of the way in which prosodic characteristics vary across languages, dialects and speech styles.

Acknowledgements

The research reported here was carried out with the support of COST 258, and the author would like to thank the organisers and other members of this network for their encouragement and for many interesting and fruitful discussions during the COST meetings and workshops.

References

Astésano, C., Espesser, R., Hirst, D.J., and Llisterri, J. (1997). Stylisation automatique de la fréquence fondamentale: une évaluation multilingue. *Actes du 4e Congrès Français d'Acoustique* (pp. 441–443). Marseilles, France.

Boersma, P. and Weenink, D. (1995–2000). Praat: a system for doing phonetics by computer. htttp://www.fon.hum.uva.nl/praat/

Campbell, W.N. (1992). *Multi-level Timing in Speech*, PhD Thesis, University of Sussex.

[1] http://www.lpl.univ-aix.fr/projects/sprosig.

Campione, E., Flachaire, E., Hirst, D.J., and Véronis, J. (1997). Stylisation and symbolic coding of F0, a quantitative approach. *Proceedings ESCA Tutorial and Research Workshop on Intonation*. Athens. pp. 71–74.

Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinas, G., Kvale, L., Lamel, L., Lindberg, L., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., and Zeiliger, J. (1995). EUROM: A spoken language resource for the EU. *Proceedings of the 4th European Conference on Speech Communication and Speech Technology, Eurospeech '95*, Vol. I (pp. 867–880). Madrid.

Courtois, F., Di Cristo, Ph., Lagrue, B. and Véronis, J. (1997). Un modèle stochastique des contours intonatifs en français pour la synthèse à partir des textes. *Actes du 4ème Congrès Français d'Acoustique* (pp. 373–376). Marseilles.

Di Cristo, A. & Hirst, D.J. 1986. Modelling French micromelody: analysis and synthesis. Phonetica 43, 11–30.

Dalsgaard, P. Andersen, O., and Barry, W. (1991). Multi-lingual alignment using acoustic-phonetic features derived by neural-network technique. *Proceedings ICASSP-91* (pp. 197–200).

Di Cristo, A., Di Cristo, P., and Véronis, J. (1997). A metrical model of rhythm and intonation for French text-to-speech. *Proceedings ESCA Workshop on Intonation: Theory, Models and Applications*. Athens pp. 83–86.

Di Cristo, Ph. and Hirst, D.J. (1997). Un procédé d'alignement automatique de transcriptions phonétiques sans apprentissage préalable. *Actes du 4e Congrès Français d'Acoustique* (pp. 425–428). Marseilles.

Dutoit, T. (1997). *An Introduction to Text-to-Speech synthesis*. Kluwer Academic Press.

Fujisaki, H. (2000). The physiological and physical mechanisms for controlling the tonal features of speech in various languages. *Proceedings Prosody 2000: Speech Recognition and Synthesis*. Kraków, Poland.

Hirst, D.J. & Espesser, R. 1993. Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix 15, 71-85.

Hirst, D.J. (1999). The symbolic coding of segmental duration and tonal alignment: An extension to the INTSINT system. *Proceedings Eurospeech* (pp.). Budapest.

Hirst, D.J. (2000a). ProZed: A multilingual prosody editor for speech synthesis. *Proceedings IEE Colloquium on State-of-the-Art in Speech Synthesis* (pp.). London.

Hirst, D.J. (2000b). Optimising the INTSINT coding of F0 targets for multi-lingual speech synthesis. *Proceedings ISCA Workshop: Prosody 2000 Speech Recognition and Synthesis* (pp.). Kraków, Poland.

Hirst, D.J. and Di Cristo, A. (eds) (1998). A survey of intonation systems. In D.J. Hirst and A. Di Cristo (eds), *Intonation Systems: A Survey of Twenty Languages* (pp. 1–44). Cambridge University Press.

Hirst, D.J., Di Cristo, A., Le Besnerais, M., Najim, Z., Nicolas, P., and Roméas, P. (1993). Multi-lingual modelling of intonation patterns. *Proceedings ESCA Workshop on Prosody* (pp. 204–207). Lund, Sweden.

Hirst, D.J., Di Cristo, A., and Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. In M. Horne (ed.), *Prosody: Theory and Experiment* (pp.). Kluwer Academic Publishers.

Malfrère, F. and Dutoit, T. (1997). High quality speech synthesis for phonetic speech segmentation. *Proceedings of EuroSpeech 97*. Rhodes, Greece.

Mixdorff, H. (1999). A novel approach to the fully automatic extraction of Fujisaki model parameters. *ICASSP 1999*.

Mixdorff, H. and Fujisaki, H. (2000). Symbolic versus quantitative descriptions of F0 contours in German: Quantitative modelling can provide both. *Proceedings Prosody 2000: Speech Recognition and Synthesis*. Kraków, Poland.

Mora, E., Hirst, D., and Di Cristo, A. (1997). Intonation features as a form of dialectal distinction in Venezuelan Spanish. *Proceedings ESCA Workshop on Intonation: Theory, Models and Applications*. Athens.

Talkin, D. and Wightman, C. (1994). The aligner. *Proceedings ICASSP 1994*.

Véronis, J., Hirst, D.J., Espesser, R., and Ide, N. (1994). NL and speech in the MULTEXT project. *Proceedings AAAI '94 Workshop on Integration of Natural Language and Speech* (pp. 72–78).

Vorsterman, A., Martens, J.P., and Van Coile, B. (1996). Automatic segmentation and labelling of multi-lingual speech data. *Speech Communication*, *19*, 271–293.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

33

Automatic Speech Segmentation Based on Alignment with a Text-to-Speech System

Petr Horák
Institute of Radio Engineering and Electronics, AS CR
Chaberská 57, 182 51 Praha 8, Czech Republic
horak@ure.cas.cz

Introduction

Automatic phonetic speech segmentation, or the alignment of a known phonetic transcription to a speech signal, is an important tool for many fields of speech research. It can be used for the creation of prosodically labelled databases for research into natural prosody generation, for the automatic creation of new speech synthesis inventories, and for the generation of training data for speech recognisers. Most systems for automatic segmentation are based on a trained recognition system operating in 'forced alignment' mode, where the known transcription is used to contrain the recognition of the signal. Such recognition systems are typically trained on hidden Markov models of phoneme realisations. Such models are trained from many realisations of each phoneme in various phonetic contexts as spoken by many speakers.

An alternative strategy for automatic segmentation, of use when a recognition system is not available or when there is insufficient data to train one, is to use a text-to-speech system to generate a prototype realisation of the transcription and to align the synthetic signal with the real one. The idea of using speech synthesis for automatic segmentation is not new. Automatic segmentation for French is thoroughly described by Malfrère and Dutoit (1997a). The algorithm developed in this article is based on the idea of Malfrère and Dutoit (1997b) as modified by Strecha (1999) and by Tučková and Strecha (1999). Our aim in pursuing this approach was to generate a new prosodically labelled speech corpus for Czech.

Speech Synthesis

In this study, phonetically labelled synthetic speech was generated with the Epos speech synthesis system (Hanika and Horák, 1998 and 2000). In Epos, synthesis is based on the concatenation of 441 Czech and Slovak diphones and vowel bodies (Ptáček, *et al.*, 1992; Vích, 1995). The sampling frequency is 8 kHz. To aid alignment, each diphone was additionally labelled with the position of the phonetic segment boundary. This meant that the Epos system was able to generate synthetic signals labelled at phones, diphones, syllables, and intonational units from a text. The system is illustrated in Figure 33.1.

Segmentation

The segmentation algorithm operates on individual sentences, therefore both text and recording are first divided into sentence-sized chunks and labelled synthetic versions are generated for each chunk. The first step of the segmentation process is to generate parametric acoustic representations of the signals suitable for aligning equivalent events in the natural and synthetic versions.

The acoustic parameters used to characterize each speech frame fall into five sets. The first set of parameters defines the representation of the local speech spectral envelope – these are the cepstral coefficients c_i obtained from linear prediction analysis of the frame (Markel and Gray, 1976).

$$c_0 = \ln(\sqrt{\alpha}), \tag{1}$$

$$c_n = -a_n - \frac{1}{n}\sum_{k=1}^{n-1}(n-k)c_{n-k}a_k \text{ for } n > 0, \tag{2}$$

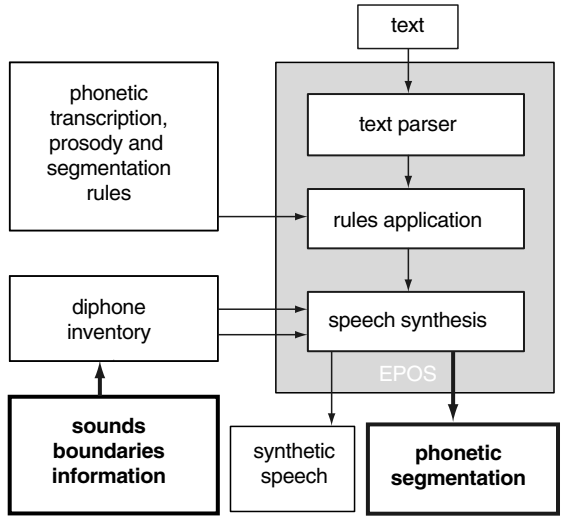


Figure 33.1 Epos speech synthesis system enhanced for segmentation
Note. The bold parts were added

where:

α ... linear prediction gain coefficient
$a_0 = 1$ and $a_k = 0$ for $k > M$
M ... order of linear prediction analysis.

The delta cepstral coefficients Δc_i form the second set of coefficients:

$$\Delta c_0(i) = c_0(i), \tag{3}$$

$$\Delta c_n(i) = c_n(i) - c_n(i-1), \tag{4}$$

where: $c_j(i)$ is j^{th} cepstral coefficient of i^{th} frame.

The third set of parameters is formed by the short time energy and its first difference (Rabiner and Schafer, 1978):

$$E(i) = \sum_{m=-\infty}^{\infty} (x(m)w(i \cdot N \cdot (1-\mu) - m))^2, \tag{5}$$

$$\Delta E(i) = E(i) - E(i-1), \tag{6}$$

where:

x ... speech signal
i ... frame number
N ... frame length
μ ... frame overlapping
$w(a) = \begin{cases} 1: & 0 \le a < N \\ 0: & \text{otherwise} \end{cases}$

Finally, the zero-crossing rate and the delta zero-crossing rate coefficients form the last set of parameters.

$$Z(i) = \sum_{m=-\infty}^{\infty} f(x(m)x(m-1))w(i \cdot N \cdot (1-\mu) - m), \tag{7}$$

$$\Delta Z(i) = Z(i) - Z(i-1), \tag{8}$$

where:

x ... speech signal
i ... frame number
N ... frame length
μ ... frame overlapping
$w(a) = \begin{cases} 1: & 0 \le \mathrm{a} < \mathrm{N} \\ 0: & \text{otherwise} \end{cases}$
$f(a) = \begin{cases} 1: & a < k_z \ (k_z < 0) \\ 0: & \text{otherwise.} \end{cases}$

All the parameters are normalized to the interval $\langle 0, 1 \rangle$. The block diagram of the phonetic segmentation process is illustrated in Figure 33.2.

The second step of the process is the segmentation itself. It is realized with a classical dynamic time warping algorithm with accumulated distance matrix **D**.

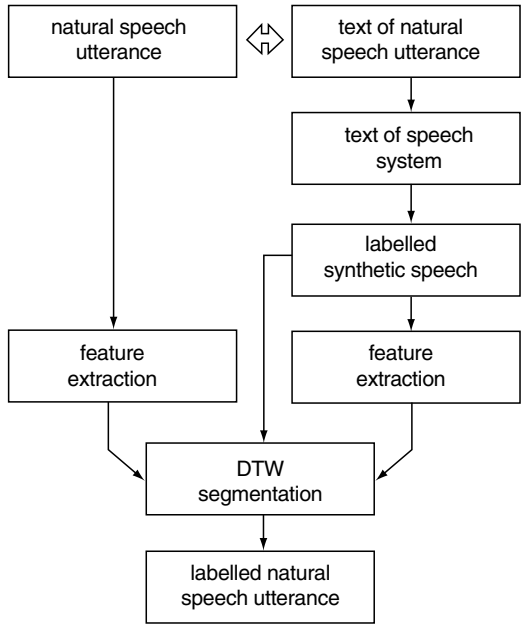


Figure 33.2 Phonetic segmentation process

$$\mathbf{D} = \begin{pmatrix} D(1,J) & D(2,J) & \text{L} & D(I,J) \\ D(1,J-1) & D(2,J-1) & \text{L} & D(I,J-1) \\ \text{M} & \text{M} & D(i,j) & \text{M} \\ D(1,2) & D(2,2) & \text{L} & D(I,2) \\ D(1,1) & D(2,1) & \text{L} & D(I,1) \end{pmatrix} \tag{9}$$

where:

$I \ldots$ number of frames of the first signal,
$J \ldots$ number of frames of the second signal.

This DTW algorithm uses symmetric form of warping function weighting coefficients (Sakoe and Chiba, 1978). The weighting coefficients are described in Figure 33.3.

In the beginning the marginal elements of the distance matrix are initialized (see equations 10–12). Other elements of the distance matrix are computed by equation 13.

$$D(1,1) = d(x(1), y(1)) \tag{10}$$

$$D(i,1) = D(i-1,1) + d(x(i), y(1)),\ i = 1 \ldots I \tag{11}$$

$$D(1,j) = D(1,j-1) + d(x(1), y(j)),\ j = 1 \ldots J \tag{12}$$

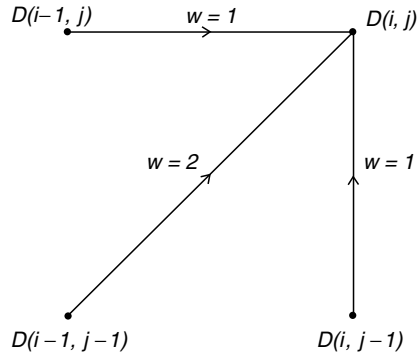


Figure 33.3 Weighting coefficients w for dynamic programming

$$D(i,j) = \begin{pmatrix} D(i-1,j) + d(x(i),y(j)) \\ D(i-1,j-1) + d(x(i),y(j)) \\ D(i,j-1) + d(x(i),y(j)) \end{pmatrix} \tag{13}$$

$$i = 1 \ldots I;\ j = 1 \ldots J$$

where: $d(x(i), y(j)) \ldots$ distance between the i^{th} frame of the first signal and the j^{th} of the second signal (see equation 14) and $MIN(*) \ldots$ minimum function.

The distance $d(x,y)$ is a weighted combination of a cepstral distance, an energy distance and a zero-crossing rate distance used to compare a frame from the natural speech signal x and a frame from the synthetic reference signal y.

$$\begin{aligned} d(x,y) = {} & \alpha_0 \sum_{i=0}^{n_{cep}} (c_i(x) - c_i(y))^2 + \beta \sum_{i=0}^{n_{cep}} (\Delta c_i(x) - \Delta c_i(y))^2 + \gamma (E(x) - E(y))^2 \\ & + \delta (\Delta E(x) - \Delta E(y))^2 + \varphi (Z(x) - Z(y))^2 + \eta (\Delta Z(x) - \Delta Z(y))^2 \end{aligned} \tag{14}$$

Values for the weights in equation (14) and other coefficients of the distance metric were found by an independent optimisation process leading to the following values:

- frames of 20 ms with a $n = 0.7$ (14 ms) overlap;
- linear predictive analysis order: $M = 8$;
- $\alpha = 1.5$; $\beta = 1.25$; $\gamma = 1.5$; $\delta = 1$; $\varphi = 1$; $\eta = 1.5$;
- zero-crossing rate constant $k_z = -20000$.

An example of accumulated distance matrix with minimum distance trajectory is shown in Figure 33.4. The next section shows the results of the first experiments performed with our segmentation system.

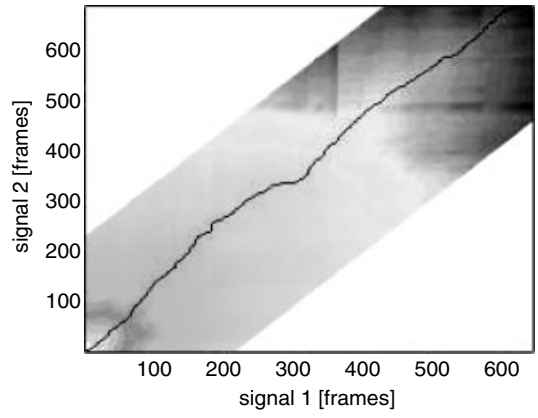


Figure 33.4 An example of a DTW algorithm accumulated distance matrix

Results

The system presented in the previous section was evaluated with one male and one female Czech native speaker. Each speaker pronounced 72 sentences, making a total of 3,994 phonemes per speaker. Automatic segmentation results were then compared with manual segmentation of the same data. Segmentation alignment errors were computed for the beginning of each phoneme and are analysed below under 10 phoneme classes:

vow – short and long vowels [a, ɛ, ɪ, ɔ, o, ʊ, a:, e:, i:, o:, u:]
exv – voiced plosives [b, d, ɟ, g]
exu – unvoiced plosives [p, t, c, k]
frv – voiced fricatives [v, z, ʒ, ɦ, ř]
fru – unvoiced fricatives [f, s, ʃ, x, r̝̊]
afv – voiced affricates [d͡z, d͡ʒ]
afu – unvoiced affricates [t͡s, t͡ʃ]
liq – liquids [r, l]
app – approximant [j]
nas – nasals [m, n, ŋ, ɲ]

Table 33.3 shows the percentage occurrences of each phoneme class.

Phoneme onset time errors as a function of the absolute value of their magnitude are given in Table 33.1 (male voice). Phoneme duration errors are presented in Table 33.2 (male voice). The same error data for the female voice are given in Tables 33.4 and 33.5. In most cases, segmentation results were superior for the female voice even though the male speech synthesis voice was used (Table 33.6).

As we can see from the tables, the average segmentation error for vowels is the smallest one among all the speech sound groups (see Figure 33.5). Very good results were also obtained for the class of unvoiced plosives. This is probably because the spectral patterning of these sounds is quite distinct, with a clear closure at the onset and with a release that often remains separate from the following speech sound. On the other hand, fricatives showed larger alignment errors. This

Table 33.1 Error rates (%) as a function of the segmentation magnitude error in ms for phoneme onsets for male voice

t(ms)	< 5	< 10	< 20	< 30	< 40	< 50	≥ 50
n_{all} (%)	19.3	37.6	64.5	79.4	86.3	95.2	4.8
n_{vow} (%)	19.1	35.5	61.2	76.0	83.0	95.0	5.0
n_{exv} (%)	21.5	45.0	75.7	87.4	91.2	95.9	4.1
n_{exu} (%)	21.2	42.2	73.5	86.7	93.1	97.9	2.1
n_{frv} (%)	20.9	41.4	68.9	83.6	91.4	95.9	4.1
n_{fru} (%)	11.5	26.2	54.4	74.1	87.2	96.7	3.3
n_{afv} (%)	0.0	10.0	40.0	70.0	70.0	100.0	0.0
n_{afu} (%)	26.1	51.1	75.0	88.0	92.4	100.0	0.0
n_{liq} (%)	18.5	37.6	66.5	85.3	91.2	98.1	1.9
n_{app} (%)	13.5	27.0	52.7	76.4	83.8	92.6	7.4
n_{nas} (%)	22.9	42.5	63.3	73.5	79.3	86.2	13.8

Table 33.2 Error rates (%) as a function of the segmentation magnitude error in ms for phoneme duration for male voice

t[ms]	< 5	< 10	< 20	< 30	< 40	< 50	≥ 50
n_{all} (%)	19.4	37.5	64.8	80.0	87.2	90.9	9.1
n_{vow} (%)	19.1	35.7	61.5	76.4	83.7	87.6	12.4
n_{exv} (%)	21.5	45.4	75.7	87.4	91.2	94.0	6.0
n_{exu} (%)	21.2	42.2	73.5	86.9	93.2	95.9	4.1
n_{frv} (%)	21.3	41.8	69.3	84.0	91.8	94.7	5.3
n_{fru} (%)	11.5	26.2	54.8	74.4	87.5	93.8	6.2
n_{afv} (%)	0.0	10.0	40.0	70.0	70.0	80.0	20.0
n_{afu} (%)	26.1	51.1	75.0	88.0	92.4	97.8	2.2
n_{liq} (%)	18.5	37.6	66.8	85.9	92.8	95.0	5.0
n_{app} (%)	14.2	27.7	53.4	77.7	85.8	89.2	10.8
n_{nas} (%)	23.2	43.4	64.9	76.5	83.4	87.6	12.4

may be because the initial and final parts of fricatives, as opposed to plosives, overlap with adjacent speech sounds, especially with vowels. The voiced affricates have the poorest alignment, however there were very few occurrences of these sounds in the corpus. Borders between nasals and other sonorants also showed larger than average alignment error.

The automatic segmentation algorithm seems to be robust to mistakes in transcription. In places where the natural speech utterance and synthetic speech utterance are not the same, the algorithm skips the unequal parts and continues to correctly align the other parts of the signal.

Applications

The main application of the Czech speech segmentation system is the creation of a prosodically labelled speech database to be used for further research on prosody

Table 33.3 Distribution of phoneme occurrences by phoneme class

phoneme class	Number of occurrences	Occurrence (%)
total	4066	100.0
short and long vowels	1736	42.7
voiced plosives	317	7.8
unvoiced plosives	533	13.1
voiced fricatives	244	6.0
unvoiced fricatives	305	7.5
voiced affricates	10	0.2
unvoiced affricates	92	2.3
liquids	319	7.8
approximant	148	3.6
nasals	362	8.9

Table 33.4 Error rates (%) as a function of the segmentation magnitude error in ms for phoneme onsets for female voice

t(ms)	< 5	< 10	< 20	< 30	< 40	< 50	≥ 50
n_{all} (%)	22.4	40.5	66.0	80.3	87.1	95.5	4.5
n_{vow} (%)	20.9	38.6	63.2	77.4	84.1	94.6	5.4
n_{exv} (%)	31.5	55.5	81.1	89.6	94.6	97.2	2.8
n_{exu} (%)	24.0	42.2	70.4	85.4	91.4	97.9	2.1
n_{frv} (%)	29.9	53.7	75.8	86.9	93.0	97.1	2.9
n_{fru} (%)	10.8	21.3	52.1	74.4	87.9	96.4	3.6
n_{afu} (%)	20.0	50.0	80.0	100.0	100.0	100.0	0.0
n_{afu} (%)	32.6	58.7	82.6	90.2	94.6	100.0	0.0
n_{liq} (%)	26.3	48.6	75.2	90.3	95.3	98.7	1.3
n_{app} (%)	23.6	37.8	58.1	76.4	82.4	93.2	6.8
n_{nas} (%)	17.7	30.4	55.2	69.1	76.2	89.2	10.8

modelling, especially for the training of neural nets for automatic pitch contour generation (Horák, *et al.* 1996; Tučková and Horák, 1997) and also for the analysis and synthesis of pitch contours performed in our lab (Horák, 1998). Research into Czech phoneme duration (motivated by Bartkova and Sorin, 1987) was started with the use of the segmentation system.

The speech segmentation tool has also been used for the transplantation of pitch contours between natural and synthetic utterances in order to evaluate our speech-coding algorithm. The block structure of our pitch transplantation tool is given in Figure 33.6.

Future applications for this approach to automatic segmentation will be to accelerate the creation of new voices for existing speech synthesizers on the basis of an existing voice (Portele *et al.*, 1996).

Table 33.5 Error rates (%) as a function of the segmentation magnitude error in ms for phoneme durations for female voice

t(ms)	< 5	< 10	< 20	< 30	< 40	< 50	≥ 50
n_{all} (%)	22.5	40.7	66.4	81.0	88.1	91.9	8.1
n_{vow} (%)	21.0	38.9	63.6	77.9	85.0	89.5	10.5
n_{exv} (%)	31.9	55.8	81.4	90.2	95.3	97.2	2.8
n_{exu} (%)	24.0	42.2	70.4	85.7	92.1	95.5	4.5
n_{frv} (%)	29.9	53.7	75.8	87.3	93.4	95.9	4.1
n_{fru} (%)	11.1	21.6	52.5	74.8	88.5	93.1	6.9
n_{afv} (%)	20.0	50.0	80.0	100.0	100.0	100.0	0.0
n_{afu} (%)	32.6	58.7	82.6	90.2	94.6	96.7	3.3
n_{liq} (%)	26.3	48.9	75.5	90.6	96.2	97.8	2.2
n_{app} (%)	23.6	38.5	59.5	79.1	85.1	89.9	10.1
n_{nas} (%)	17.7	30.4	56.1	71.0	79.3	84.3	15.7

Table 33.6 Average durations of phoneme classes for manual and automatic segmentation for both male and female speakers

Phoneme class	Male speaker		Female speaker	
	manu	auto	manu	auto
total	84.7	88.0	77.0	79.2
short and long vowels	83.3	82.9	81.6	81.3
voiced plosives	77.6	72.9	69.6	65.7
unvoiced plosives	88.4	87.5	78.4	76.9
voiced fricatives	78.2	95.8	66.5	72.7
unvoiced fricatives	113.7	129.8	90.4	105.2
voiced affricates	120.0	141.7	99.9	108.9
unvoiced affricates	126.0	127.3	113.5	115.1
liquids	59.3	63.2	48.4	50.3
approximant	74.0	77.2	61.8	65.6
nasals	87.5	101.1	76.8	88.2

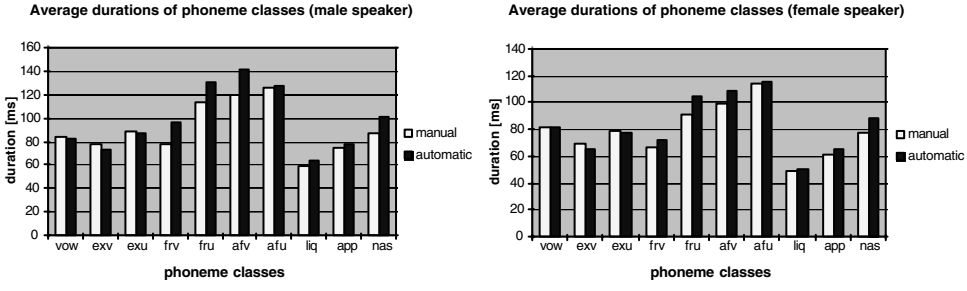


Figure 33.5 Average durations of phoneme classes for manual and automatic segmentation for both male and female speakers

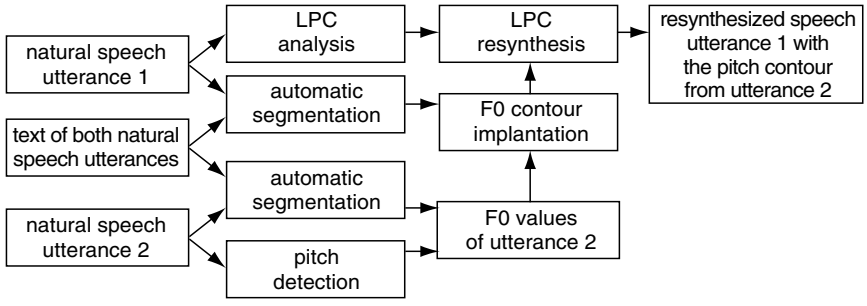


Figure 33.6 Transplantation of a pitch contour using automatic segmentation

Conclusion

The preliminary evaluation of the automatic segmentation algorithm shows that the accuracy of the automatic segmentation is sufficient for creating prosodically labelled speech corpora and for prosody transplantation, but it is not yet adequate for unit inventory creation. However, the automatic segmentation algorithm could be used for a new unit inventory creation, if supplemented with a manual or semiautomatic adjustment.

New speech corpora from several speakers have been recorded. We are working now on the manual labelling of these speech corpora for a better evaluation of the presented system. We plan to use the described automatic segmentation system for creation of a new 16 kHz diphone inventory which could be used for 16 kHz automatic segmentation algorithm. We also plan to extend the new diphone inventory by CC diphones and by Czech consonants missing in the current 8 kHz diphone inventory (ŋ, r̝) (Palková, 1994).

The Epos speech system is a free multilingual speech synthesis system (Horák and Hanika, 1998) which can be used for automatic segmentation of other languages (e.g. German). The Epos speech system can be freely downloaded from http://epos.ure.cas.cz/. We plan to make the automatic segmentation software a free addition to the system.

Acknowledgements

This work was supported by the grant No 102/96/K087 'Theory and Application of Speech Communication in Czech' of the Grant Agency of the Czech Republic and by the Czech Ministry of Education, Youth and Physical Training supply for the COST 258 project. Special thanks to Guntram Strecha from TU Dresden for his effort on automatic segmentation during his stay in our lab and to Betty Hesounová from our lab for a lot of manual work on segmentation and comparison.

References

Bartkova, K. and Sorin, C. (1987). A model of segmental duration for speech synthesis in French. *Speech Communication, 6*, 245–260.

Deroo, O., Malfrère, F. and Dutoit, T. (1998). Comparison of two different alignment systems: Speech synthesis vs. hybrid HMM/ANN. *Proceedings of the European Conference on Signal Processing (EUSIPCO '98)* (pp. 1161–1164). Rhodes, Greece.

Hanika, J. and Horák, P. (1998). Epos – A new approach to the speech synthesis. *Proceedings of the First Workshop on Text, Speech and Dialogue – TSD '98* (pp. 51–54). Brno, Czech Republic.

Hanika, J. and Horák, P. (2000). *The Epos Speech System: User Documentation ver. 2.4.43*. Available at *http://epos.ure.cas.cz/epos.html*.

Horák, P. (1998). The LPC analysis and synthesis of F0 contour. *Proceedings of the First Workshop on Text, Speech and Dialogue – TSD '98* (pp. 219–222). Brno, Czech Republic.

Horák, P. and Hanika, J. (1998). Design of a multilingual speech synthesis system. *Sprachkommunikation No. 152, 9. Konferenz Elektronische Sprachsignalverarbeitung* (pp. 127–128). Dresden, Germany.

Horák, P., Tučková, J. and Vích, R. (1996). New prosody modelling system for Czech text-to-speech. In D. Mehnert (ed.), *Studientexte zur Sprachkommunikation. No. 13. – Elektronische Sprachsignalverarbeitung* (pp. 102–107). Berlin.

Malfrère, F. and Dutoit, T. (1997a). Speech synthesis for text-to-speech alignment and prosodic feature extraction. *Proceedings of the ISCAS 97* (pp. 2637–2640). Hong Kong.

Malfrère, F. and Dutoit, T. (1997b). High-quality speech synthesis for phonetic speech segmentation. *Proceedings of the EuroSpeech '97* (pp. 2631–2634). Rhodes, Greece.

Markel, J.D. and Gray, A.H. Jr. (1976). *Linear Prediction of Speech*. Springer-Verlag.

Palková, Z. (1994). *The Phonetics and Phonology of the Czech Language*. Charles University, Prague (in Czech).

Portele, T., Stöber, K.-H., Meyer, H. and Hess, W. (1996). Generation of multiple synthesis inventories by a bootstrapping procedure. *Proceedings of ICSLP '96* (pp. 2392–2395). Philadelphia.

Ptáček, M., Vích, R. and Víchová, E. (1992). Czech text-to-speech synthesis by concatenation of parametric units. *Proceedings of URSI ISSSE '92* (pp. 230–232). Paris.

Rabiner, L.R. and Schafer, R.W. (1978). *Digital Processing of Speech Signals*. Bell Laboratories Inc.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Proc*., Vol. ASSP-26, 43–49.

Strecha, G. (1999). *Automatic Segmentation of Speech Signal*. Pre-diploma stay final report, IREE Academy of Sciences, Czech Republic (in German).

Tučková, J. and Horák, P. (1997). Fundamental frequency control in Czech text-to-speech synthesis. *Third Workshop on ECMS* (pp. 80–83). Université Paul Sabatier, Toulouse, France.

Tučková, J. and Strecha, G. (1999). Automatic labelling of natural speech by comparison with synthetic Speech. *Proceedings of the 4th International Workshop on Electronics, Control, Measurement and Signals ECMS '99* (pp. 156–159). Liberec, Czech Republic.

Vích, R. (1995). Pitch synchronous linear predictive Czech and Slovak text-to-speech synthesis. *Proceedings of the 15th International Congress on Acoustics ICA 95*, Vol. III (pp. 181–184). Trondheim, Norway.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

34

Using the COST 249 Reference Speech Recogniser for Automatic Speech Segmentation

Narada D. Warakagoda and Jon E. Natvig
Telenor Research and Development, Norway,
P. O. Box 83,
2027 Kjeller,
Norway
jon-email.natvig@telenor.com

Introduction

In the operation of a TTS system, the duration of a segment (usually a phoneme) is determined typically as a function of several linguistic factors such as phoneme identity, stress, phrasal position, surrounding phones, syllable length etc.. One popular methodology for duration modelling is the so called *data-driven* or *corpus-based* approach. In this kind of approach, duration rules are represented by some sort of function approximation device, such as a neural network or a CART (Classification and Regression Tree) and these devices are trained on an annotated database. Labelled and segmented speech databases therefore play a key role. If annotation (labelling and segmentation) needs to be carried out manually by phoneticians, this is a highly time-consuming and tedious task even for moderate size speech databases (Kvale, 1993). Therefore, research on the automation of these processes has attracted a considerable interest in the speech community. In this chapter we consider only the segmentation process, assuming that transcription has already been performed.

Because of its similarity to Automatic Speech Recognition (ASR), segmentation can be performed automatically using slightly modified speech recognisers. Hidden Markov Model (HMM) based approaches seem to be the most attractive approach, for the same reasons as those that caused their popularity in ASR.

In the experiments reported here the task was the segmentation of a relatively small speech database called PROSDATA (Natvig, 1998). This is a manually annotated, studio quality Norwegian database, specifically developed for the study of prosody. Training of a speech recogniser specifically for this task would require a much larger database with relevant speech material. In our case, the PROSDATA database was not expected to be sufficient for that purpose. The main objective of this work was therefore to investigate the possibility of adapting readily available recognisers like the COST 249 reference system to the segmentation task, based on a fairly limited speech material. For the experiments reported in this paper, the reference recogniser developed in connection with the European cooperation project COST 249 was used (Johansen *et al.*, 2000). This is an HMM based system implemented as a PERL script using the Hidden Markov Model Toolkit (HTK) (Young *et al.*, 1999). As the name implies, the original purpose of the system was for use as a calibration recognizer for a European multilingual database known as SpeechDat (Hoge *et al.*, 1999), collected via fixed and mobile telephone networks.

Description of the System

The COST 249 reference recogniser is a collection of recogniser structures, which includes both monophone, triphone and tied state triphone based architectures. Each recogniser-structure for Norwegian contains 42 HMMs representing 40 Norwegian SAMPA phonemes, silence and 'tee' models. All those models are standard three-state left-to-right HMMs, with the number of mixture components varying from 1 to 32 in the experiments (except the tee model). The system uses 39-dimensional feature vectors, each of which consists of 12 Mel Frequency based Cepstral Coefficients (MFCCs), the zeroth order cepstral coefficient and their delta and acceleration coefficients. Feature extraction in pre-training has been done with a Hamming window of length 25 ms which by 10 ms for each frame. For this work, a pre-trained version of the recogniser on the Norwegian SpeechDat database is used. Data used for pre-training has been collected via a fixed telephone network, from 1016 speakers and recorded with a sampling rate of 8 kHz (Johansen and Amdal, 1997).

The database used for segmentation experiments, PROSDATA, is a collection of 503 Norwegian sentences read by a female speaker. These data have been recorded in a studio, using a sampling rate of 16 kHz. The database contains a total of 39,328 segments (excluding silence). The phoneme inventory of PROSDATA consists of 54 SAMPA symbols and 3 symbols representing the silence (sil, fp = filled pause and pust = breath). They are reduced to the 42 phoneme set used in the COST 249 recogniser, by the mapping procedure shown in Table 34.1. The COST 249 recogniser does not have models for the diphthongs /ɑi/, /ɔy/, /ɑʉ/ and /øy/, so these were modelled as a sequence of phonemes as indicated in Table 34.1. In order to segment a given sentence, the recogniser is run in the *forced alignment mode*. In this mode we simply run the Viterbi algorithm on the trellis formed by concatenating the HMMs corresponding to the phonemes embedded in the sentence. A transition point between any two HMMs in the trellis is thus a segment boundary.

Table 34.1 Editing PROSDATA phonemes

SAMPA phoneme	IPA symbol	HTK phoneme(s)	SAMPA phoneme	IPA symbol	HTK phoneme(s)
2	ø	eu	2:	øː	eu:
2y	øy	eu j	=l	ɻ	l
=m	m̩	m	=n	n̩	n
=r	r̩	r	=rn	ɳ̩	rn
?	ʔ	sp	@	ə	e
A	ɑ	A	A:	ɑː	A:
Ai	ɑi	Aj	A}	ɑʉ	{ uh
C	ç	C	N	ŋ	N
O	ɔ	O	O:	ɔː	O:
Oy	ɔy	Oj	S	ʃ	S
b	b	b	d	d	d
e	e	e	e:	eː	e:
ep	-	sp	f	f	f
fp	-		g	g	g
h	h	h	i	i	i
i:	iː	i:	j	j	j
k	k	k	l	ɪ	l
m	m	m	n	n	n
p	p	p			
rd	ɖ	rd	rl	ɭ	l
rn	ɳ	rn	rt	t	rt
s	s	s	s	-	s
t	t	t	u	u	u
u:	uː	u:	v	v	v
y	y	y	y:	yː	y:
{	æ	ae	{:	æː	ae:
{i	æi	aei	}	ʉ	uh
}:	ʉː	uh:			

Even though the COST 249 reference recogniser has already been trained, it has been done on a database with (obviously) different statistical properties. Therefore *retraining* it on the PROSDATA database can improve the segmentation performance. This kind of retraining is simply achieved by running a few Baum-Welch iterations on the available models, with the help of HERest tool in HTK (Young *et al*., 1999). *Adaptation* can be considered an alternative to retraining, where the system parameters are tuned to a new, typically small data set. Note that in a deeper sense, retraining and adaptation refer to the same thing, namely transforming system parameters using a new dataset. However, their technical and algorithmic details are different and hence they can lead to different results. For our task, an off-line, supervised adaptation strategy involving both MAP (Maximum A Posteriori) and MLLR (Maximum Likelihood Linear Regression) approaches, is employed using the HADapt tool in HTK(Young *et al*., 1999).

Experiments and Results

Four main classes of experiments are performed, as shown below:

1. Direct application of the recogniser on PROSDATA.
2. Adjusting the feature extraction stage parameters.
3. Adaptation of the recogniser to PROSDATA.
4. Retraining the recogniser on PROSDATA.

Results of these experiment-classes are shown separately in the following subsections. In all these experiments, the quality of the automatic segmentation can evaluated by comparing the automatically obtained phoneme boundaries with the manually marked boundaries already available in PROSDATA. The difference between Δ_i between ith automatically marked boundary and the corresponding manually obtained boundary, is computed for all i and given a threshold p an accuracy measure can be defined as,

$$A(p) = \sum_i \frac{\varepsilon_i}{N} \times 100 \tag{1}$$

where N is the total number of boundaries, and

$$\varepsilon_i = \begin{cases} 1, & \Delta_i \leq p \\ 0, & otherwise \end{cases} \tag{2}$$

In the subsequent experiments, we report the results only for $p = 20$ ms, since this is somewhat standard in expressing segmentation accuracy in the literature.

Direct Use of the Recogniser

The mismatch between training data (SpeechDat) and test data (PROSDATA) is very clear and hence poor performances are expected. Consequently this class of experiments is performed mainly for its academic interest. The results are shown in Table 34.2, where the quality of segmentation is expressed using the accuracy measure as defined in equation 1. Different recogniser-structures are named according to the following convention,

$$(phonemodel)_(\#m)_(\#i),$$

where (*phonemodel*) can be *mono*, *tri* or *tied* representing monophone, untied triphone or tied triphone models respectively, (#*m*) represents the number of Gaussian mixture components in each state of the model and (#*i*) gives the number of Baum-Welch iterations run in the pre-training phase. We see the following two main patterns in these results:

1. Monophone models perform slightly better than pure/tied triphone models.
2. In low-parameter systems, for example monophone models with one or two mixture components, more iterations in pre-training leads to better segmentation performance. For high parameter systems the converse seems to be true.

Table 34.2 Segmentation results for direct use of the COST 249 recogniser on PROSDATA

System	Accuracy A(p)			
	p = 10 ms	p = 20 ms	p = 30 ms	p = 50 ms
Mono_1_4	19.37	52.0	73.20	90.09
Mono_1_5	20.92	55.4	76.02	91.66
Mono_1_6	22.24	58.86	76.17	91.36
Mono_1_7	23.56	58.35	76.98	91.43
Mono_2_1	24.39	59.17	77.31	91.17
Mono_2_2	24.96	60.03	77.74	91.07
Mono_4_1	24.43	60.17	77.92	91.24
Mono_4_2	24.01	59.35	77.28	91.08
Mono_8_1	23.79	58.40	76.57	90.52
Mono_8_2	23.48	57.60	75.95	90.13
Mono_16_1	23.11	56.76	75.75	89.89
Mono_16_2	22.71	56.77	75.86	89.92
Mono_32_1	22.52	56.85	76.13	90.32
Mono_32_2	22.03	56.21	76.04	90.82
tri_1_2	22.83	55.62	73.17	86.01
tied_1_1	22.94	56.87	75.84	89.43
tied_1_2	22.74	56.60	75.14	88.55
tied_2_1	22.73	56.16	74.34	87.84
tied_2_2	22.78	56.52	74.27	87.80

These patterns can be explained by taking the mismatch between SpeechDat and PROSDATA into account. The larger the number of model parameters, the higher the mismatch will be, and hence the inferiority of segmentation results. Since there are many more triphone models than monophone models, a system with triphone models contains a larger number of parameters and hence exhibits a higher mismatch. This explains why monophone models do better. Further, a system with a higher number of parameters becomes more SpeechDat-specific as more training is done on it. Hence the mismatch can increase, resulting in poor results on PROSDATA. However, if the number of parameters is low, more training can learn more general patterns in the data, and hence increasing the segmentation quality.

Adjusting the Feature Extraction Stage Parameters

Before employing more complex adaptation procedures we look for simple procedures which can reduce the mismatch between the recogniser and PROSDATA. One obvious difference is the sampling frequency of the data used to train the recogniser (8 kHz) and that of PROSDATA (16 kHz). The mismatches can be reduced by down-sampling PROSDATA to 8 kHz. A set of experiments was carried out with input data of reduced sampling frequency using some selected models. Results of those experiments are shown in column 2 of Table 34.3.

Table 34.3 Segmentation results for use of COST 249 recogniser on PROSDATA with different signal processing and model adjustment procedures

System	% Accuracy A(20 ms)			
	Down-sampled	Frame rate = 5 ms	Adaptation	Re-estimation
mono_1_7	65.91	74.75	80.82	82.75
mono_2_1	66.78	74.89	79.36	81.12
mono_2_2	67.51	74.77	79.56	81.08
mono_4_1	68.38	74.14	79.36	80.36
mono_4_2	68.29	73.42	78.80	79.92
mono_8_1	68.23	72.52	78.50	80.33
mono_8_2	67.73	71.92	78.46	80.38
tri_1_1	67.10	72.18	79.12	77.17
tri_1_2	67.23	71.72	78.92	76.94
tied_1_1	65.12	72.76	80.49	82.81
tied_1_2	64.73	71.98	80.51	82.70
tied_2_1	64.52	71.81	79.40	80.79
tied_2_2	65.02	71.79	79.26	80.92

From these results it is clear that mismatches have been reduced and hence the segmentation accuracies have been improved, since we now use a sampling frequency equal to that used in training. However, we still see that monophone models perform better than triphone models.

Since equal sampling rates (in segmentation and training) give better results in segmentation, in all the subsequent experiments we use the down-sampled (8 kHz) version of PROSDATA.

Another factor which influences the segmentation performances is the window size and frame rate used in parameter (MFCC) extraction. In speech recognition, a window size of 25 ms and a frame rate corresponding to 10 ms are often used, and as mentioned earlier, the COST recogniser also uses these values. But in segmentation, we would like to detect the variations in spectral properties with as high resolution as possible. Therefore a higher frame rate (i.e. a lower shift) can function favourably in our case. To investigate this hypothesis, a frame rate corresponding to a shift of 5 ms (instead of the original 10 ms shift) was used in feature extraction and the above experiment was rerun with this modification. Results are shown in column 3 of Table 34.3.

These results, when compared to those in column 2 of Table 34.3, clearly show the advantage of higher frame rate in feature extraction. Encouraged by these improvements, we use the frame rate corresponding to 5 ms, in all of the subsequent experiments.

Adaptation of the Recogniser to PROSDATA

In this set of experiments, each HMM-system under consideration is adapted to the PROSDATA speech data, in an off-line fashion, as follows.

- A global MLLR adaptation was performed to obtain an intermediate model.
- The intermediate model is further adapted using MAP strategy, with a scaling factor 12 to obtain the final adapted model (Young *et al.*, 1999).

Note, however, that only mean vectors of the HMMs are adapted. The final adapted model was then used for segmentation as usual. Results are shown in column 4 of Table 34.3. Comparing these results to those in column 3 of Table 34.3, we observe the advantage of adaptation. For almost all systems, adaptation leads to an improvement of about 6% in segmentation accuracy.

Re-estimation of the Recogniser Parameters on PROSDATA

As an alternative to adaptation, we can consider the re-estimation of model parameters on PROSDATA, starting from the ready-made models provided by the recogniser. In column 5 of Table 34.3, a set of results are shown, where the segmentation has been carried out with a set of models which have been re-estimated using four Baum-Welch iterations.

Comparing columns 5 and 4 in Table 34.3, we can conclude that re-estimation can offer better segmentation accuracies than adaptation. However these improvements are small, and for some systems we get even inferior results (e.g. tri_1_1 and tri_1_2). The inferiority of re-estimation (to adaptation) in these cases can be explained by considering the number of parameters of the system. If the number of parameters is very high, as in pure triphone systems, then they are difficult to re-estimate reliably if the available dataset is small. However, adaptation techniques handle the situation better, because they are designed for that purpose.

A final experiment is performed on the best systems so far, namely mono_1_7 and tied_1_1, where both adaptation and re-estimation were utilized. That is, the original model is adapted to PROSDATA first as described above and the resulting model is used for re-estimation as in the previous section. The results are shown in Table 34.4, and from that we see such an approach can further improve the segmentation quality slightly.

Segmentation Performance

The system which gives the best segmentation accuracy, mono_1_7 after adaptation and re-estimation, was selected to analyse the segmentation boundaries and durations for different phonemes. First, we observe that the segmentation accuracy

Table 34.4 Segmentation results for use of the COST 249 recogniser after adaptation and re-estimation of the recogniser parameters on PROSDATA

System	% Accuracy A(20 ms)
mono_1_7	84.41
tied_1_1	83.41

Table 34.5 Analysis of duration accuracy for the system mono_1_7, where the entries are based on durations not the boundary positions

% Accuracy $A(p)$			
$p = 20$ ms	$p = 25$ ms	$p = 30$ ms	$p = 35$ ms
74.8	82.6	87.2	90.5

obtained with this method is close to the performance reported in Kvale (1993). Since we have duration modelling in mind, we compared the *duration* of phoneme segments obtained automatically with our method to the manual values. In Table 34.5, we show the accuracy measure defined by equations (1) and (2) computed for duration errors for some values of the threshold p.

In a more detailed analysis, the average durations for each phoneme for the manual and automatic procedures were comparatively studied. A quantity of interest in this study was the relative error defined as the percentage of the duration error with respect to the average manual duration, where duration error is the difference between automatically and manually obtained average durations. We observed that, typically, relative errors are in the order of 20%.

Another quantity we considered in this study was the number of gross errors. A *gross error* is detected when the automatically labelled segment has no overlap with the corresponding manual segment. The total number of gross errors with respect to the total number of phonemes in the database amounts to 1.11% which is comparable to results reported in Kvale (1993). The seven phonemes /ə/, /h/, /r/, /n/, /t/. /l/ and /v/ are responsible for 75% of the gross errors. For the remaining 39 phonemes, the average gross error is 0.27%.

Conclusion

The COST 249 reference recogniser trained on the Norwegian SpeechDat database performs reasonably well as a segmentation device on PROSDATA database. However in its raw form, the recogniser has a limited value as a segmentation apparatus, as far as PROSDATA is concerned. Simple modifications like equalizing the sampling frequencies and increasing the frame rate improves performance significantly. Considerable further improvements is achieved through adaptation and/or re-estimation by running few Baum-Welch iterations on the recogniser. This further optimised version of the COST 249 recogniser leads to a segmentation, where the durations of the phonemic segments matches reasonably well with the manually obtained durations. To achieve further improvements we intend to apply intelligent post-processing rules which implement phonetic knowledge about the specific transitions at hand.

An important conclusion is that systems with a smaller number of parameters perform better. Thus the single mixture monophone system became the best system. Pure and tied triphone-based systems generally gave inferior results when compared with the corresponding monophone systems. In addition to the higher

number of parameters, another problem which weakened the triphone-based systems is that not all the triphones of the testdata (PROSDATA) were available in the COST 249 recogniser. This was remedied by substituting each unavailable triphone model with the corresponding monophone model. A better remedy might be to use a regression tree to synthesise the unseen triphones (Young *et al.*, 1999).

References

Hoge, H., Draxler, C., Heuvel, V.D., Johansen, F.T., Sanders, E., and Tropf, H.S. (1999). SpeechDat multilingual speech databases for teleservices: Across the finish line. *European Conference on Speech Communication and Technology*, Vol. 6 (pp. 2699–2702). Budapest, Hungary.

Johansen, F.T., and Amdal, I. (1997). SPEECHDAT-Norwegian speech database for the fixed telephone network. *Technical Report No. N 5/98*. Kjeller, Norway: Telenor Research and Development.

Johansen, F.T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kacic, Z., Zgank, A., Elenius, K., and Salvi, G. (2000). The COST 249 SpeechDat multilingual reference recognizer. Paper presented at *Language Resources and Evaluation Conference*. Athens, Greece.

Kvale, K. (1993). *Segmentation and labelling of speech*. Doctoral thesis, Norwegian Institute of Technology, Trondheim, Norway.

Natvig, J.E. (1998). A Speech database for study of Norwegian prosody. *Technical Report No. N 56/98*. Kjeller, Norway: Telenor Research and Development.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchew, V., and Woodland, P. (1999). *HTK book* (version 2.2). Entropic Research Ltd.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Part V

Future Challenges

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

35

Future Challenges

Eric Keller
Laboratoire d'analyse informatique de la parole (LAIP)
IMM-Lettres, University of Lausanne, 1015 Lausanne, Switzerland
Eric.Keller@imm.unil.ch

The preceding chapters have documented some of the key elements that have contributed to the improvement of naturalness in speech synthesis in the recent past. It is now the moment to move forward and to consider where this field will take us next.

Anyone who has worked with speech synthesis for any length of time will agree that the ancient dream of making machines speak like humans will be a reality soon enough. To a limited degree, it is a reality even now. In our laboratory, we've had the same synthesised recording on our answering machine for about three years now, and we have received quite a number of messages whose content suggests that the callers think that the recording was performed by a human speaker, not a synthesiser. In a context where synthesised messages are still rare, where the telephone transmission quality and the mechanical rhythm of the typical answering machine message camouflage the weaknesses of the synthesiser, and where the listener doesn't really pay very much attention to the message, a synthesiser can now temporarily 'pass for a human', i.e. pass the Turing test. So the question is thus no longer whether the speaking computer can pass the Turing test or not, but rather, for how long and under which circumstances.

An interesting distinction can be made in this context between 'speaking' and 'talking'. Speaking is more formal than talking. In the past, we have by and large taught computers to *speak*, not to talk. This makes speech synthesisers awkward in less formal contexts, such as man-machine interactions or virtual acting roles. It is interesting to consider that the film industry has recently begun producing full-length digitally-created movies, but that most *voices* in such movies are still human. Work on expressive voices and the implementation of subjectivity in synthesis is a high priority, and will certainly be among the most rewarding new directions of the speech synthesis domain. In that sense, Geneviève Caelen-Haumont's contribution in the next part extends some of the notions previously discussed in this volume in the Styles of Speech section by pointing out both the conceptual roots of subjectivity in speech, and its realisation in prosodic space.

Another area of further evolution of the field concerns the integration of speech synthesis into the wider contexts of multimodal systems and virtual realities.

Andrew Breen's contribution shows how TTS-systems are beginning to be integrated into much larger communication systems and interfaces. In these contexts, speech synthesis devices (or 'services') can be fed a much richer set of inputs than plain text. These machine-generated semantic contexts will let TTS systems derive automatically much richer inflectional nuances than is generally possible now.

Yet another powerful motor for speech synthesis in years to come will be virtual humanoids, functioning in various forms of virtual space. Virtual guides that can be made to say the same thing at all kiosks throughout a city are a contemporary example, but virtual autonomous teachers offering help when needed in a project performed in collaborative multi-user virtual space could well be a much richer example of future applications. Past research has shown that realistic integration between a virtual humanoid's orofacial and vocal activities is crucial to good understanding of speech, but is particularly difficult to achieve. The contribution by Beskow and colleagues demonstrates some of the key synergetic mechanisms that must be implemented in this context.

As these examples illustrate, the field of speech synthesis has clearly become a fairly extensive enterprise during the past twenty years. Gone are the days of Klatt and Fant where a single person built a system from A to Z. Nowadays, teams of researchers have various intelligent information systems communicate with each other via interfaces designed to be increasingly interchangeable and transparent in use. Gudrun Flach's contribution to this volume illustrates some of the parameters that are currently employed for speech synthesis interfaces in commercially viable systems. These listings are useful indicators of the considerable range of speech synthesis services that are available now, or will soon be available in most of the world's major languages.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

36

Towards Naturalness, or the Challenge of Subjectiveness

Geneviève Caelen-Haumont
Lab. CNRS Parole et Langage, Université de Provence
29 av. Schuman, 13621 Aix en Provence, France, Genevieve.
Caelen@lpl.univ-aix.fr

Introduction

It is now well accepted that linguistic structures cannot completely account for the full variation that one observes in speech. This variation is nevertheless an essential component of communication. When various fragments of spontaneous speech (and to some extent 'intelligent' reading as well) are submitted to analysis, a set of characteristics may be found, in short, variability, adaptation to communication context and addressees, and ultimately, subjectively identified speech characteristics at every level, from acoustics to semantics. Therefore, in order to produce more natural synthetic speech, it is necessary to model this variability.

Based on experience in the analysis of reading and spontaneous speech, the grounding hypotheses of this work are: (1) the speaker needs to make the message known (both making it heard and understood); (2) in addition, the speaker needs to make the message believed; (3) to be believed, a message has to supply a subjective dimension; and (4) a great part of the subjective dimension lies in the F0 excursion within lexical items (and other related prosodic cues).

Further studies are necessary regarding the cues of interactions at different levels and the role of emotion underlying any utterance with its prosodic correlates, notably at the lexical level. This requires investigating the function of psychological investment in speech, in other words, the personal, i.e., social and individual characteristics of speech.

How to Prosodically Converge towards more Naturalness in Speech Synthesis

Considering the great amount of research done and underway in the field of prosody, speech analysis should not only deal with linguistics but also with pragmatics,

taking into account all the elements of speech situation and conditions, and in this domain, the speaker's (i.e. subjective) point of view. More explicitly, each domain, linguistics and pragmatics, may claim to integrate the scope of the other domain, as the foreground facts of a domain are also background or related facts of the other. What remains unexplained within a domain is treated as variability and is accounted for as statistical variance. Conversely, from the viewpoint of the other paradigm, it may be taken into account as a significant aspect of speech reality. Many examples of this discrepancy between perspectives could be proposed, e.g. syntax in reading conditions versus syntax in spontaneous speech.

Both in human and automated speech analysis, pragmatics is certainly a better paradigm. Not only does it not deny the linguistic outputs, but it places them on a more reliable basis. If we take it for granted that pragmatics encompasses a linguistic perspective, evidently in certain conditions of communication, or in particular moments of speech, pragmatic requirements would correspond to nothing else than pure linguistic constraints. This approach leads to the idea that the main perspective in prosodic analysis might be essentially oriented towards the speaker's point of view. It might be the link between linguistics and pragmatics, and might help to unify these different perspectives. Natural communication is a matter of person-to-person relation, not a relation between conceptual systems, and in this relation speakers have at their disposal a great deal of resources and tools, which include the linguistic ones, of course, but also para- and extralinguistic ones such as pause, prosodic effects, disruptions, dysfluencies, etc. If we really intend to come as close as possible to a natural speech expression, we need to encapsulate these characteristics in speech synthesis.

Indeed, the choice of words, phrase ordering, sentence structures, i.e. the semantic and syntactic means, contribute to framing and casting meaning in the most appropriate way. In addition, paralinguistic and extralinguistic material is superimposed to clarify, clearly disambiguate, and capture meaning in a subtle, personal way. This material is structured in terms of shared codes; however, its use, occurrence and combination in the actual performance stand for an accurate and personal interpretation of meaning.

Thus, this interpretation outlines a sort of subjective space, whereby the only way to subjectively express meaning is to modify, release, or set here and there the prosodic material *against* the well-framed organisation of linguistic units: for instance, by using unexpected prominence with respect to the syntactic status of the word, or opposing a prosodic grouping (and/or pause) to the syntactic one. Obviously – except in the case of very grave speech disorders – the linguistic organisation may never be broken in practice, because it is a social convention and therefore a reality independent from its actual realisation, which remains apart from the prosodic outputs. This gives a measure of the relatively great freedom given to each speaker to prosodically modify (i.e. capture) the links between linguistic forms (and to some extent, contents) in speech. Even though this linguistic organisation may not be a straitjacket, as the speaker is free to choose lexical items, contexts and combinations, it remains a social convention, something still external and somehow impersonal (Caelen-Haumont, 1997; Caelen-Haumont and Bel, 2001). In fact this impersonality reflects the present situation of our synthetic speech outputs.

Linguistic Structures, Prosody and the Capture of Meaning

In our experience of reading and spontaneous speech in French, prosody encompasses two ways of expressing meaning: in addition to the intonation line which conveys a linguistic framework (essentially the grouping function), the melodic excursion in the local (lexical) domain conveys a subjective relation to meaning. In this relation, the speaker's freedom consists of attributing relative prosodic prominence to lexical items, notably, F0 prominence in terms of maximum and range. In brief, the more the F0 line in the lexical unit deviates from the mean F0 line of the whole prosodic phrase (i.e. intonation), the more this lexical unit (and therefore the phrase) expresses a subjective interpretation of meaning and/or the speaker's communication intention (Caelen-Haumont, 1994).

This F0 local excursion in the word signals the speaker's discrepancy, *or not*, *vis-à-vis* the canonical prosodic expression of linguistic structures. It defines a space of relative freedom at the level of words in the prosodic phrase. Of course these two processes may be combined when, in French for instance, such a word coincides with the right boundary of a phrase. In my opinion, whether they appear combined or not in the output signal, these two processes have distinct origins. If the prosodic expression of linguistic structures has prevailed in synthesis until recently, the consideration of the subjective capture of meaning would significantly contribute to personalising the artificial speech outputs, thereby improving naturalness.

Let us give more precision to this analysis. In French and in other languages, prosody plays two main functions to convey meaning. First, it expresses well-known linguistic functions such as syntactic and semantic ones, both structural (for instance in the semantic domain, theme-rheme organisation). These functions belong to the domain of intonation (F0, but also timing and intensity), shaping sentences and phrases. At this stage, the speaker does not invest him/herself in prosodically reformulating the linguistic links between units (and therefore specifying a subjective meaning). The speaker only gives way to their own linguistic competence. This conventional prosody may however, fit with their voluntary or involuntary purpose. Second, prosody expresses subjective meaning, whose domain is local. It concerns the lexical organisation of melody and related other prosodic parameters already mentioned.

As in other fields, in speech a person establishes their identity by discarding common behaviour to some extent. A space remains free for each speaker, *given the linguistic rules and intonational background*, to disrupt and break down (or conversely to support and even to focus) the syntactic links between units (Caelen-Haumont, 1981; Zellner, 1997). More precisely, this space is prosodically outlined by the F0 range within words (in fact $|\Delta F0|$ because in this space the relevant information lies in the difference between F0 maximum / minimum, not in the direction of the F0 slope), and associated cues such as F0 maximum, duration, and occasionally, intensity, pause, downstepping.

Thus, untying the linguistic link between items is a way of expressing subjective meaning – which might be what 'meaning' stands for in the first place. This process makes sense, because the actual linguistic structure and intonation are interpreted in terms of the speaker's/listener's linguistic and prosodic knowledge. This intimate

use of meaning is a sort of a game-playing around linguistic units; nonetheless, an essential process for complete understanding.

In fact, prosody can be seen as a sort of trade-off between two antagonistic forces, on the one hand, a trend towards social convention, structure, norm, an external point of view, and on the other hand, a tendency towards subjective expression, local viewpoints, rupture, identity, emotion, present realisation. The first one provides cohesive strength, and at the same time, the second one tends to disrupt this continuity (dissociative strength). The first one refers to the social norms of linguistics and its prosodic counterpart, intonation, which both cannot be altered without compromising understanding. The second one corresponds to the subjective use of the linguistic structures, which works at the local/lexical layer. While speakers resort to social norms to prosodically and locally recompose the linguistic links between words at their convenience, the choice of the items to be more or less prosodically focused, even involuntary, is their own. It is a space of subjective freedom, a way of capturing meaning.

This capture of meaning is all the more important, because in spontaneous speech, for instance, background information is not supplied as it would be in a text or a prepared speech. This lack of conceptual accuracy is compensated for by the paralinguistic and extralinguistic precision delivered all at once in speech, in which prosody plays an essential role.

In reading, the speaker is not necessarily the author. Thus the only possibility of investing him/herself in the conceptual domain of the text is to set up a distance *vis-à-vis* linguistic structures, a space of one's own, in which everything is checked and recast in function of personal interpretation and feeling. In other terms, when a reader, or more generally a speaker, is talking and conveying their own point of view, as said before, they can only deliver their own feeling against the background of linguistic structures, simply because the language system allows this personal interpretation. This instantaneous, actual and short-lived speaker's filtering of linguistic meaning is an essential prosodic function. In the second part of the chapter, examples will support this hypothesis.

The next section concerns the melodic excursion within words, from a local and subjective perspective. Figure 36.1 is a fragment of reading in French. In this chunk, for instance, the speaker's prosody obviously recasts the syntactic structure.

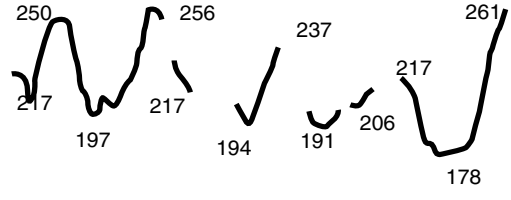


Figure 36.1 Female speaker. Fragment of reading in French extracted from the sentence: '*Ces longs vers prospèrent sur le plancher marin des zones sous-marines profondes*.' ('These long worms are prospering in the deep areas of sea bed'). The numbers correspond to the minimum and maximum F0 values in Hz

If the right boundary of the NP1 ('*vers*') is highlighted, nevertheless the F0 range (|ΔF0|) is smaller than the one of the right boundary ('*marin*') of the prepositional NP, which, moreover, is syntactically of a minor level and dependent (i.e. embedded). In the same fragment, due to the semantic field in progress (the 'giants worms' isotopy, which is the main theme of the text), a wide range is given to the lexical word '*longs*'. The widest one is attributed to the word '*marin*' which is the first occurrence of an unexpected information (i.e. a very deep sea bed is a hospitable place for worms). This chunk of speech displays a relevant example of linguistic structures captured and linguistic links reshaped: pragmatic (and semantic) considerations, such as, for instance, taking into account their addressees, are in the foreground, and prosody enables to highlight this process.

In my opinion, this play (or if you wish, this 'dialogue') between, on the one hand, subjectivity, and on the other hand, linguistic structure, is the closest way of describing the real nature of speech, that is properly its subjective and effective dimension.

Main Prosodic Functions in Speech Generation: Making Known and Making Believed

In this section, I would like to indicate, or define more precisely, some prosodic functions that seem to be important in the context of communication and understanding.

As we have seen, the main function of prosody, which governs the other ones, is pragmatic, and subjective expression is a window opened in this field. In such a perspective, prosody may be viewed as exerting two main functions. One of them is '*making known*'. In this domain, the goals of speakers and generation systems are twofold: on the one hand, making the linguistic units *well heard*, that is to say, allowing a proper demarcation at the phonetic/word level (for instance, in French, prosody, pitch and timing are crucial in left and especially right boundaries), but also at the phrase and sentence levels, which is properly the role of intonation. Here the pragmatic function of prosody fits best with the linguistic one. On the other hand, the second sub-function is to make utterances *well understood*. This goal requires particular focus on specific units from the overall stream, the very ones that seem to carry the main information (whatever it could be) from the speaker's point of view. The following illustrates these considerations.

Figure 36.2 displays an example of a word which is not common ('*jacuzzi*'). This word belongs to a syntactic phrase (the third one from a sequence of six, each of which has a noun in final position). Interestingly, the range of this word is the greatest (229 Hz) of the whole sequence of the six syntactic phrases. According to the needs of their own expression, or the estimated needs of their addressees, the speaker then adjusts the pitch range to the relevance of words in the current pragmatic and subjective conceptual model.

In the example in Figure 36.3, the word '*sources*' ('springs') is given the biggest pitch range, as it is unexpected in the context of deep sea beds. The end of the NP1 '*chaudes*' ('hot') is not highlighted at all: comparatively, its pitch range is rather flat (four-eights of a tone), and there is no pause after it. This prosodic function is given some more concrete evidence, such as for instance in French, when a lexical

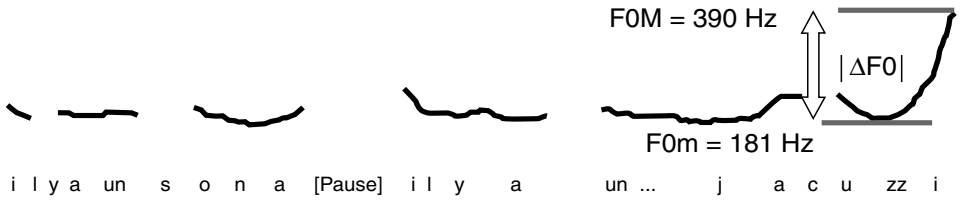


Figure 36.2 Female speaker. Fragment of spontaneous speech: '*... il y a un sauna, il y a un jacuzzi...*' (... 'there is a sauna, there is a jacuzzi' ...). |ΔF0| is expressed in Hz

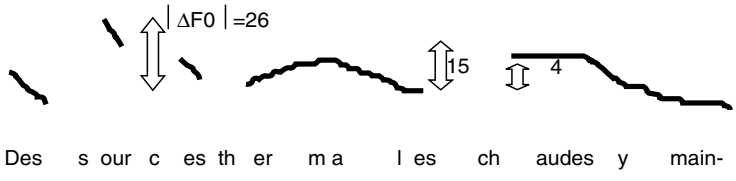


Figure 36.3 Male speaker Fragment of reading: '*Des sources thermales chaudes y maintiennent une température moyenne élevée.*' ('Hot springs keep a high mean temperature') The pitch range (|ΔF0|) is calculated in 1/8th tones

item which is highlighted does not coincide with the right boundary of a phrase, which is usually accentuated. Indeed, both prosodic events may come together at this place.

It remains that in speech communication, 'making known' information is not enough, as a main dimension is the expression of beliefs. Thus, an important prosodic function is *making believed*. Though 'making believed' and 'making known' refer to two different functions, nevertheless they cannot be isolated in the prosodic process: moreover all the examples presented in this chapter illustrate simultaneously these two functions. The only way to be believed is to make known which lexical units convey best our belief and personal truth. Prosody needs to be convincing of its own, on top of the linguistic structure. *The more the speaker invests her/himself in speech, the more they try to be convincing and the more they do so by the way of prosody, thereby evading the regular linguistic framework*.

If prosody is just used as a sort of acoustic paraphrase of linguistic structures, of course the meaning may be available, but it is ill-instantiated in actual speech conditions, and no information is supplied to guide its interpretation. In this situation, the speaker cannot – or will not – deliver a personal interpretation of the linguistic structures. Sometimes this prosodic expression may lead to a better understanding, when the listener acknowledges the speaker's intent and prosodic compliance with speech conditions. This is the case when the speaker refers to an external authority's text or discourse. Here the listener may not expect a subjective meaning conveyed by prosody. Anyway, even though this style might sound correct according to the situation, it becomes rapidly unpleasant and boring. In fact, this effect of prosodic weariness seems to be reached not only because of the repetition of the same syntactic patterns, but also because the person is perceived to some extent as 'absent-minded' in their speech. For instance, belief, a component of a motivated speech,

and – consciously or not – perceived as a strong expression of the speaker's personality, is entrusted to another person or authority in the weary style of speaking. So, pragmatic conditions are not performed exactly, and interest is not aroused.

Moreover, interest in speech is aroused when a person's belief is conveyed and when some kind of innovation takes place. Prosody, in fact, has to say more than syntactic structures can. Furthermore, it brings to the foreground unpredictable meaning at the very moment the listener is decoding utterances, by focusing, or by lowering the importance of words. This process represents at the same time the condition of linking subjectivity to speech and of improving understanding by listener(s), as subjectivity (feelings, beliefs) is made accessible and offered to be shared. Beyond linguistics, a communication process is at work between two persons who recognise each other because they share the same psychoprosodic use and the same rules.

To be more precise, natural speech, when motivated, never departs from a kind of 'emotional' expression. The focus is on 'ordinary' emotion underlying human communication, even in the absence of strong ones. This emotional expression is often under conscious control, but sometimes it is not, of course. According to my experience, in both situations, the differences between melodic ranges in words are mainly filtered by this emotional component, as the speaker expresses a feeling, a personal point of view superimposed on the linguistic stream. Speaker identity and subjectivity prevail, they stand in the foreground. This perspective fits best with other studies in the field of prosody and emotion (Zei, 1995; see also Zei & Archinard, Chapter 24, this volume).

This means that prosody supplies implicit meanings superimposed on the linguistic meaning (and the activation of associated semantic networks). First, referring to the conditions and context of speech, an implicit meaning could be translated in terms such that: 'here this word expresses my feeling that . . .' The contents of this feeling might be such as for instance: 'no doubt that I'm right', or 'mind, you don't expect this word', or 'just consider this word, it will be important later', or otherwise, 'don't mind this one, it has no relevant meaning, it is simply a bridge to the next one' . . .

Second, beyond this function of conveying the expression of one's own feelings or taking care (or not) of addressees' feelings, prosody may also express other implicit meanings, in the case of attitudes, for instance, irony, and especially in dialogue conditions, in the case of indirect speech acts. Figure 36.4 displays an example of irony. In the previous sentence, the speaker was mentioning a street

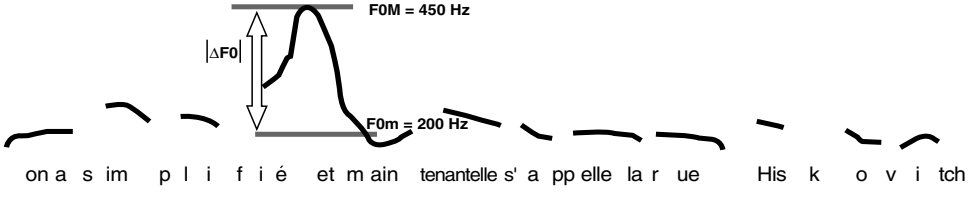


Figure 36.4 Female speaker. Example of irony. Fragment of spontaneous speech. '*on a simplifié, et maintenant elle s'appelle la rue Hiskovitch*' ('[the name] has been simplified and now it is called the street Hiskovitch'). (|ΔF0| is expressed in Hz)

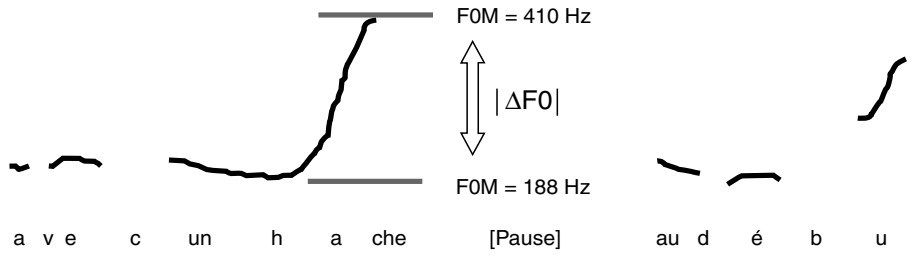


Figure 36.5 Same speaker, same corpus, same sentence . . . '*avec un h au début*' ('with an h at the beginning'). |ΔF0| is expressed in Hz

previously called '*rue de Lyon*'. The prosodic mechanism of irony works at two levels: first, the word '*simplifié*' gets a wide range (250 Hz), second, the following sequence is clearly lowered. We notice that even the informative part, i.e. the new name of the street ('*Hiskovitch*'), is not only focused but lowered as well. This is another clear illustration of the distribution of roles between linguistic structures, which convey information, and lexical prosody which puts an attitude in the foreground.

Figure 36.5 is the next sequence of the same sentence. In this example, the informative part of the sentence (phonetically '*hache*', i.e. 'h'), which expresses a metalinguistic purpose clarifying the spelling of the name '*Hiskovitch*', is again associated with a wide pitch excursion ($|\Delta F0| = 222Hz$) as expected, and a pause (*P*). This melodic range is wider than that of the syntactic phrase boundary '*début*' which is nevertheless hierarchically higher and independent.

In the area of indirect speech act in dialogue, within a given linguistic context (for instance: /it is hot here/), prosody makes it possible to identify an illocutionary act as a question or a statement, and for instance to prompt the listener to act (for instance, here, to open the window). Thus, in the case of irony or indirect speech act, for instance, prosody alone, possibly, or with the support of situation, may convey meaning beyond linguistic items, and even, in their place. In such a function, prosody works precisely as a paraphrase or an antiphrase.

Towards Some Improvements: The Challenge of Subjectivity and the Interpretation of Meaning

As already mentioned before, some characteristics of naturalness in speech may be expressed in terms of variability, diversity, spontaneity, adaptability and subjective capture. The consideration of these properties leads to some recommendations to improve synthetic speech. Following our preceding remarks, these directives deal with three dimensions of speech: context situation, linguistic and subjective levels. Only a few of them will be exemplified relative to these levels.

As explained, standardised speech seems to diverge considerably from natural speech, insofar of course as all the ingredients composing a concrete communicative situation tend to disrupt standardisation. Thus, in my opinion, in synthesis,

a model only based on average structures or average speakers leading to standardisation, is irrelevant if it is not enriched by specific options that deviate from standard output. In fact, models need to be enlarged to encompass *singularity*, which is an important characteristic of natural utterances. Singularity, in turn, may be reached in the subjective space of prosody at the local/lexical level. An interesting challenge is to try to reproduce this inner prosodic trade-off between linguistic structure and subjective expression, which is the private game of taking a distance from lexical items, or to appropriating their meanings. A way of approaching this intimacy is to be carefully sensitive to the speaker's subordinate and superordinate goals and feelings. Goals and feelings are one of the main roads that lead to subjective expression. They are also effective in constructing a classification between lexical items for the specific needs of speech synthesis.

Another way of expressing prosodic subjectivity is to alternate linguistic and/or subjective models. According to what has been observed in spontaneous dialogue (Caelen-Haumont, 1997; Caelen-Haumont and Bel, 2001) and intelligent reading (Caelen-Haumont, 1994), speakers base prosodic expression, and especially pitch range, on those underlying linguistic (syntactic or semantic) or subjective (feeling of complexity of the word contents, lexical field continuity, or unexpected information…) structures or networks that they are sensitive to at the moment of speaking. At that moment, meaning is not yet definitively established by linguistic structures, but is in the process of being determined both productively and receptively. Prosody contributes to isolate and capture one meaning among possible other ones, reinforcing the linguistic one, or conversely operating a double layer of meaning, where prosody, by leaning on linguistic sense and intonation backgrounds, evokes the subjectively appropriated one. This idea illuminates the role of *prosodic structure variance* with respect to linguistic context (Caelen-Haumont, 1994).

These results are in line with other studies in psycholinguistics, based on semantic purposes more than on syntactic ones (Kintsch and Van Dijk, 1978; Le Ny *et al*., 1981–2), for instance, the aspects of 'transitory understanding', and with the idea of competitiveness between several fields of information in speech (Hupet and Cotermans, 1981–2).

Conclusion

To summarise, prosody plays a linguistic function when it highlights phonetic, morphosyntactic, syntactic or semantic *structure*. In this case, the pragmatic function of prosody is restricted to the linguistic one. The reference to speaker subjectivity may then be minimal. This style of prosody may be useful when the speaker cannot – or does not want to – invest or express his feelings, but it is insufficient, or even irrelevant, when speech is subjectively motivated. In this case, another prosodic line is woven into the lexical dimension of the linguistic stream superimposed on the intonation baseline, and the F0 range ($|\Delta F0|$) is assigned the main role. By this very fact, this prosodic style is enhanced in the expression of belief, and it becomes greatly subjective. It contains the prosodic signals (and impulses) for giving rise, among addressees, to interaction.

This chapter aims to identify factors of naturalness in speech generation and synthesis by reintroducing the speaker's point of view concerning linguistic structures, and by superimposing them on the final output. This consideration leads to the idea that once the syntactic level is prosodically fixed, another process can be undertaken to modify local lexical pitch range, places and levels of some F0 maxima values, pause and duration according to users' goals. This process can be grounded in the lexical domain by taking into account semantic and pragmatic considerations during the generation phase.

In my opinion, the new challenge for speech synthesis may be expressed as follows: how to provide an accurate, lively, convenient, or personal meaning to reconfigure the links between linguistic units *without sacrificing anything* to the linguistic meaning, and without being artificial.

After the phase of linguistic structure learning (basic learning), which is the current focus of speech synthesis, another phase might be to break free from a strong dependency on normative linguistic links, in order to adapt these forms to a more subjective expression. This is what infants do, albeit simultaneously, in the course of the acquisition of their first language.

Acknowledgements

Kind thanks to B. Bel and E. Keller for their help with the English expression. Thanks to the European Community and the COST 258 organisation for their support of our meetings and research activities.

References

Caelen-Haumont, G. (1981). *Structures prosodiques de la phrase énonciative simple et étendue*. Hamburger Phonetische Beiträge, Band 34, Hamburger Buske.

Caelen-Haumont, G. (1994). Synthesis: Semantic and pragmatic predictions of prosodic structure. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech recognition* (pp. 271–293). J. Wiley and Sons.

Caelen-Haumont, G. (1997). Du faire-savoir au faire-croire: aspects de la diversité prosodique. *Traitement Automatique des Langues*, *38*, *n*°1, 5–26.

Caelen-Haumont, G. and Bel, B. (2001). Le caractère spontané dans la parole et le chant improvisés: de la structure intonative au mélisme. Revue PArole, 15.

Hupet, M. and Costermans, J. (1981–2). Et que ferons-nous du contexte pragmatique de l'énonciation? *Bull. de Psychologie*, *XXXV, 356*, 759–766.

Kintsch, W. and Van Dijk, T.A. (1978). Toward a model of discourse comprehension and production. *Psychological Review*, *85*, 363–394.

Le Ny, J.-F., Carfantan, M., and Verstiggel, J.-C. (1981–2). Accessibilité en mémoire de travail et rôle d'un retraitement lors de la compréhension de phrases, *Bull. de Psychologie, XXXV, 356*, 627–634.

Zei, B. (1995). Au commencement était le cri. *Le Temps Stratégique*, 96–103.

Zellner, B. (1997). La fluidité en synthèse de la parole. In E. Keller and B. Zellner (eds), *Les défis actuels en synthèse de la parole, Etudes de Lettres*, (pp. 47–78). University of Lausanne, Switzerland.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

37

Synthesis Within Multi-Modal Systems

Andrew Breen
Nuance Communications Inc.
The School of Information Systems
University of East Anglia, Norwich, NR4 7TJ, United Kingdom
andrew.breen@sys.uea.ac.uk

Introduction

There are a number of challenges facing researchers working in the field of speech technology. These challenges stem from the growth of the telecommunications industry and the expectation that speech technology is 'just around the corner'. Researchers working in academia and industry are encouraged to undertake their work with the aim of developing practical systems. This has proved to be a mixed blessing. The pressure to develop high-quality systems has encouraged researchers to investigate efficient and practical solutions, but at some cost to basic research.

This chapter suggests that the compromise is well worth making, and moreover, that recent developments in speech technology and related disciplines, such as multi-media interfaces, will force speech researchers to re-evaluate many basic assumptions about what constitutes recognition and synthesis systems.

Over the last two decades, researchers have regularly predicted the imminent appearance of speech technology in our everyday lives. This technology, however, still has a long way to go before it can become widespread in society. Some services are starting to appear, but this slow change in fortune has more to do with dramatic changes in the computer and telecommunications industries, than with any significant breakthrough in speech technology. The simple fact is, that while some applications can now be handled adequately with the current generation of systems, the advanced applications, those wanted by the vast majority of people, are still many years away.

An obvious question to ask at this point is 'So what do people really want?'. The simplest way to answer this question, is to divide speech technology into three broad application areas:

1. Natural discourse with a machine.

2. Data summarisation.
3. Domain-sensitive database retrieval.

Natural Discourse with a Machine

In its widest sense, this represents the Holy Grail for many speech researchers, where the term 'natural' in this context is defined as meaning a man–machine interaction which exhibits all the components of a conversation between two humans speaking on the some topic and in the same speaking style.

Discourse in its most general sense contains both verbal and non-verbal cues. A great deal of information is born in the speech signal, but a significant amount of information is also contained in pausing and gestures. Truly natural conversation with a machine will only be possible once all the components, both verbal and non-verbal, are considered and mastered. Simply stated, the computer must be sensitive to the emotional and physical state of the human interlocutor. To do this, it must have access to as many different modalities as possible.

Data Summarisation

Many applications require, or are enhanced through, the use of data summarisation. Document summary, for example, is a growth area, but the current generation of summarisers do not attempt to analyse and re-interpret the text. Instead, they use statistical techniques to extract highly relevant portions of existing text. True summarisation would attempt to understand the contents of the document and restate it in a way appropriate to the discourse and the communication medium. As an example, consider the short e-mail given below:

Hi Tom,
  Got your E-mail regarding the meeting on Friday with the director. I'll be there.
Regards,
Jerry.

Consider now a brief spoken dialogue, taken from an imaginary advanced automated E-mail enquiry system:

Human user: Do I have any E-mail from Jerry about the directors meeting?
E-mail system: Yes, you have one from Jerry. He says he can make the meeting on Friday.

Here the e-mail system has interpreted the user's request and generated an appropriate reply.

The example above demonstrates that improved naturalness can be achieved through the appropriate use of summarisation and language generation. In general, there is a significant difference between written and spoken language. Advanced spoken language systems must be sensitive to these differences. Spoken language contains many false starts, filled pauses and grammatically ill-formed utterances. In contrast, written language is typically grammatically well formed and rich in

grammatical structure. In addition, written language when performed by a speaker tends to suffer from fewer false start and filled pauses.

Domain-Sensitive Database Retrieval

Many texts should not be summarised, but still need to be treated differently depending upon their specific characteristics. For example, an early application of speech synthesis was as reading machines for the blind. The current generation of reading machines are insensitive to the type of material they are reading. A book is read in the same impassive style as a newspaper. The desired solution is to have a system, which interprets the text and provides an intoned reading.

There is a common theme running through all the application areas described above, which is the need to provide synthesis systems with the ability to control the style and emotion of synthetically generated spoken language. Speech synthesis should not be seen as an isolated component independent of the language generated, but as part of the larger process of expressing meaning and emotion through spoken language.

The Growth of Multi-Media Applications

The traditional application areas for speech technology are changing faster than the component technologies. Such changes are making many of the traditional assumptions about the market for speech technology out of date before they have even been developed! The growth of multi-media, particularly within the telecommunications industry, has forced many applications developers to reconsider the basic requirements of the component technology. In addition, with the expanding research areas of intelligent user interfaces and augmented reality, the goals of speech recognition and synthesis research are also changing. In that sense, speech technology can no longer afford to divorce its self from other disciplines such as artificial intelligence and kinetics. In fact, this trend towards a broader appreciation of the requirements of the technology is already visible in the variety of topics being presented at international spoken language conferences. Unfortunately, most if not all of the traditional problems associated with speech technology remain. It is still the case that for many applications the underlying performance of both speech recognition and speech synthesis is not yet at an acceptable level. Continuous speech, speaker-independent, large vocabulary recognition systems are currently unable to cope with applications in which untrained users wish to interact with a system on an ill defined task. Most large vocabulary recognition systems rely heavily on language models. Provided a speaker remains within the expectations of these models, the system will perform well. However, performance suffers as soon as a speaker deviates from these models.

Speech synthesis systems still have a long way to go before they start to compare with human performance. The current generation of 'corpus-based' synthesis systems have achieve highly natural voice quality but at the cost of flexibility. Such systems work by selecting stretches of text from very large inventories of speech. When the requested text closely matches the content and underlying style of these

inventories, the resulting synthesis is of a very high calibre. However, as soon as the text deviates from the coverage of stored sounds, the perceived quality falls. In addition, to maintain high voice quality the minimum amount of signal processing is applied to the concatenated segments of speech data. This approach works well provided the inventory of speech has sufficient coverage of sounds, but is in reality a developmental cul-de-sac, as any manipulation of the speech data results in a perceived drop in the overall quality of the synthesised utterance.

This chapter will concentrate on issues facing researchers working in the area of speech synthesis, however, many of the conclusions are relevant to all areas of speech research. It is suggested that far from being an extra burden on speech technology, the requirements imposed by the need to develop advanced multimodal systems represent in fact an essential step in the development of speech technology. The rest of this chapter will consider some of the issues facing speech synthesis researchers, and suggest that the most effective way of advancing the technology is to consider the requirements imposed by the application areas described above.

The Future of Speech Generation

The introduction suggested that the goal of speech technology was to provide a 'natural' interface to machines, and that this goal may be expressed through three different but related application areas. What differs most between these application areas is the type of data that is presented to the speech synthesis system, and the style of speech produced by the system. Consequently, our speech synthesis systems must be rendered sensitive to these differences.

The current generation of text-to-speech systems have been designed to accept plain text as input. The majority of typographical ambiguities found in unrestricted text are handled by a process known as text normalisation, while trivial cases may be handled using a limited set of escape sequences. Plain text does not contain sufficient information to correctly resolve many types of ambiguity. As a result, a large number of the decisions made by the text normalisation process are in reality arbitrary. This problem is not restricted to the process of text normalisation, but seriously affects all aspects of the speech synthesis process. To illustrate this point, consider the following example. Imagine that as part of some dialogue, an automated system has generated the question given below:

Did you say fifty pence?

The meaning of this question changes with word emphasis. For example, if emphasis was placed on the word *pence*, the system is asking the user to confirm that the amount was in pence rather than pounds, whereas, if emphasis was place of the word *fifty*, the system is asking the user to confirm that the number was *fifty* as opposed to some other amount. Be default, with only plain text to work with, text-to-speech systems would typically place greatest prominence on 'pence'.

This simple example could easily be handled by attaching a synthesiser-dependent emphasis flag to the desired word, e.g.

Did you say <esc> E fifty pence?

The above example demonstrates this using an invented escape sequence. In this example, the escape sequence 'E' is used to trigger emphasis on the following word. Such escape sequences or flags are commonly used to a greater or lesser degree by most text-to-speech systems. Flags are often used to modify the behaviour of the text normalisation and word pronunciation processes, but only comparatively few systems have flags to modify the emphasis applied to a particular word.

To date, embedded flags of the sort considered above are not sophisticated enough to provided a realistic mechanism for significantly modifying the behaviour of a text-to-speech system. However, a number of researchers are starting to investigate more advanced flag sets (Sproat, *et al.* 1998). Many of the proposed systems are based on the widely used XML standard.

Figure 37.1 shows diagrammatically how researchers and developers have envisaged the interface to TTS systems. Figure 37.1a shows the traditional approach to speech synthesis, where plain text is presented to the system. A TTS system may be asked to convert text from a news article, a poem, a discourse or even a train timetable. Typically little or no pre-processing will be performed on the text. Is it any wonder, then, that the quality of the speech produced by such systems is so stilted? TTS researchers are faced with two choices: either to increase the level of linguistic analysis conducted on the text within the TTS system, or to encourage users to extend the type of information presented to a speech synthesis system.

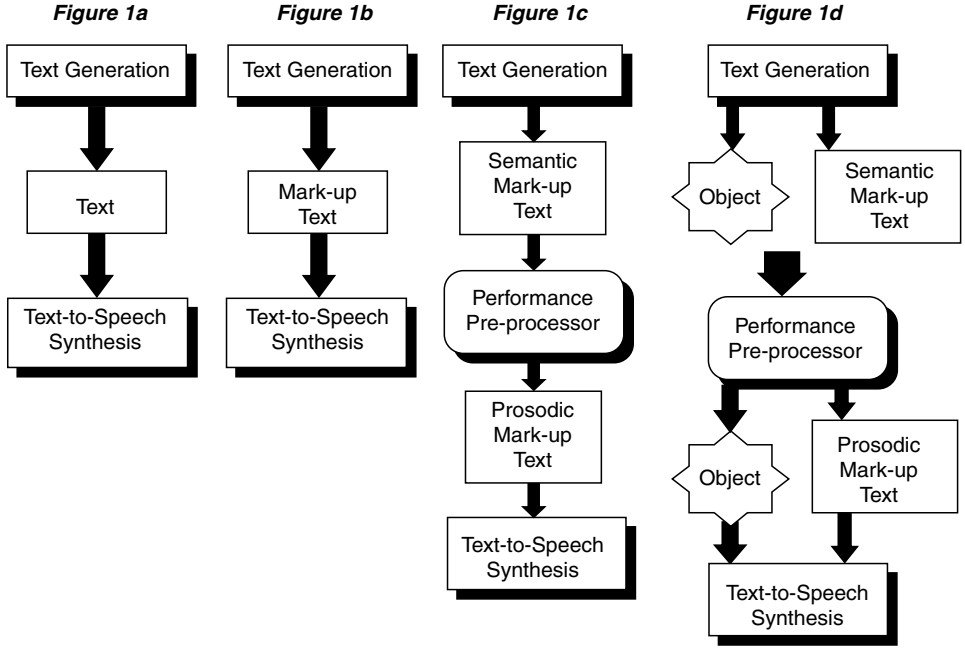


Figure 37.1 Text generation to TTS interface

If the first choice is taken, TTS systems will be forced to balloon in size, effectively taking on the majority of the tasks involved in the interpretation and presentation of information. This is an unpalatable choice for many synthesis researchers, as linguistic analysis, while being a necessary requirement for the production of synthetic speech, is by no means a sufficient requirement. That is, knowing what to say is not the same as knowing how to say it. Even if it was considered desirable to included textual interpretation within a system, it is difficult to imagine how such a system could be constructed. Such a system would through necessity need to understand the context in which the text was being spoken. In other words, the TTS system would need to either control or have full access to information on the application in which it was embedded.

If the second choice is taken, as shown in Figure 37.1b, researchers are obliged to investigate effective methods of encoding linguistic and paralinguistic information, which in itself is currently an ill-defined and complex task.

Figure 37.1b shows a system, which has marked-up text presented to the interface. However, it is unclear from this diagram what information is contained within the mark-up. Are we to assume that a comprehensive and complete XML language will be devised that can address all the issues facing synthesis developers? In other words, is this language going to contain structures to handle semantic, pragmatic, syntactic, prosodic and stylistic information? If so, to what level of detail? Will the prosodic information contain fundamental frequency contours and segmental duration information? If it were, then surely nothing would have been gained. Also, which component within a system is responsible for ensuring that the information presented to the synthesis system is both complete and not contradictory? If a synthesiser receives incomplete information it has two choices. First of all, it can fail, thus passing the responsibility of completeness back to the calling process. Alternatively, it could attempt to 'fill in' missing information. Such an approach is fraught with problems; for example, information generated within the synthesis system may contradict or invalidate information presented within the marked-up text. Finally, if an application developer, through mark-up, overrides the preferred setting of the synthesiser, who is responsible for the reduced quality of the system?

Clearly, there is a need for a clear distinction between an applications programmer's interface, and a researchers interface. The question is, should these interfaces be resident within the same mark-up language?

Figure 37.1c shows a two-stage approach. In this approach, an attempt has been made to differentiate between semantic/pragmatic information, and prosodic information. The process involves two levels of mark-up; the first high-level description is interpreted by a pre-processor, which re-formulates the requirements into a prosodic description of some form. However, such an approach does not address the problematic issue of what an appropriate descriptive level is.

Figure 37.1d addresses the final issue facing researchers wishing to develop standard interfaces for synthesis systems. The figure suggests that mark-up is not the only way of interfacing with a component. Specifically, it is likely to be a verbose and inefficient way of passing complex information across an interface layer. With the growth of OO-based design methods, and middle-ware inter-operability standards such as CORBA and DCOM, alternative, more efficient and flexible interfaces can be envisaged.

XML is clearly an important development is speech synthesis research, since it offers an extensible framework for annotating complex linguistic information and sophisticated tools to manipulate such information. The problem stems not so much from XML itself, but from the attempts by researchers to devise a single complete interface standard to synthesis systems. The author believes that, far from aiding research, an overemphasis on the development of interface standards will obscure the basic issues facing researchers. In other words, a great deal of fundamental research is needed to define what information should be present within such interfaces before any hard and fast decisions can be made regarding a standard.

Multi-Modal Speech Synthesis

It is clear from the previous section that the traditional interpretation of what constitutes a text-to-speech system must change. Plain text does not contain sufficient information. Some form of extended text interface or structured interface which contains supplementary information is needed. The definition of such interfaces does not lead to an immediate increase in the quality of synthesis. A great deal of work is needed to translate abstract information into concrete improvements in voice quality. Nor is it the case that the current generation of synthesisers, given the appropriate information, whatever that may be, is good enough to generate truly natural speech. However, the trend is clearly back towards concept-to-speech synthesis.

The chapter also suggests that the drive to develop complex multi-modal systems is a necessary next step in the development of synthesisers. The belief that this is the case stems directly from the arguments outlined above. Given that plain text can no longer be viewed as an acceptable interface to high quality synthesis systems, and that it is impractical to imagine that synthesis systems will balloon to accommodate significant amounts of text analysis, only one path remains, which is the definition of a rich interface. Such an interface is only useful if some external process first interprets and generates information appropriate to this interface. Such a process would inevitably form part of a complete system, which would be either application-specific or so sophisticated as to accommodate domain independence. A synthesis system embedded within such a framework would have the advantage of concrete requirements. In addition, the consequence of any modification to the interface or to the algorithms used within the various sub-systems could be directly evaluated through experimentation. The author believes that only in this way will truly flexible standards be developed. Also, undertaking research into embedded systems will highlight particular deficiencies in component technologies.

Figure 37.2 shows a simplified model of how information may be passed to a synthesis system. The figure is based on the structure incorporated within the Maya multi-modal dialogue system (Downey *et al.* 1998). Notice that data may be generated from a number of different sources, and that each source would have specific characteristics. Data derived from a text generator would be rich in information, while data from other sources would have varying degrees of supplementary information. In this example, it is the task of the information agent to organise and forward information to the relevant service and additionally to the appropriate

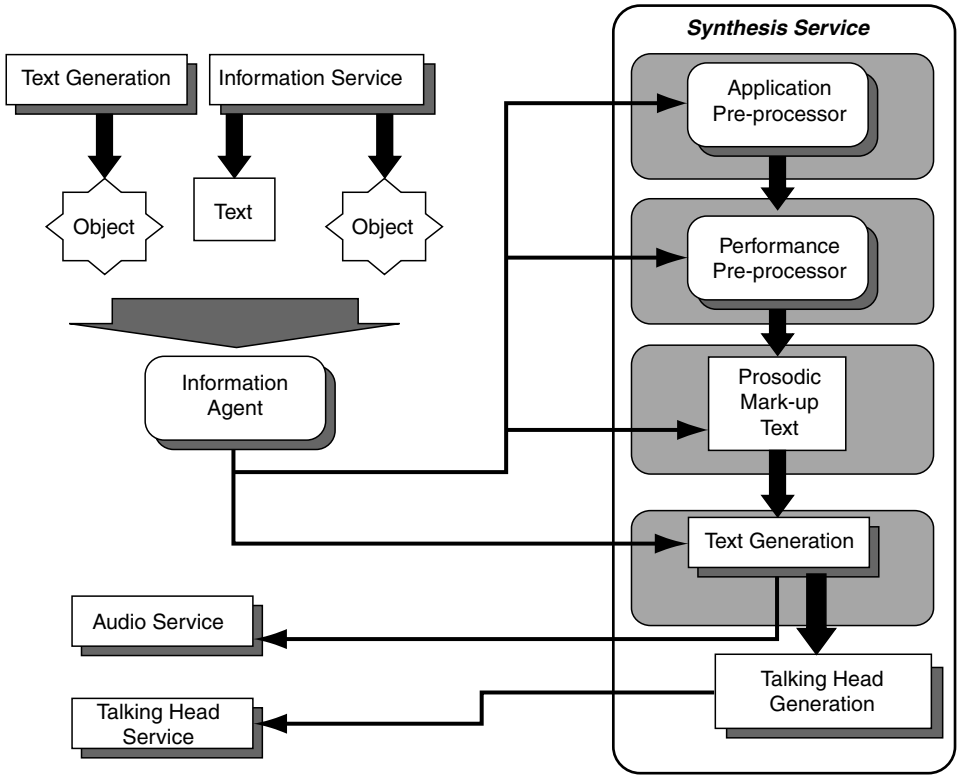


Figure 37.2 How information is passed to a synthesis system

interface level. Notice also that the synthesis system has been renamed a synthesis service, which is composed of a number of components, where the lowest-level interface is represented by the traditional interpretation of a TTS system, while the highest level is dedicated to application pre-processors, e.g. e-mail filters.

The synthesis service no longer passes the output directly to a sound source. Instead, the information produced by the service may be forwarded to other agents or services for further processing. The output of such a service is not an unstructured sequence of samples, but a complex data structure. This is demonstrated by the addition of a talking head component. In fact, the next generation of synthesiser will not only be expected to produce audio they may well be expected to generate the visual correlates of the speech as well. With the introduction of visual speech, many of the arguments considered in the earlier section are further exacerbated as the data presented to the synthesiser may contain information specific to the visual aspects of the synthesis process as well as the acoustic.

Conclusion

This chapter suggests that the next generation of synthesis systems will inevitably take more account of the type of information being presented to them; that the

interfaces to such systems will become more generic, and that the type of processing conducted as part of the synthesis process will become more diffuse and data orientated. The chapter also suggested that advances in speech synthesis could best be achieved when developed within a complete multi-modal framework.

Acknowledgements

The author wishes to thank COST 258 for providing an effective forum for discussion of important issues in the development of multilingual synthesis systems.

References

Downey, S., Breen, A.P., Fernandez, M., and Kaneen, E. (1998). Overview of the Maya spoken language system. *Proceedings of ICSLP '98* (paper 391). Sydney.

Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K., and Edgington, M. (1998). SABLE: A standard for TTS Markup. *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis* (pp. 27–30). Jenolan Caves, Australia.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

38

A Multi-Modal Speech Synthesis Tool Applied to Audio-Visual Prosody

Jonas Beskow, Björn Granström and David House
Centre for Speech Technology, Dept. of Speech, Music and Hearing, KTH, Stockholm, Sweden
beskow | bjorn | davidh@speech.kth.se

Introduction

Speech communication is inherently multi-modal in nature. While the auditory modality often provides the phonetic information necessary to convey a linguistic message, the visual modality can qualify the auditory information and provide segmental cues on place of articulation, prosodic information concerning prominence and phrasing and extralinguistic information such as conversational signals for turn-taking, emotions and attitudes. Although these observations are not novel, prosody research applied to speech synthesis has largely ignored the visual modality. One reason is the primary status of auditory speech; another has been the absence of flexible tools for the synthesis of visual speech.

The visible articulatory movements are mainly those of the lips, jaw and tongue. However, these are not the only visual information carriers in the face during speech. Much information related to e.g. phrasing, stress, intonation and emotion are expressed by head movements, raising and shaping of the eyebrows, eye movements and blinks, for example. These kinds of facial actions should also be taken into account in a visual speech synthesis system, not only because they may transmit important non-verbal information, but also because they make the face look alive, thus adding to naturalness. These movements are more difficult to model in a general way than the articulatory movements, since they are optional and highly dependent on the speaker's personality, mood, purpose of the utterance, etc. (Cavé *et al*., 1996). Recent examples of prosodic information in facial animation systems can be found in Massaro *et al*. (2000) and Poggi and Pelachaud (2000).

Apart from the basic research issues in multi-modal speech synthesis, there is currently considerable interest in developing 3D-animated agents for use in multi-modal spoken dialogue systems (Cassell, 2000) and in computer-aided language

learning (CALL) applications (Badin *et al.* 1998; Cole *et al.*, 1999). As the systems become more conversational in nature and the 3D-animated agents become more sophisticated in terms of visual realism, the demand for natural and appropriate prosodic and conversational signals (both verbal and visual) is clearly increasing.

This chapter is concerned with prosodic and conversational aspects of visual speech synthesis. A distinction can be made in visual synthesis between articulatory and prosodic cues. The visual articulatory cues are related to the production of the speech segments (e.g. lip aperture size, lip movement, jaw rotation, tongue position) and provide information primarily on place of articulation, vowel features, vowel–consonant alternation and syllable timing. Prosodic cues (e.g. head, gaze and eyebrow movement) overlie the segmental phonetic information of the articulatory cues, and can in the same way as verbal prosody provide information on prominence and phrasing as well as extralinguistic information.

In this chapter, an advanced research tool is first described which is used to experiment with prosodic signals on both a parametric and symbolic level. A pair of audio-visual perception experiments is then presented in which the aim is to quantify to what extent upper face movement cues can serve as independent cues for the prosodic functions of prominence and phrasing. The use of audio-visual prosodic signals in two types of applications (spoken dialogue systems and automatic language learning) is then discussed and exemplified.

Modelling and Synthesis of Talking Faces

Animated synthetic talking faces and characters have been developed using a variety of techniques and for a variety of purposes during the last two decades. Our own approach is based on parameterised deformable 3D facial models, controlled by rules within a text-to-speech framework (Carlson and Granström, 1997). The rules generate the parameter tracks for the face from an augmented representation of the text, taking coarticulation into account (Beskow, 1995).

We employ a generalised parameterisation technique to adapt a static 3D-wireframe of a face for visual speech animation (Beskow, 1997). Based on concepts first introduced by Parke (1982) we define a set of parameters that will deform the wireframe by applying weighted transformations to its vertices. One critical difference from Parke's system, however, is that we have de-coupled the model definitions from the animation engine, thereby greatly increasing flexibility. For example, this allows us to define and edit the weighted transformations using a graphical modelling interface, rather than hand-coding them into e.g. C-source files. The parameterisation information is then stored, together with the rest of the model, in a specially designed file format. For all our models developed to date, we have decided to conform to the same set of basic control parameters for the articulators that was used in Beskow (1995). This has the advantage of making the models independent of the rules that control them during visual speech synthesis; all models that conform to the parameter set will produce compatible results for any given set of parameter tracks. The animation engine and the modelling interface currently run under Windows and several UNIX dialects.

Parameter Manipulation

For stimuli preparation and explorative investigations, we have developed a control interface that allows fine-grained control over the trajectories for acoustic as well as visual parameters. The interface is implemented as an extension to the WaveSurfer application (www.speech.kth.se/wavesurfer) (Beskow and Sjölander, 2000), which is a tool for recording, playing, editing, viewing, printing, and labelling audio data. The interface makes it possible to start with an utterance synthesised from text, with all the parameters generated by rule, and then interactively edit the parameter tracks for any parameter, including F0, visual (non-articulatory) parameters as well as the durations of individual segments in the utterance to produce specific effects. An example of the user interface is shown in Figure 38.1. In the top box a text can be entered either in Swedish or English. This creates a phonetic transcription that can then be edited. On pushing 'Synthesize', rule-generated parameters will be created and displayed in different panes below. The selection of parameters is user-controlled. The lower section contains segmentation and the acoustic waveform. The talking face is displayed in a separate window. The acoustic synthesis can be exchanged for a natural utterance and synchronised to the face synthesis. This is useful for different experiments on multimodal integration and has been used in the Teleface project (Agelfors *et al.*, 1999), aiming at a telephone device for hard-of-hearing persons, (an automatically produced demonstration of

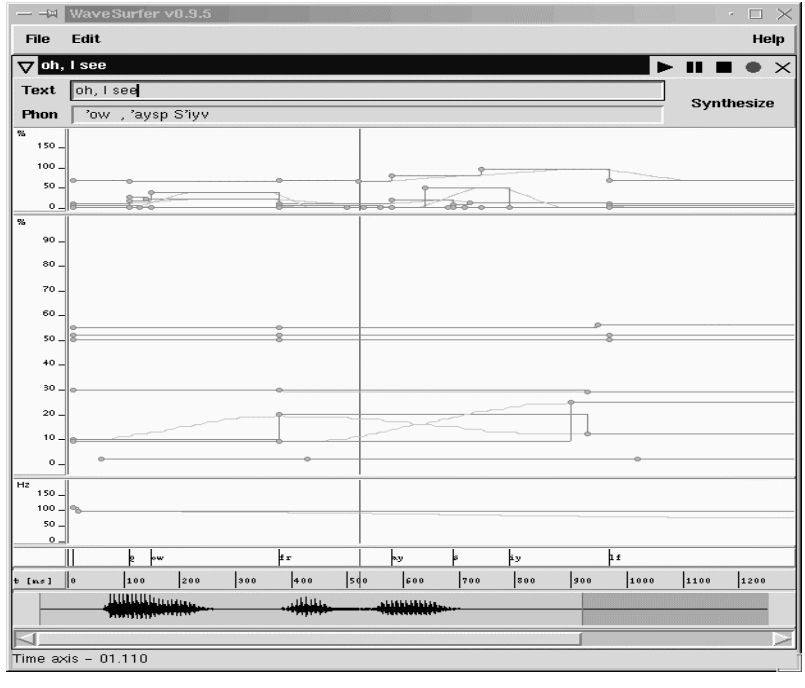


Figure 38.1 The WaveSurfer user interface for parametric manipulation of the multi-modal system

this can be seen in video-clip A). In automatic language learning and pronunciation training applications, it could be used to add to the naturalness of the tutor's voice in cases when the acoustic synthesis is judged to be inappropriate.

Symbolic Representation

The parametric manipulation tool described in the previous section is used to experiment with and to define different gestures. A gesture library is under construction, containing procedures with general emotion settings and non-speech specific gestures, as well as some procedures with linguistic cues. These procedures serve as a base for the creation of new communicative gestures in future animated talking agents, used in multi-modal spoken dialogue systems and as automatic tutors.

For example, we have included a set of markers for emotional settings for the conversational agent used in the August project (Gustafson *et al.*, 1999; Lundeberg and Beskow, 1999). To enable display of the agent's different moods, six basic emotions similar to the six universal emotions defined by Ekman (1979) were implemented in a way similar to that described by Pelachaud *et al.* (1996).

We are at present developing an XML-based representation of visual cues that facilitates description of the visual cues at a higher level. These cues could be of varying duration like the above-mentioned emotions that could sometimes be regarded as settings appropriate for an entire sentence or conversational turn, or be of a shorter nature like a qualifying comment to something just said. Some cues relate to e.g. turntaking or feedback, and need for that reason not be associated with speech acts, but can occur during breaks in the conversation. It is important that there exists a one-to-many relation between the symbols and the actual gesture implementation to avoid stereotypic agent behaviour. Currently, a weighted random selection between different realisations is used.

Phrasing and Prominence Perception Experiments

The interaction between acoustic and visual cues in communication has been discussed previously. Specifically the interaction between acoustic intonational gestures (F0) and eyebrow movements have been studied in for example, Cavé *et al.* (1996). The tool described above makes experiments in this area possible. In video-clip B, different examples of F0/eyebrow relationship can be seen for an example sentence. A preliminary conclusion is that a direct relation is very unnatural, but that prominence and eyebrow movement may co-occur.

To investigate this more formally, the following pair of experiments was designed to examine the independent contribution of eyebrow movements to the audio-visual perception of phrasing and prominence in a single test sentence. The test sentence used to create the stimuli for the experiments was ambiguous in terms of an internal phrase boundary. Acoustic cues and lower face visual cues were the same for all stimuli. Articulatory movements were created by using the text-to-speech rule system. The upper face cues were eyebrow movement where the eyebrows were raised on successive words in the sentence.

The movements were created by hand editing the eyebrow parameter using the synthesis parameter editor. The degree of eyebrow raising was chosen to create a subtle movement that was distinctive, although not too obvious. The total duration of movement was 500 ms and comprised a 100 ms dynamic raising part, a 200 ms static raised portion and a 200 ms dynamic lowering part. The synthetic face *Alf* with neutral and with raised eyebrows is shown in Figure 38.2.

The same 21 subjects participated in the two experiments described below. All were part of a speech technology class taught at KTH. No one reported any hearing loss or visual impairment. 14 subjects had Swedish as their mother tongue. All except one reported that they had a central Swedish (Stockholm) dialect.

Seven subjects had other mother tongues than Swedish (1 Finnish, 2 French, 2 Italian and 2 Spanish), but all had working competence in Swedish (attending a masters level class given in Swedish at KTH). In the results section, the results of the total group as well as for these subgroups are presented.

Experiment 1 – Phrasing

In a previous study concerned with prominence and phrasing, using acoustic speech only, ambiguous sentences were used (Bruce *et al*., 1992). In the present experiment we used one of these sentences:

(1) När pappa fiskar stör, piper Putte
(When dad is fishing sturgeon, Putte is whimpering)
(2) När pappa fiskar, stör Piper Putte
(When dad is fishing, Piper disturbs Putte)

Figure 38.2 The synthetic face *Alf* with neutral eyebrows (left) and with eyebrows raised (right)

Hence, 'stör' could be interpreted as either a noun (1) or a verb (2); 'piper' (1) is a verb, while 'Piper' (2) is a name.

In the stimuli, the acoustic signal is always the same, and synthesised as one phrase, i.e., with no phrasing prosody disambiguating the sentences. In Bruce *et al.* (1992), different segmental and prosodic disambiguation strategies are discussed. In the present series of experiments the possibility of visual disambiguation was investigated. Six different versions were included in the experiment: one with no eyebrow movement and five where eyebrow rise was placed on one of the five content words in the test sentence. In the test list of 20 stimuli, each stimulus was presented three times in random order. The first and the last item of the list were dummies and not part of the data analysis.

All subjects participated in the same session. The audio was presented via loudspeakers and the face image was shown on a projected screen, four times the size of a normal head. The viewing distance was 3 to 6 metres, simulating a normal face-to-face conversation distance of 0.75 to 1.5 metres. In this range of distances the visual intelligibility is judged to be close to constant (Neely, 1956).

The subjects were instructed to listen as well as to speech-read. Two seconds before each sentence, an audio beep was played to give subjects time to look up and focus on the face. No mention was made of eyebrows. The subjects were made aware of the ambiguity in the test sentence and were asked to mark the perceived interpretation for each sentence.

In Figure 38.3 the results from experiment 1 can be seen. It is obvious that there is a bias for all the stimuli to more often (about 60%) be perceived with a phrase boundary after 'stör', i.e. interpretation (1).

This is possibly also the default interpretation of the sentence without speech for most subjects, since Piper is a rather uncommon name. On the whole, very little difference is seen between the different stimulus conditions.

The non-Swedish subjects behaved very much like the Swedes, perhaps with one exception. For the Swedish subjects there was a small increase in the (1) interpretation when there was an eyebrow rise on p/Piper. One possible explanation could be that an eyebrow movement could be associated with a phrase onset, but on the whole there is rather limited evidence in this experiment that eyebrow movements contributed to phrasing information.

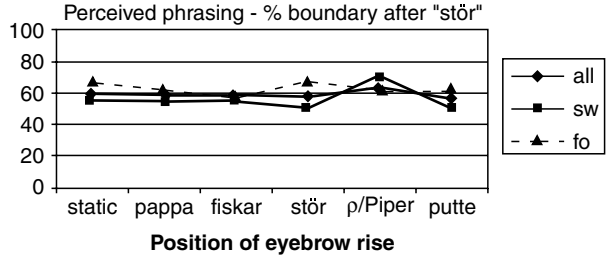


Figure 38.3 Result of the phrasing/disambiguation experiment
Note: Interpretation (1) i.e. a phrase boundary after 'stör' (rather than after 'fiskar'). sw : 14 subjects with Swedish as their mother tongue. fo: 7 non-Swedish subjects

Experiment 2 – Prominence

In the second experiment we used the same stimulus material as in experiment 1, but the question now concerned prominence. The subjects were asked to circle the word that they perceived as most stressed/most prominent in the sentence. The results are shown in Figure 38.4. Figure 38.4/static refers to judgements when there is no eyebrow movement at all.

The distribution of judgements varies with both subject group and word in the sentence. This could be related to phonetic information in the auditory modality, since the intonational default synthesis used here put a weak focal accent on the first and the last word in a sentence. This could explain the many votes for the first and the last word, 'pappa' and 'Putte' in Figure 38.4/static. However, it may well be related to prominence expectations. In experiments where subjects are asked to rate prominence on words in written sentences, nouns tend to get higher ratings than verbs (Fant and Kruckenberg, 1989). This is supported by our data, since 'stör' has the default interpretation of a noun and p/Piper the default interpretation

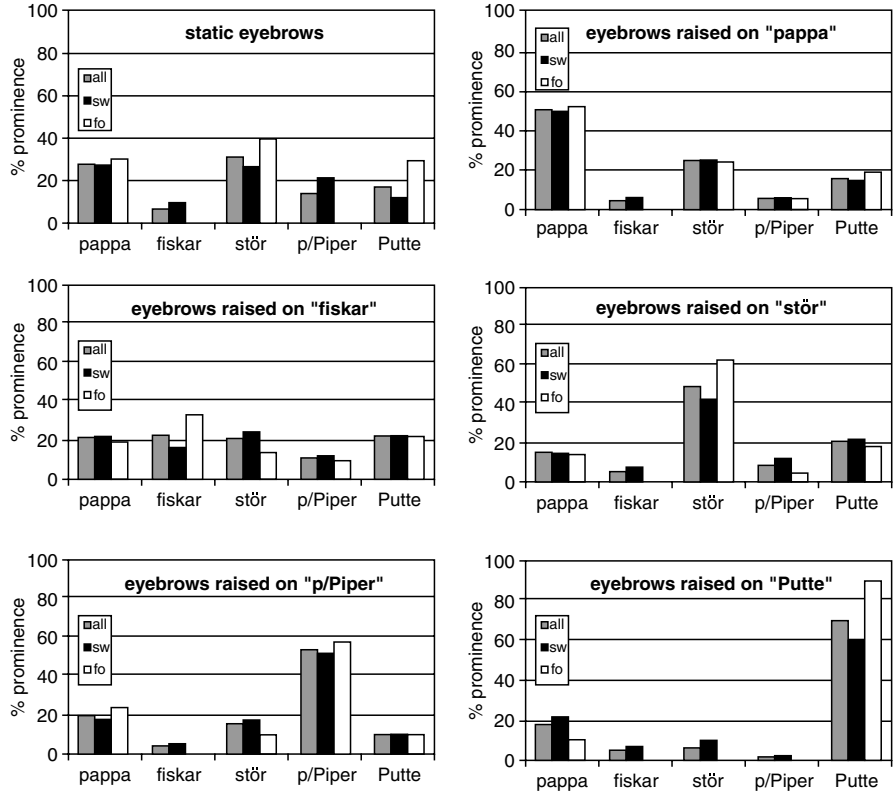


Figure 38.4 Prominence responses in percentage for each word and each stimulus condition

Note: Subjects are grouped as all, Swedish (sw) and foreign (fo)

of a verb in experiment 1, while 'fiskar' is always a verb in these contexts. The non-Swedish subjects seem to behave slightly differently in this experiment, since no prominence votes are given to 'fiskar' and 'p/Piper'.

The results of the prominence experiment indicate that eyebrow raising can function as a perceptual cue to word prominence, independent of acoustic cues and lower face visual cues. In the absence of strong acoustic cues to prominence, the eyebrows may serve as an F0 surrogate or they may signal prominence in their own right. While there was no systematic manipulation of the acoustic cues in this experiment, a certain interplay between the acoustic and visual cues can be inferred from the results. As mentioned above, a weak acoustic focal accent in the default synthesis falls on the final word 'Putte'. Eyebrow raising on this word (Figure 38.4/Putte) produces the greatest prominence response in both listener groups. This could be a cumulative effect of both acoustic and visual cues, although compared to the results where the eyebrows were raised on the other nouns, this effect is not great.

In an integrative model of visual speech perception (Massaro, 1998), eyebrow raising should signal prominence when there is no direct conflict with acoustic cues. In the case of 'fiskar' (Figures 38.4/static and 38.4/fiskar) the lack of specific acoustic cues for focus and the linguistic bias between nouns and verbs, as mentioned above, could account for the absence of prominence response for 'fiskar'. Further experimentation where strong acoustic focal accents are coupled with and paired against eyebrow movement could provide more data on this subject.

It is interesting to note that the foreign subjects in all cases responded more consistently to the eyebrow cues for prominence, as can be seen in Figure 38.5. This might be due to the relatively complex Swedish F0 stress/tone/focus signalling and the subjects' non-native competence. It could be speculated that eyebrow motion is a more universal cue for prominence.

The relationship between cues for prominence and phrase boundaries is not unproblematic (Bruce *et al.*, 1992). The use of eyebrow movement to signal phrasing may involve more complex movement related to coherence within a phrase rather than simply as a phrase delimiter. It may also be the case that eyebrow raising is not an effective independent cue for phrasing, perhaps because of the complex nature of different phrasing cues.

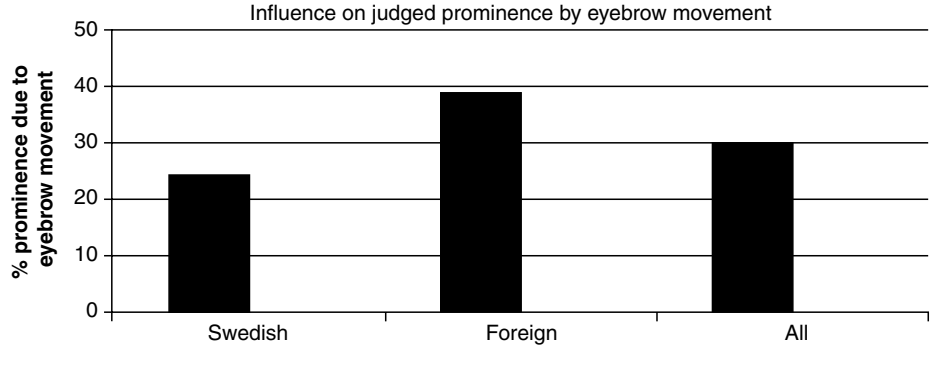


Figure 38.5 Mean increase in prominence judgement due to eyebrow movement

This experiment presents evidence that eyebrow movement can serve as an independent cue to prominence. Some interplay between visual and acoustic cues to prominence and between visual cues and word class/prominence expectation is also seen in the results. Eyebrow raising as a cue to phrase boundaries was not shown to be effective as an independent cue in the context of the ambiguous sentence. Further work on the interplay between eyebrow raising as a cue to prominence and eyebrow movement as a visual signal of speaker expression, mood and attitude will benefit the further development of visual synthesis methods for interactive animated agents in e.g. spoken dialogue systems and automatic systems for language learning and pronunciation training.

Implementation of Visual Prosody in Talking Agents

In the course of our work at KTH dealing with developing multi-modal dialogue systems, we are gaining experience in implementing visual prosody in talking agents (e.g. Lundeberg and Beskow, 1999). We have found that when designing a talking agent, it is of paramount importance that it should not only be able to generate convincing lip-synchronised speech, but also exhibit a rich and reasonably natural non-verbal behaviour including gestures which highlight prosodic information in the speech such as prominent words and phrase boundaries, as in the experiment just described. As mentioned above, we have developed a library of gestures that serve as building blocks in the dialogue generation. This library consists of communicative gestures of varying complexity and purpose, ranging from primitive punctuators such as blinks and nods to complex gestures tailored for particular sentences. They are used to communicate such non-verbal information as emotion and attitude, conversational signals for the functions of turn taking and feedback, and to enhance verbal prosodic signals.

Each gesture is defined in terms of a set of parameter tracks which can be invoked at any point in time, either during a period of silence between utterances or synchronised with an utterance. Several gestures can be executed in parallel. Articulatory movements created by the TTS will always supersede movements of the non-verbal gestures if there is a conflict. Scheduling and coordination of the gestures are controlled through a scripting language.

Having the agent augment the auditory speech with non-articulatory movements to enhance accentuation has been found to be very important in terms of the perceived responsiveness and believability of the system. The main guidelines for creating the prosodic gestures were to use a combination of head movements and eyebrow motion and to maintain a high level of variation between different utterances. To avoid gesture predictability and to obtain a more natural flow, we have tried to create subtle and varying cues employing a combination of head and eyebrow motion. A typical utterance from the agent can consist of either a raising of the eyebrows early in the sentence followed by a small vertical nod on a focal word or stressed syllable, or a small initial raising of the head followed by eyebrow motion on selected stressed syllables. A small tilting of the head forward or backward often highlights the end of a phrase. In the August system, we used a number of standard gestures with typically one or two eyebrow raises and some head

motion (video-clip C). The standard gestures work well with short system replies such as 'Yes, I believe so,' or 'Stockholm is more than 700 years old.'

For turn-taking issues, visual cues such as raising of the eyebrows and tilting of the head slightly at the end of question phrases were created. Visual cues are also used to further emphasise the message (e.g. showing directions by turning the head). To enhance the perceived responsiveness of the system, a set of listening gestures and thinking gestures was created. When a user is detected, by e.g the activation of a push-to-talk button, the agent immediately starts a randomly selected listening gesture, for example, raising the eyebrows. At the release of the push-to-talk button, the agent changes to a randomly selected thinking gesture like frowning or looking upwards with the eyes performing a searching gesture.

Our talking agent has been used in several different demonstrators with different agent appearances and characteristics. This technology has also been used in several applications representing various domains. An example from the actual use of the August agent, publicly displayed in the Cultural House in Stockholm (Gustafson *et al.*, 1999) can be seen on video-clip D. A multi-agent installation where the agents are given individual personalities is presently (2000–2001) part of an exhibit at the Museum of Science and Technology in Stockholm (video-clip E). An agent 'Urban' serving as a real estate agent is under development (Gustafson *et al.*, 2000) (video-clip F).

Finally, the use of animated talking agents as automatic language tutors is an interesting future application that puts heavy demands on the interactive behaviour of the agent (Beskow *et al.*, 2000). In this context, conversational signals not only facilitate the flow of the conversation but can also make the actual learning experience more efficient and enjoyable. One simulated example where stress placement is corrected, with and without prosodic and conversational gestures, can be seen on video-clip G.

Acknowledgements

The research reported here was carried out at CTT, the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations. We are grateful for having had the opportunity to discuss and develop this research within the framework of COST 258.

References

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Salvi, G., Spens, K.-E., and Öhman, T. (1999). Synthetic visual speech driven from auditory speech. *Proceedings of AVSP '99* (pp. 123–127). Santa Cruz, USA.

Badin, P., Bailly, G., and Boë, L.-J. (1998). Towards the use of a virtual talking head and of speech mapping tools for pronunciation training. *Proceedings of ESCA Workshop on Speech Technology in Language Learning (STiLL 98)* (pp. 167–170). Stockholm: KTH.

Beskow, J. (1995). Rule-based visual speech synthesis. *Proceedings of Eurospeech '95* (pp. 299–302). Madrid, Spain.

Beskow, J. (1997). Animation of talking agents. *Proceedings of AVSP '97, ESCA Workshop on Audio-Visual Speech Processing*. (pp. 149–152). Rhodes, Greece.

Beskow, J., Granström, B., House, D., and Lundeberg, M. (2000). Experiments with verbal and visual conversational signals for an automatic language tutor. *Proceedings of InSTiL 2000* (pp. 138–142). Dundee, Scotland.

Beskow, J. and Sjölander, K. (2000). WaveSurfer – a public domain speech tool. *Proceedings of ICSLP 2000*, Vol 4 (pp. 464–467). Beijing, China.

Bruce, G., Granström, B., and House, D. (1992). Prosodic phrasing in Swedish speech synthsis. In G. Bailly, C. Benoit, and T.R. Sawallis (eds), *Talking Machines: Theories, Models, and Designs* (pp. 113–125). Elsevier.

Carlson, R. and Granström, B. (1997). Speech synthesis. In W. Hardcastle and J. Laver (eds), *The Handbook of Phonetic Sciences*, (pp. 768–788). Blackwell Publishers Ltd.

Cassell, J. (2000). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds), *Embodied Conversational Agents* (pp. 1–27). The MIT Press.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. In H.T. Bunnell and W. Idsardi (eds), *Proceedings ICSLP 96* (pp. 2175–2178). Philadelphia.

Cole, R., Massaro, D.W., de Villiers, J., Rundle, B., Shobaki, K., Wouters, J., Cohen, M., Beskow, J., Stone, P., Connors, P., Tarachow, A., and Solcher, D. (1999). New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. *Proceedings of ESCA/Socrates Workshop on Method and Tool Innovations for Speech Science Education (MATISSE)* (pp. 45–52). University College London.

Ekman, P. (1979). About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepinies and D. Ploog (eds), *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium* (pp. 169–248). Cambridge University Press.

Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR 2/1989*, 1–80.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., and Wirén, M. (2000). AdApt – a multimodal conversational dialogue system in an apartment domain. *Proceedings of ICSLP 2000*. Vol. 2 (pp. 134–137). Beijing, China.

Gustafson, J., Lindberg, N., and Lundeberg, M. (1999). The August spoken dialogue system. *Proceedings of Eurospeech '99* (pp. 1151–1154). Budapest, Hungary.

Lundeberg, M. and Beskow, J. (1999). Developing a 3D-agent for the August dialogue system. *Proceedings of AVSP '99* (pp. 151–156). Santa Cruz, USA.

Massaro, D.W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. The MIT Press.

Massaro, D.W., Cohen, M.M., Beskow, J., and Cole, R.A. (2000). Developing and evaluating conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds), *Embodied Conversational Agents* (pp. 287–318). The MIT Press.

Neely, K.K. (1956). Effects of visual factors on intelligibility of speech. *Journal of the Acoustical Society of America*, *28*, 1276–1277.

Parke, F.I. (1982). Parameterized models for facial animation. *IEEE Computer Graphics*, *2(9)*, 61–68.

Pelachaud, C., Badler, N.I., and Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science*, *28*, 1–46.

Poggi, I. and Pelachaud, C. (2000). Performative facial expressions in animated faces. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds), *Embodied Conversational Agents* (pp. 155–188). The MIT Press.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

39

Interface Design for Speech Synthesis Systems

Gudrun Flach
Institute of Acoustics and Speech Communication, Dresden University of Technology
Mommsenstr. 13, 01069 Dresden, Germany
flach@eakss2.et.tu-dresden.de

Introduction

Today speech synthesis has become an increasingly important component of human–machine interfaces. For this reason, speech synthesis systems with different features are needed. Most systems offer internal control functions for varying speech parameters. These control functions are also needed by the developers of human–machine interfaces to realise suitable voice characteristics for different applications. The design of speech synthesis interfaces gives a cue to the control and can be realised in different ways, as shown in this contribution.

Description of Features for Speech Utterances

A speech synthesis system requires quite a number of control parameters. The primary parameters are those for physical and technical control of the variation of the speech quality. In J.E. Cahn (1990) we find a set of such control parameters, used for the generation of expressive voices by means of the synthesis system DECtalk3:

Accent Shape: pitch variation for accented words
Final Lowering: steepness of pitch fall at the end of an utterance
Pitch Range: difference (in Hz) between the highest and the lowest pitch value
Reference Line: reference value (default value) of the pitch
Speech Rate: number of syllables or words uttered per second (influences the duration of pauses and phoneme classes)
Average Pitch: average pitch value for the speaker
Breathiness: describes the aspiration noise in the speech signal
Brilliance: weakness or strengthening of the high frequencies (excited or calm voice)
Laryngealisation: degree of creaky voice
Loudness: speech signal amplitude and subglottal pressure

Contour Slope: general direction of the pitch contour (rising, falling, equal)
Fluent Pauses: pauses between intonation clauses
Hesitation Pauses: pauses in intonation clauses
Stress Frequency: frequency of word accents (pitch accents)
Pitch Discontinuity: the form of pitch changes (abrupt vs. smooth changes)
Pause Onset: smoothness of word ends and the start of the following pause
Precision: the range of articulation styles (slurry vs. exact articulation)

By means of combination and control of these parameters different types of voice characteristics (happy, angry, timid, sad, ...) could be realised.

Standard Software

Various industry-standard software interfaces (application programming interfaces, or APIs) define a set of methods for the integration of speech technology (speech recognition and speech synthesis). For instance we find special API-Standards for speech recognition and speech synthesis in the following packages:

- MS SAPI (Microsoft Speech API)
- SRAPI (Speech Recognition API)
- JSAPI (JAVA Speech API)
- SSIL (Arkenstone Speech Synthesizer Interface Library)
- S.100 R1.0 Media Services API specification.

These APIs specify methods for integrating speech technology in applications, typically written in C/C++, JAVA or Visual Basic (listed in Table 39.1). Table 39.1 shows some control functions used in Speech API's. The first column represents the function type, the second column shows selected values for this types and the third column gives an interpretation of the function type value. The investigation of theses standard software packages shows that the following control parameters are generally manipulated in speech synthesis systems:

- Pitch: values for average (or minimum and maximum) pitch

Table 39.1 Selected control functions in Speech APIs

device control	GetPitch/SetPitch	value of F0 baseline
	GetSpeed/SetSpeed	value of speech rate
	GetVolume/SetVolume	value of intensity
navigation	GetWord	position of the last spoken word
	WaitWord	position of word after speaking
	Pause	pauses the speech
	Resume	resumes the speech
	IsSpeaking	test of activity
lexicon handling	DlgLexicon	lexicon handling

- Speech rate: absolute speech rate in words (syllables) per second; increasing or decreasing of the current speech rate
- Volume: intensity in percentage of a reference value; increasing or decreasing the current volume value
- Intonation: the raise and fall of the declination line between phrase boundaries
- Control (Start, Pause, Resume, Stop): control over the state of the speech synthesis device
- Activity: status information on the internal conditions of the system
- Synchronization: the reading position in the text, for instance to synchronize multimedia applications
- Lexicon: the pronunciation lexicon for special tasks and/or user-defined pronunciation lexica
- Mode: reading mode (text, sentences, phrases, words or spelling)
- Voice: selection of a specific voice (male, female, child)
- Language: selection of a specific language, appropriate databases and processing models
- Text mode: selection of adapted intonation models for several text types (weather report, lyrics, addresses, ...).

Platform-Independent Standards

Text-to-Speech systems need information about the structure of the texts for a right pronunciation. The platform-independent interface standards provide a set of markers for the description of the text structure and for the control of the synthesisers. They are based on several kinds of mark-up languages. At the current time we find the following standards:

- SSML (Speech Synthesis Markup Language) (Taylor and Isard, 1996)
- STML (Spoken Text Markup Language) (Sproat *et al.*, 1997)
- JAVA™ Speech Markup Language (Java Speech Markup Language Specification, 1997)
- Extended Information (VERBMOBIL) (Helbig, 1997)

A small example for using SSML shows the principal possibilities:

<ssml>
<define word="holyrood" phonemes="h o1 l ii – r uu d">
<phrase><emph>I</emph>saw the man in the holyrood park</phrase>
<phrase>I saw the man in the park<phrase>with the telescope</phrase>
<voice name="male2">
<phrase>I saw the man in the park<phrase>with the telescope</phrase>
<sound src="laughter.au">
</ssml>

Some tags, like <ssml> and >/ssml>, are used as brackets for text portions. Definition tags, like <define word...> and information and action tags, like <sound src...>, define a sound source and initiate an action ('play this sound').

Table 39.2 Tags in the Extended Information Concept

tag	value	interpretation
voice	speaker	selects a voice
turn	number	pointer for navigation
utterance	begin/end	semantically completed unit
sentence	begin/end	sentence boundaries
ClauseType	quest/final	clause boundaries
PhonTrans	SAMPA	phonetic transcription
PhraseBound	b1 . . . b4	value for phrase bound weighting
Prominence	0 . . . 31	word weighting

The extended information concept contains similar tags, as shown in Table 39.2. This table gives an impression of the control facilities of the extended information concept. This concept gives cross-control parameters like the ones shown in the first column. The second column represents possible values for the control parameters and the third column gives an interpretation for the tag values.

In the framework of the mark-up languages we find a wide range of description possibilities. On the one hand, there are 'cross'-descriptors for general control like voice type or phrase boundary type, and, on the other, there are very detailed descriptors for the realisation type of the sounds in the given articulatory context, for instance.

Systems

For the definition of an interface standard we have investigated some speech synthesis systems (commercial and laboratory systems), represented on Internet:

- AntaresTM Centigram® TTS (TrueVoiceTM)
- Antares L&H TTS (tts2000/T)
- AntaresTM Telefónica TTS
- DECtalkTMPC2 Text-to-Speech Synthesizer
- DECtalk Express Speech Synthesizer
- DECtalk PC Software
- Bell Labs Text-to-Speech System
- INFOVOX 500, PC board
- INFOVOX 700
- TrueTalk
- ProVoice for Windows
- SoftVoice Text-to-Speech System
- ETI-ELOQUENCE
- WinSpeech 3.0(N)
- Clip&Talk 2.0
- EUROVOCS
- Festival
- ProVerbe Speech Engine (ELAN Informatique).

We compared the systems with regard to the description features for speech utterances mentioned in paragraph 1 and the control parameters of the standard software interfaces mentioned above. We found the following external system control parameters:

- Lexicon: Special dictionaries or user-defined pronunciation lexica can be selected.
- Rate: The speech rate can be changed, as measured in words or syllables per second.
- Pitch: The pitch of the actual voice can be defined or changed i.e., average pitch, or the lowest or highest value.
- Voice: A voice of a set of voices can be selected (i.e., 'abstract' voices in formant synthesis systems, like male-young, male-old, or 'concrete' voices in time-domain speech synthesis systems, like Jack or Jill).
- Mode: The reading mode can be defined (text, phrase, word, letter).
- Intensity: The intensity of the speech can be modified (mostly in the sense of loudness).
- Language: The speech system can synthesise more than one user-selectable language, with activation of appropriate databases and the processing algorithms.
- Pauses: The length of the different pauses (after phrases) can be defined.
- Navigation: Navigation in the text is possible (forward, backward, repeat), and in some cases, the position of the actual word is furnished for the synchronization of several processes.
- Punctuation: The system behaviour at punctuation marks can be defined (length of pauses, raising and falling of the pitch).
- Aspiration: The aspiration value of the voice can be specified.
- Intonation: A predefined intonation model can be selected.
- Vocal tract (formant synthesiser): Some parameters of the system's model of the vocal tract can be changed.
- Text mode: An appropriate intonation model can be selected for different types of text. The models include for instance special preprocessing algorithms, pronunciation dictionaries and intonation models.

Figures 39.1 and 39.2 show how many of the systems make a variation of the above mentioned parameters available.

Proposal of a Set of Control Parameters for TTS-Systems

The evaluation of these investigations suggests a tri-level interface of a speech synthesis system, consisting of global, physical and linguistic parameters.

Global Parameters

The global parameters describe voice, language and genre for the application. The system has a set of internal parameters and databases to realise the chosen global parameter. The user cannot vary the internal parameters shown in Table 39.3.

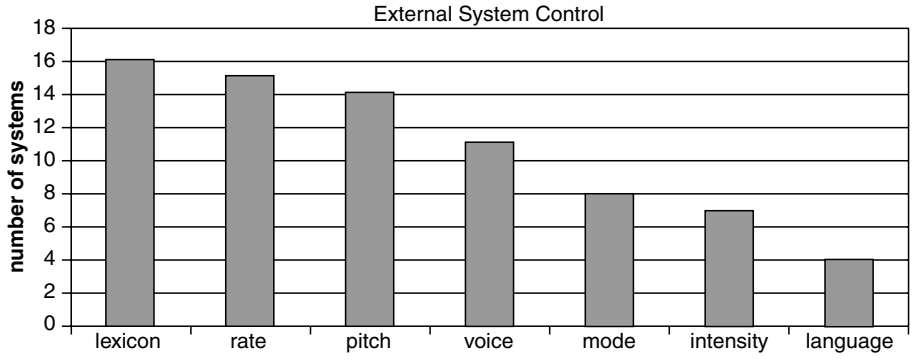


Figure 39.1 Part 1 of the external system control parameters

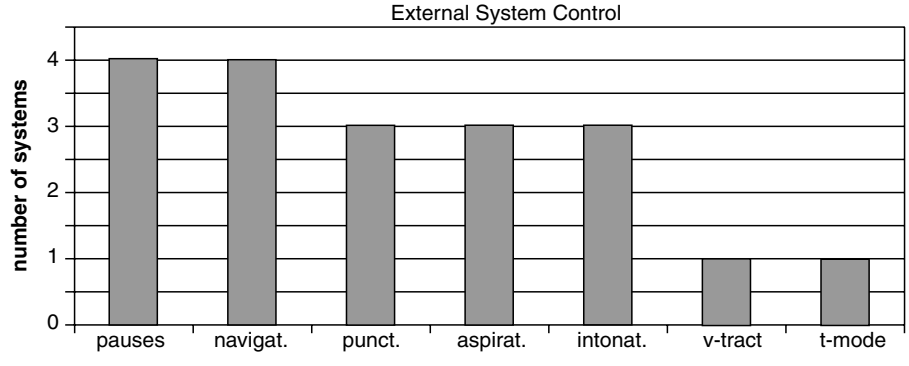


Figure 39.2 Part 2 of the external system control parameters

The parameters shown here describe the global behaviour of the synthesis system. A voice, a language or a genre can be chosen from a given range for this parameters.

Table 39.3 Global parameters

Parameter	Range	Example
Voice	Voice marker	Speaker name Young Female
Language	Language marker	English German Bavarian dialect
Genre	Genre marker	Weather report Lyrics List

Physical Parameters

The physical parameters (Table 39.4) describe the concrete behavior of the acoustic synthesis. For this description we need minimally values for the pitch and its variation range, the speech rate and the intensity. The word position is used for the multimedia synchronization of several applications. The speech mode controls the size of the synthesised phrases. For each speech mode we need specialised intonation models.

Linguistic Parameters

The linguistic parameters (Table 39.5) control the text preprocessing of the speech synthesis system. The application or user-defined pronunciation dictionary guarantees the right pronunciation of application-specific words, abbreviations or phrases. The punctuation level defines how the punctuation marks are realised (including pauses and pronunciation descriptions). The parameter text mode selects predefined preprocessor algorithms and intonation models for special kinds of text.

Table 39.4 Physical parameters

Parameter	Range	Interpretation
Pitch (average or lowest value)	60–100 Hz (male speaker)	index value
	100–140 Hz (female speaker)	
Pitch variation		lower and upper
average v.	1 Hz – 300 Hz	boundary for the
lowest v.	low val. – 300 Hz	pitch values
Speech rate	min: 75	words per
	max: 500	minute
Intensity	0–100	scale value
Word position	yes/no	used for synchronization in multimedia applications
Speech mode	text sentence word letter	utterance types

Table 39.5 Linguistic Parameters

Parameter	Range	Example/interpr.
Lexicon	lexicon marker	symbolic name file name
Punctuation level	punctuation characters	describes a pronunciation model
Text mode	standard mathematics	description of special prosodic
	list addresses	models

Conclusion

We have seen that for the practical use of speech synthesis technology, the specification of an interface standard is very important. Users are interested in developing a variety of applications via standard interfaces. For that reason, the developers of speech synthesis technology must develop complex internal controls for their devices by means of such interfaces. Current development in this area incorporates several strategies for the solution of this problem. The first key is the development of libraries that put special interface functions at the disposal of the user. The second strategy is to make available synthesis systems with simple interfaces that are used by the application. Via such interfaces, only a small set of parameters can be varied by the application. Of assistance are the mark-up languages for speech synthesis systems which allow the embedding of control information in symbolic form in the synthesis text.

Acknowledgements

This work was supported by the Deutsche Telekom BERKOM Berlin. I also want to extend my thanks to the organisation committee and all the participants of the COST 258 action for their interest and the fruitful discussions.

References

Cahn, J.E. (1990). *Generating Expression in Synthesized Speech. Technical Report*, M.I.T., Media Laboratory, Massachusetts Institute of Technology.

Helbig, J. (1997). Erweiterungsinformationen für die Sprachsynthese. *Digitale Erzeugung von Sprachsignalen zum Einsatz in Sprachsynthetisatoren, Anliegen und Ergebnisse des Projektes X243.2 im Rahmen der Deutsch-Tschechischen wissenschaftlich-technischen Zusammenarbeit*, TU Dresden, Fak. ET, ITA.

Sproat, R., Taylor, P.A., Tanenblatt, M., and Isard, A. (1997). A markup language for text-to-speech synthesis. *Proceedings Eurospeech 97*, Vol. 4 (pp. 1747–1750). Rhodes, Greece.

Sun Microsystems (1997). *Java Speech Markup Language Specification, Version 0.5*.

Taylor, P.A. and Isard, A. (1996). SSML: A speech synthesis markup language. *Speech Communication*, *21*, 123–133.

Improvements in Speech Synthesis. Edited by E. Keller et al.
Copyright © 2002 by John Wiley & Sons, Ltd
ISBNs: 0-471-49985-4 (Hardback); 0-470-84594-5 (Electronic)

Index

accent, 168, 170
accents, 207
accentual, 154, 155, 159
adaptation, 341, 342, 344
affect, 252
affective attributes, 256
Analysis-Modification-Synthesis Systems, 39
annotation, 339
aperiodic component, 25
arousal, 239
aspiration noise, 255
assessment, 40
assimilation, 228
automatic alignment, 322

Bark, 240
Baum-Welch iterations, 341, 342, 345, 346
benchmark, 40
boundaries, 205

Classification and Regression Tree, 339
concatenation points, smoothing of, 82
configuration model, 238
COST 249 reference system, 340
Cost 258 Signal Generation Test Array, 82
covariance model, 237
Czech, 129

dance, 155, 157, 158, 159
data-driven prosodic models, 176
deterministic/stochastic decomposition, 25
differentiated glottal flow, 274
diplophonia, 255
Discrete Cepstrum, 31
Discrete Cepstrum Transform, 30
distortion measures, 45
duration modeling, 340
  corpus based approach, 340
duration, 77, 129, 322, 323
durations, 154, 156, 159, 160, 161
Dutch, 204
dynamic time warping, 322

emotion, 253
emotions, 237
English, 204
enriched temporal representation, 163
evaluation, 46
excitation strength, 274

F0 global component, 147
F0 local components, 147
fast speech, 206
flexible prosodic models, 155
forced alignment mode, 340
formant waveforms, 34
formants, 77
formatted text, 309
French, 166, 167, 168, 170, 171, 174
Fundamental Frequency Models, 322
fundamental frequency, 77

Galician accent, 219
Galician corpus, 222
German, 166, 167, 168, 169, 170, 171, 174, 204
glottal parameters, 254, 274
glottal pulse skew, 275
glottal source variation voice quality, 253
glottal source variation
  cross-speaker, 280–2
  segmental, 275–8
  single speaker, 275–80
  suprasegmental, 279
glottal source, 253, 273
glottis closure instant, 77
gross error, 346

Hidden Markov Model, 220, 339, 340
HNM, 23
HTML, 317
hypoarticulation, 228

implications for speech synthesis, 232
intensity, 129

INTSINT, 323, 324
inverse filter, 77, 274
inverse filtering, 254, 274

KLSYN88a, 255

labelling word boundary strength, 179
labelling word prominence, 179
LaTeX, 317
lattice filter, 79
LF model, 254, 274
linear prediction, 77
linguistics
  convention, norms, 354–6
  framework, 355, 358
  patterns, 358
  semantics, 353
  social, 353, 354, 356
  structure, 353–62
  syntax, 354
lossless acoustic tube model, 80
low-sensitivity inverse filtering (LSIF), 80
LPC, 77
LPC residual signal, 77
LPC synthesis, 77
LP-PSOLA, 81
LTAS, 240

major prosodic group, 168, 171
mark-up language, 227
Mark-up, 297, 308
MATE Project, 299
MBROLA System, 301
Mbrola, 322
melodic, 155
modelling, 155, 160
minor prosodic group, 170, 171
Modulated LPC, 36
MOMEL, 322, 323, 324
monophone, 342
mood, 253
multilingual (language-independent) prosodic models, 176
music, 155, 157, 158, 159

nasalisation, 229
natural, 157, 161, 164
naturalness, 129

open quotient, 255, 275

pause, 168, 173
phonetic gestures, 228
phonetic segmentation and labelling, 177
phonetics, 166, 167, 168, 174
phonological level, 221
phonology, surface vs underlying, 321
phonostylistic variants, 131
physiological activation, 238
pragmatics
  making believed, 357, 358
  making heard, 353, 357
  making known, 353, 357, 358
  making understood, 353, 357
predicting phone duration, 176
predicting word boundary strength, 176
predicting word prominence, 176
principal component analysis, 47
PROSDATA, 340–7
prosodic mark-up, 311
prosodic modelling, 218
prosodic structure, 219
prosodic parameters, 129
prosodic transplantation, 42
prosody manipulation, 76
prosody, 154, 157, 204, 328, 334, 337
prosody, expressive, 304
prosody
  cohesive strength, 356
  F0, 353, 355, 357, 361, 362
  grouping function, 355
  implicit meaning, 359
  intonation, 355, 356, 357, 361
  melody, pitch, 355, 357, 360, 361, 362
  pitch range, ΔF0, F0 range, F0 excursion, 353, 355, 356, 360
ProSynth project, 302
ProZed, 324
punctuation mark, 166, 171
punctuation, 208

Reduction, 228
RELP, 77
representing rhythm, 155, 157
resistance to reduction and assimilatory effects, 231
retraining, 341
return phase, 274
rhythm rule, 209
rhythmic information, 159
rhythmic structure, 171

RTF, 317
Rules of Reduction and Assimilation, 234

SABLE Mark-Up, 300
segment duration, 168
segmentation, 340–6
  accuracy measure, 342, 346
  automatic segmentation, 342
shape invariance, 22
sinusoidal model, 23
slow speech, 204
source-filter model, 77
speaker characteristics, 76
speaking styles, 218
spectral tilt, 255
speech rate, 204
speech rhythm, 154, 155, 156, 157, 159, 161
speech segmentation, 328, 334, 335
speech synthesis, 155, 156, 163, 215, 328, 329, 333, 337, 354, 361, 362
speech synthesiser, 154, 156, 161
SpeechDat database, 340, 346
speed quotient, 255
SRELP, 77
SSABLE Mark-Up, 300
Standards, 308
stress, 154, 157, 166, 168, 170, 172, 173, 174
subjectivity
  belief, 358, 359, 361
  capture of meaning, appropriation, 354, 355, 356, 357, 361
  emotion, 353, 356, 359
  intention, 355
  interpretation, 356, 358, 360
  investment, 353, 355, 356, 358
  lexical, local, 353, 354, 355, 356, 357, 358, 360
  meaning, 354, 355, 358, 359, 361
  naturalness, 355, 360, 362
  personality, singularity, 353, 359, 361
  point of view, 354, 356, 357, 359, 362
  psychological, 353
  space, 354, 355, 356, 361
  speaker, 353, 354, 355, 356, 357, 358, 359, 360
  subjectivity, 353, 354, 355, 356, 357, 358, 359, 360, 361
Swiss High German, 165
syllable, 168, 169, 170, 171, 173, 174

Tags, 314
TD-PSOLA, 23, 76
Telephony Applications, 308
tempo, 156, 158, 159, 160, 161
temporal component, 154, 155, 157
temporal patterns, 156, 159, 160
temporal skeleton, 160, 161, 163
text types, 309
tied triphone, 342, 346
timing model, 166, 167, 168, 171, 174
ToBI, 323
tone of voice, 252

unit selection, 76
untied triphone, 342

valence, 237
Vector Quantization, 221
vioce quality, 82, 237
voice quality acoustic profiles, 253–5
voice source parameters, 255, 275
voice source, 253, 274
VoiceXML Mark-Up, 300

word, 168, 170, 173, 174

XML, 297, 317