

This book presents the development of a prosody system for text-to-speech (TTS) applications. The prosody is responsible for a communicative intention and guarantees some naturalness in the uttered speech. The prosodic features consist in the imposition of the timing, characterized by the segmental durations and pauses, the intonation, characterized by the fundamental frequency (F0) curve, and by the intensity curve. The proposed prosody model consists of several sub-models, namely, the duration model to predict the segmental durations and the model to predict the F0 pattern. The segmental durations model consists of one ANN carefully selected concerning its architecture and type as well as the set of input features with the objective of minimizing the error between predicted and measured durations. One alternative model, is based on same considerations but uses one dedicated ANN for each phoneme. The alternative model, with dedicated ANNs, improved the final performance. The proposed model to predict the F0 contour is based on the Fujisaki model and consists of two sub-models. One predicts the Phrase Commands parameters and the other predicts the Accent Commands parameters.

Prosody Generation Model for TTS Systems



João Paulo Teixeira

Prosody Generation Model for TTS Systems

Segmental Durations and F0 Contours with Fujisaki
Model



João Paulo Teixeira

Prof. João P. Teixeira obtained his PhD in Electrical and Computers Engineering at FEUP in 2004 and worked in several projects related with TTS and Prosody. He teaches at the IPB since 1995 mainly in the area of Signal Processing and is author of several scientific publications in journals and conference proceedings about signal processing and ANN.



978-3-659-16277-0

Teixeira

LAP **LAMBERT**
Academic Publishing

João Paulo Teixeira

Prosody Generation Model for TTS Systems

João Paulo Teixeira

**Prosody Generation Model for TTS
Systems**

**Segmental Durations and F0 Contours with Fujisaki
Model**

LAP LAMBERT Academic Publishing

Impressum / Imprint

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Alle in diesem Buch genannten Marken und Produktnamen unterliegen warenzeichen-, marken- oder patentrechtlichem Schutz bzw. sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Inhaber. Die Wiedergabe von Marken, Produktnamen, Gebrauchsnamen, Handelsnamen, Warenbezeichnungen u.s.w. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this works is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Coverbild / Cover image: www.ingimage.com

Verlag / Publisher:

LAP LAMBERT Academic Publishing

ist ein Imprint der / is a trademark of

AV Akademikerverlag GmbH & Co. KG

Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Germany

Email: info@lap-publishing.com

Herstellung in Deutschland (siehe letzte Seite) /

Printed in the U.S.A. / U.K. by (see last page)

ISBN: 978-3-659-16277-0

Zugl. / Approved by: Bragança, Polytechnic Institute of Bragança, Diss., 2012

Copyright © 2012 AV Akademikerverlag GmbH & Co. KG

Alle Rechte vorbehalten. / All rights reserved. Saarbrücken 2012

Table of Contents

Abbreviations	v
1 INTRODUCTION	1
1.1 Foreword	2
1.2 What Is This Book About?	4
1.3 Motivation and Objectives	6
1.4 FEUP TTS System for European Portuguese.....	9
1.4.1 Pre-processing of text module	11
1.4.2 Linguistic analysis	11
1.4.3 Phonetic transcription of text	12
1.4.4 Prosody pattern determination	12
1.4.5 Production of speech signal waveform	12
1.5 Organization Aspects of the Book	15
1.6 Original Contributions	17
2 PREPARATORY WORK	19
2.1 Introduction	20
2.2 Preliminary Prosodic Study of the Tonic Syllable	21
2.2.1 Introduction.....	21
2.2.2 Method	21
2.2.3 Analysis and results	23
2.2.4 Comments and conclusion	30
2.3 Speech Corpus - FEUP-IPB Database.....	32
2.3.1 Introduction.....	32
2.3.2 Speech corpus	33
2.3.3 Sound segmentation and labelling	33

2.3.4	Characteristics	36
2.3.5	Phonetic changing phenomena in database	40
2.3.6	Final remarks	43
2.4	Syllabification	44
2.4.1	Introduction	44
2.4.2	Syllable splitting of written text	47
2.4.3	Syllabic splitting of spoken text	50
2.4.4	Analysis and results	54
2.4.5	Conclusions	54
2.5	Phonetic Transcription from Text	56
2.5.1	Dedicated ANN to transcribe graphemes <a> and <e>	58
2.5.2	Rules to transcribe graphemes <a>, <e>, <o> and <x>	59
2.5.3	Co-articulation rules or post-lexical rules	67
2.5.4	Final remarks	69
3	DURATION MODEL	71
3.1	Introduction	72
3.2	Other Duration Models	74
3.2.1	The Klatt model	74
3.2.2	Sum-of-Products models	75
3.2.3	The Jan van Santen model	76
3.2.4	The Keller-Zellner algorithm	78
3.2.5	The Campbell model	81
3.2.6	The Barbosa-Bailly model – Inter-Perceptual-Centre-Groups	82
3.2.7	Model for the Hungarian language	85
3.2.8	Model for the Galician language	85
3.2.9	Model for the Castilian language	86
3.3	Duration Model for Standard European Portuguese	87
3.3.1	Considerations on the speech database	88
3.3.2	Network architecture	89
3.3.3	Neural network training	91
3.3.4	Features	96
3.4	Model Evaluation	108
3.4.1	Standard deviation (σ) or (std)	108

3.4.2	Mean absolute error (δ)	109
3.4.3	Linear correlation coefficient (r).....	109
3.4.4	Results and discussion	109
3.5	Alternative Model.....	118
3.5.1	Alternative model results	118
3.6	Pauses.....	123
3.6.1	Pause occurrence.....	123
3.6.2	Pause duration	125
3.6.3	Final considerations on studying pauses	127
3.7	Conclusion	129
4	FUNDAMENTAL FREQUENCY	131
4.1	Introduction	132
4.2	The Fujisaki Model	137
4.2.1	Phrase component	139
4.2.2	Accent component	141
4.3	Parameters Estimation of Fujisaki Model	144
4.3.1	Tool to support the manual estimation of Fujisaki model parameters	145
4.3.2	Parameters estimation process	149
4.3.3	Evaluation of the estimated F0 contour in the Database	152
4.4	Application of the Model.....	155
4.5	Phrase Commands.....	157
4.5.1	PC positions in text	158
4.5.2	Evaluation of preliminary inserted PC.....	163
4.5.3	Prediction of Ap and T0a parameters	167
4.5.4	Evaluation of the prediction of Ap and T0a	173
4.5.5	Results of the PC model.....	175
4.6	Accent Commands.....	177
4.6.1	ANN architectures	178
4.6.2	Training.....	180
4.6.3	Features	181

4.6.4	Results of prediction with ANNs	187
4.6.5	Results of AC model	195
4.7	Results of the Predicted F0 Contour.....	198
4.7.1	F0 model.....	198
4.7.2	F0 model over segmental durations	198
4.8	Conclusion	201
5	PERCEPTUAL TESTS	203
5.1	Introduction	204
5.2	Perceptual Test of Duration Models	205
5.2.1	Discussion	211
5.3	Perceptual Test of F0 Models	216
5.3.1	Discussion	226
5.4	Conclusions.....	231
6	FINAL CONCLUSIONS	233
6.1	General Observations about the Tasks	234
6.2	General Conclusions.....	236
6.2.1	Preparatory work	236
6.2.2	Timing	237
6.2.3	Fundamental frequency	241
6.2.4	Complete prosody model	243
6.3	Final Considerations about the Error Contributions	247
6.4	Resume of Results and Conclusions	249
6.5	Future Work	251
	BIBLIOGRAPHY	255

Abbreviations

Aa – Amplitude of AC;
ABU – Acoustic Building Unit;
AC – Accent Command;
ANN – Artificial Neural Network;
Ap – Magnitude of phrase command;
Ca – ANN that predicts the amplitude of the AC;
CA – ANN that predicts the existence of AC associated to the syllable;
EP – European Portuguese;
F0 – Fundamental frequency;
FEUP – Faculty of Engineer of University of Porto;
FEUP-TTS – FEUP Text-To-Speech system;
MOS – Mean Opinion Score;
PC – Phrase Command;
r – Linear correlation coefficient;
rmse – Root mean squared error;
std – Standard deviation;
T0 – Onset time of PC;
T0a – Anticipation of PC;
T0_E – Beginning of accent group where PC was inserted;
T1 – Onset time of AC;
T1a – Anticipation of the onset time of the AC;
T2 – Offset time of AC;
T2a – Anticipation of the offset time of the AC;
TPML – Text Processing Markup Language;
TTS – Text-To-Speech;
 δ – Mean absolute error;
 σ – Standard deviation.

1 Introduction

This introductory chapter makes a short overview of what is prosody and describes the motivations and objectives for this work. The FEUP-TTS system for European Portuguese, which will be, in first instance, the host of the proposed prosody model, is briefly described. Finally an overview of this document and a reference to the original contributions are made.