Lecture Notes in Computer Science 5967

Anna Esposito Nick Campbell Carl Vogel
Amir Hussain Anton Nijholt (Eds.)

Development of Multimodal Interfaces: Active Listening and Synchrony

Second COST 2102 International Training School
Dublin, Ireland, March 23-27, 2009
Revised Selected Papers

Springer

Volume Editors

Anna Esposito
Second University of Naples, and IIASS
Via Pellegrino 19
84019 Vietri sul Mare (SA), Italy
E-mail: iiass.annaesp@tin.it

Nick Campbell
Carl Vogel
The University of Dublin, Centre for Language and Communication Studies
Trinity College, Dublin 2, Ireland
E-mail: {nick; vogel}@tcd.ie

Amir Hussain
University of Stirling, Department of Computing Science & Mathematics
Stirling, FK9 4LA, Scotland, UK,
E-mail: ahu@cs.stir.ac.uk

Anton Nijholt
University of Twente, Faculty of Electrical Engineering,
Mathematics and Computer Science
P.O. Box 217, 7500 AE, Enschede, The Netherlands
E-mail: anijholt@cs.utwente.nl

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.5, H.4, H.5.2, H.3, I.4, I.2.10

LNCS Sublibrary: SL 3 – Information Systems and Application,
incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-642-12396-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-12396-2 Springer Berlin Heidelberg New York

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

This book is dedicated to beauty: we seek beauty in our scientific research, and we find our efforts justified when we catch glimpses during the pursuit

Author Index

Preface

This volume brings together, through a peer-revision process, the advanced research results obtained by the European COST Action 2102: Cross-Modal Analysis of Verbal and Nonverbal Communication, primarily discussed for the first time at the Second COST 2102 International Training School on "*Development of Multimodal Interfaces: Active Listening and Synchrony*" held in Dublin, Ireland, March 23–27 2009.

The school was sponsored by COST (European Cooperation in the Field of Scientific and Technical Research, www.cost.esf.org) in the domain of Information and Communication Technologies (ICT) for disseminating the advances of the research activities developed within the COST Action 2102: "Cross-Modal Analysis of Verbal and Nonverbal Communication" (cost2102.cs.stir.ac.uk)

COST Action 2102 in its third year of life brought together about 60 European and 6 overseas scientific laboratories whose aim is to develop interactive dialogue systems and intelligent virtual avatars graphically embodied in a 2D and/or 3D interactive virtual world, capable of interacting intelligently with the environment, other avatars, and particularly with human users.

The main focus of the school was the development of multimodal interfaces. Traditional approaches to multimodal interface design tend to assume a "ping-pong" or "push-to-talk" approach to speech interaction wherein either the system or the human interlocutor is active at any one time. This is contrary to many recent findings in conversation and discourse analysis, where the definition of a "turn" or even an "utterance" is found to be very complex. People don't "take turns" to talk in a typical conversational interaction, but they each contribute actively to the joint emergence of a "common understanding." The sub-theme of the school was "Synchrony and Active Listening" selected with the idea to identify contributions that actively give support to the ongoing research into the dynamics of human spoken interaction, to the production of multimodal conversation data and to the subsequent analysis and modelling of interaction dynamics, with the dual goal of appropriately designing multimodal interfaces, as well as providing new approaches and developmental paradigms.

The themes of the papers presented in this book emphasize theoretical and practical issues for modelling human–machine interaction, ranging from the attempt in describing "the spacing and orientation in co-present interaction" to the effort for developing multimodal interfaces, collecting and analyzing interaction data and emergent behavior as well as analyzing the use of nonverbal and pragmatic elements of exchanges, implementing discourse control and virtual agents and using active listening in computer speech processing. The papers included in this book benefited from the live interactions in person among the many participants of the successful meeting in Dublin. Over 100 established and apprenticing researchers converged for the event.

The editors would like to thank the ESF COST- ICT Programme for its support in the realization of the school and the publication of this volume, and in particular the COST Science Officers Gian Mario Maggio, Francesca Boscolo, Bernie O'Neill, and Matteo Razzanelli for their constant help, guidance, and encouragement and

Sietske Zeinstrafor for supporting and guiding the publication effort. That the event was successful owes this partly to more individuals than can be named, but notably: Marcus Furlong, Jean Maypother, Alena Moison. Our special appreciation goes to Gaetano Scarpetta, Dean of the International Institute for Advanced Scientific Studies, who supported and encouraged the editorial process including collecting, reviewing, and improving the manuscripts submitted. The IIASS team, Tina Marcella Nappi, Michele Donnarumma, and Antonio Natale, are acknowledged for their precious technical support in the organization of this volume.

In addition, the editors are grateful to the contributors for making this book a scientifically stimulating compilation of new and original ideas. Finally, the editors would like to express their greatest appreciation to all the members of the COST 2102 International Scientific Committee for their rigorous and invaluable scientific revisions, for their dedication, and their priceless selection process.

Anna Esposito
Nick Campbell
Carl Vogel
Amir Hussain
Anton Nijholt

Organization

International Advisory and Organizing Committee

Anna Esposito	Second University of Naples and IIASS, Italy
Nick Campbell	Trinity College Dublin, Dublin 2, Ireland
Carl Vogel	Trinity College Dublin, Dublin 2, Ireland
Marcos Faundez-Zanuy	Escola Universitaria Politecnica de Mataro, Spain
Amir Hussain	University of Stirling, UK
Anton Nijholt	University of Twente, The Netherlands

International Scientific Committee

Uwe Altmann	Technische Universität Dresden, Germany
Hicham Atassi	Brno University of Technology, Czech Republic
Nikos Avouris	University of Patras, Greece
Ruth Bahr	University of South Florida, USA
Gérard Bailly ICP	Grenoble, France
Marian Bartlett	University of California, San Diego,USA
Štefan Beňuš	Constantine the Philosopher University, Nitra, Slovakia
Niels Ole Bernsen	University of Southern Denmark, Denmark
Jonas Beskow	Royal Institute of Technology, Sweden
Horst Bishof	Technical University Graz, Austria
Peter Birkholz	Aachen University, Germany
Jean-Francois Bonastre	Universitè d'Avignon, France
Nikolaos Bourbakis	ITRI, Wright State University, Dayton, USA
Maja Bratanić	University of Zagreb, Croatia
Antonio Calabrese	Istituto di Cibernetica – CNR, Naples, Italy
Eric Cambria	University of Stirling, Scotland, UK
Paola Campadelli	Università di Milano, Italy
Nick Campbell	Trinity College Dublin, Dublin 2, Ireland
Antonio Castro Fonseca	Universidade de Coimbra, Portugal
Aleksandra Cerekovic	Faculty of Electrical Engineering, Croatia
Josef Chaloupka	Technical University of Liberec, Czech Republic
Mohamed Chetouani	Universitè Pierre et Marie Curie, France
Gerard Chollet	CNRS-LTCI, Paris, France
Muzeyyen Ciyiltepe	Gulhane Askeri Tip Academisi, Ankara, Turkey
Anton Čižmár	Technical University of Košice, Slovakia
Nicholas Costen	Manchester Metropolitan University, UK
Vlado Delic	University of Novi Sad, Serbia
Marion Dohen	ICP, Grenoble, France
Giuseppe Di Maio	Second University of Naples, Italy
Francesca D'Olimpio	Second University of Naples, Italy

Thierry Dutoit	Faculté Polytechnique de Mons, Belgium
Laila Dybkjær	University of Southern Denmark, Denmark
Jens Edlund	Royal Institute of Technology, Sweden
Matthias Eichner	Technische Universität Dresden, Germany
Aly El-Bahrawy	Faculty of Engineering, Cairo, Egypt
Engin Erzin	Koc University, Istanbul, Turkey
Anna Esposito	Second University of Naples, and IIASS, Italy
Joan Fàbregas Peinado	Escola Universitaria de Mataro, Spain
Sascha Fagel	Technische Universität Berlin, Germany
Nikos Fakotakis	University of Patras, Greece
Marcos Faundez-Zanuy	Escola Universitaria de Mataro, Spain
Dilek Fidan	Ankara University, Turkey
Leopoldina Fortunati	Università di Udine, Italy
Carmen García-Mateo	University of Vigo, Spain
Björn Granström	Royal Institute of Technology, KTH, Sweden
Marco Grassi	Università Politecnica delle Marche, Italy
Maurice Grinberg	New Bulgarian University, Bulgaria
Mohand-Said Hacid	Universitè Claude Bernard Lyon 1, France
Jaakko Hakulinen	University of Tampere, Finland
Ioannis Hatzilygeroudis	University of Patras, Greece
Immaculada Hernaez	University of the Basque Country, Spain
Javier Hernando	Technical University of Catalonia, Spain
Wolfgang Hess	Universität Bonn, Germany
Dirk Heylen	University of Twente, The Netherlands
Rüdiger Hoffmann	Technische Universität Dresden, Germany
David House	Royal Institute of Technology (KTH), Sweden
Amir Hussain	University of Stirling, UK
Ewa Jarmolowicz	Adam Mickiewicz University, Poznan, Poland
Kristiina Jokinen	University of Helsinki, Finland
Jozef Juhár	Technical University Košice, Slovak Republic
Zdravko Kacic	University of Maribor, Slovenia
Bridget Kane	Trinity College Dublin, Ireland
Maciej Karpinski	Adam Mickiewicz University, Poznan, Poland
Eric Keller	Université de Lausanne, Switzerland
Adam Kendon	University of Pennsylvania, USA
Stefan Kopp	University of Bielefeld, Germany
Jacques Koreman	University of Science and Technology, Norway
Robert Krauss	Columbia University, New York, USA
Maria Koutsombogera	Institute for Language and Speech Processing, Greece
Bernd Kröger	Aachen University, Germany
Gernot Kubin	Graz University of Technology, Austria
Alida Labella	Second University of Naples, Italy
Yiannis Laouris	Cyprus Neuroscience and Technology Institute, Cyprus
Børge Lindberg	Aalborg University, Denmark
Saturnino Luz	Trinity College Dublin, Ireland

Wojciech Majewski	Wroclaw University of Technology, Poland
Pantelis Makris	Neuroscience and Technology Institute, Cyprus
Maria Marinaro	Salerno University and IIASS, Italy
Raffaele Martone	Second University of Naples, Italy
Dominic Massaro	University of California - Santa Cruz, USA
Olimpia Matarazzo	Second University of Naples, Italy
David McNeill	University of Chicago, USA
Nicola Melone	Second University of Naples, Italy
Katya Mihaylova	University of National and World Economy, Sofia, Bulgaria
Michal Mirilovič	Technical University of Košice, Slovakia
Peter Murphy	University of Limerick, Ireland
Antonio Natale	Salerno University and IIASS, Italy
Eva Navas	Escuela Superior de Ingenieros, Bilbao, Spain
Delroy Nelson	University College London, UK
Géza Németh	Budapest University of Technology, Hungary
Friedrich Neubarth	Research Institute for Artificial Intelligence, Austria
Giovanna Nigro	Second University of Naples, Italy
Anton Nijholt	University of Twente, The Netherlands
Jan Nouza	Technical University of Liberec, Czech Republic
Igor Pandzic	Faculty of Electrical Engineering, Croatia
Harris Papageorgiou	Institute for Language and Speech Processing, Greece
Ana Pavia	Spoken Language Systems Laboratory, Portugal
Catherine Pelachaud	Université de Paris 8, France
Paolo Pedone	Second University of Naples, Italy
Bojan Petek	University of Ljubljana, Slovenia
Harmut R. Pfitzinger	University of Munich, Germany
Francesco Piazza	Università Politecnica delle Marche, Italy
Neda Pintaric	University of Zagreb, Croatia
Isabella Poggi	Università di Roma 3, Italy
Ken Prepin	LTCI lab, CNRS, Telecom-ParisTech, France
Jiří Přibil	Academy of Sciences, Czech Republic
Anna Přibilová	Slovak University of Technology, Slovakia
Michael Pucher	Telecommunications Research Center Vienna, Austria
Jurate Puniene	Kaunas University of Technology, Lithuania
Roxanne Raine	University of Twente, The Netherlands
Giuliana Ramella	Istituto di Cibernetica – CNR, Naples, Italy
Kari-Jouko Räihä	University of Tampere, Finland
José Rebelo	Universidade de Coimbra, Portugal
Luigi Maria Ricciardi	Università di Napoli "Federico II", Italy
Maria Teresa Riviello	Second University of Naples, and IIASS, Italy
Matej Rojc	University of Maribor, Slovenia
Algimantas Rudzionis	Kaunas University of Technology, Lithuania
Vytautas Rudzionis	Kaunas University of Technology, Lithuania
Milan Rusko	Slovak Academy of Sciences, Slovak Republic

Zsófia Ruttkay	Pazmany Peter Catholic University, Hungary
Bartolomeo Sapio	Fondazione Ugo Bordoni, Roma, Italy
Yoshinori Sagisaka	Waseda University, Tokyo, Japan
Gaetano Scarpetta	University of Salerno and IIASS, Italy
Silvia Scarpetta	Salerno University, Italy
Ralph Schnitker	Aachen University, Germany
Jean Schoentgen	Université Libre de Bruxelles, Belgium
Stefanie Shattuck-Hufnagel	MIT, Cambridge, USA
Zdeněk Smékal	Brno University of Technology, Czech Republic
Stefano Squartini	Università Politecnica delle Marche, Italy
Piotr Staroniewicz	Wroclaw University of Technology, Poland
Vojtěch Stejskal	Brno University of Technology, Czech Republic
Marian Stewart-Bartlett	University of California, San Diego,USA
Jing Su	Trinity College Dublin, Ireland
Jianhua Tao	Chinese Academy of Sciences, P. R. China
Jure F.Tasič	University of Ljubljana, Slovenia
Murat Tekalp	Koc University, Istanbul, Turkey
Kristinn Thórisson	Reykjavík University, Iceland
Isabel Trancoso	Spoken Language Systems Laboratory, Portugal
Luigi Trojano	Second University of Naples, Italy
Wolfgang Tschacher	University of Bern, Switzerland
Markku Turunen	University of Tampere, Finland
Henk Van Den Heuvel	Radboud University Nijmegen,The Netherlands
Robert Vích	Academy of Sciences, Czech Republic
Klára Vicsi	Budapest University of Technology, Hungary
Leticia Vicente-Rasoamalala	Alchi Prefectural Univesity, Japan
Hannes Högni Vilhjálmsson	Reykjavík University, Iceland
Jane Vincent	University of Surrey, Guilford, UK
Carl Vogel	Trinity College Dublin, Dublin 2, Ireland
Jan Volín	Charles University, Czech Republic
Rosa Volpe	Université De Perpignan Via Domitia, France
Yorick Wilks	University of Sheffield, UK
Matthias Wimmer	Technische Universiät München, Germany
Matthias Wolf	Technische Universität Dresden, Germany
Bencie Woll	University College London, UK
Bayya Yegnanarayana	Institute of Information Technology, India
Jerneja Žganec Gros, Alpineon	Development and Research, Slovenia
Goranka Zoric	Faculty of Electrical Engineering, Croatia

Sponsors

The following organizations sponsored and supported the International Training School:

- European COST Action 2102 "*Cross-Modal Analysis of Verbal and Nonverbal Communication*" (cost2102.cs.stir.ac.uk)
- Second University of Naples, Faculty and Department of Psychology, Faculty of Science, Mathematics, and Physics and Department of Computer Engineering & Informatics, Italy
- International Institute for Advanced Scientific Studies "E.R. Caianiello" (IIASS), Italy
- Regione Campania (Italy)
- Provincia di Salerno (Italy)
- The School of Computer Science and Statistics, Trinity College Dublin
- Science Foundation Ireland (RFP 05/RF/CMS002)

Table of Contents

Microintonation Analysis of Emotional Speech

Jiří Přibil[1,2] and Anna Přibilová[3]

[1] Institute of Photonics and Electronics, Academy of Sciences CR, v.v.i.,
Chaberská 57, CZ-182 51 Prague 8, Czech Republic
[2] Institute of Measurement Science, SAS,
Dúbravská cesta 9, SK-841 04 Bratislava, Slovakia
`Jiri.Pribil@savba.sk`
[3] Slovak University of Technology, Faculty of Electrical Engineering & Information
Technology, Dept. of Radio Electronics, Ilkovičova 3, SK-812 19 Bratislava, Slovakia
`Anna.Pribilova@stuba.sk`

Abstract. The paper addresses reflection of microintonation in male and female acted emotional speech. Microintonation component of speech melody is analyzed regarding its spectral and statistical parameters. Achieved statistical results of microintonation analysis show good correlation comparing male and female voices for four emotional states (joy, sadness, anger, neutral state) portrayed by several professional actors.

Keywords: Speech analysis, emotional speech, jitter, microintonation.

1 Introduction

Emotional speech is characterized by prosodic features (F0, energy, duration) and several voice quality features (e.g. jitter, shimmer, glottal-to-noise excitation ratio, Hammarberg index) [1-4]. The voice quality parameter "jitter" describes pitch perturbations in the context of vocal expression. There exist different approaches to define vocal jitter: majority of authors use definitions resulting from perturbation in pitch period [1, 5-11], some authors define jitter as pitch frequency perturbation [12, 13]. According to [14] jitter is difficult to manipulate for actors and there is only tendency for anger portrayals to show more jitter than sadness portrayals. On the other hand, in [15] an example is reported that a speaker may increase F0 jitter for "happiness" rather than increasing the overall pitch level. For similar perturbations in the context of music performance the term "microintonation" is used [14]. However, phoneticians use also the term "microintonation" or "micromelody" describing local changes in F0 as a part of microprosody [16]. Strictly speaking, "microprosody" comprises pitch, duration, and energy short-term localized changes [17].

In our present work we analyze microintonation of male and female emotional speech representing joy, sadness, anger, and a neutral state. Our approach to microintonation estimation is somewhat similar to that of [18] where jitter related to microvariations of a pitch curve is computed as a relative number of zero crossings of a derivative pitch curve normalized by utterance duration.

A. Esposito et al. (Eds.): COST 2102 Int. Training School 2009, LNCS 5967, pp. 268–279, 2010.

Our research aimed at microintonation analysis was motivated by requirement of higher naturalness of synthetic speech with expression of emotions. We want to implement microintonation into a pitch-synchronous cepstral speech synthesizer by superimposing a small random variable to each pitch period. In this way the effect of jitter can be modeled in the voiced parts of the synthesized speech. Randomization will be controlled by statistical results of analyzed emotional speech corresponding to different emotions. This modification can be applied directly to the text-to-speech (TTS) system enabling expressive speech production [19].

On the other hand, we need to remove the microintonation component before creating a database of emotional prosodic prototypes for emotional speaking style conversion using a similar TTS system [20]. When compared with other speech melody components, the microintonation signal is a high-frequency signal and it can be filtered out of the melody contour and processed separately. Microintonation statistical and spectral analysis for several speakers expressing several emotions will be further used to synthesize a FIR filter suppressing the microintonation part of the speech signal.

2 Subject and Method

Microintonation, together with sentence melody and word melody, represents melody of speech given by F0 contour. Microintonation component of speech melody can be supposed to be a random, band-pass signal that can be described by its spectrum and statistical parameters. Fig. 1 shows the block diagram of our speech processing method of microintonation analysis.

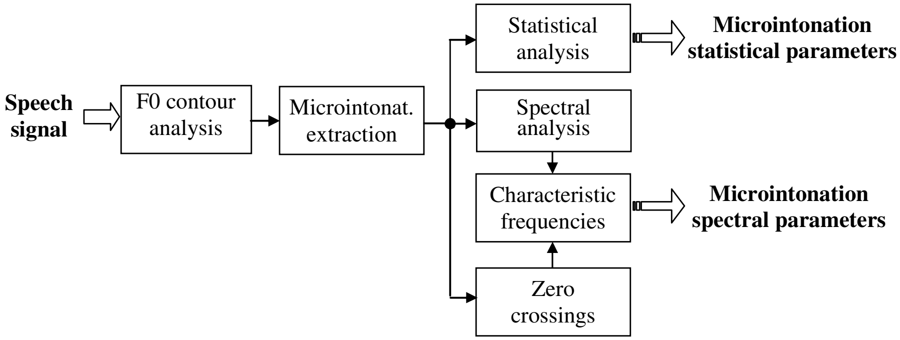

Fig. 1. Block diagram of microintonation statistical and spectral parameters estimation

Speech frames classified as voiced are analyzed separately depending on emotional state (joyous, sad, angry, and neutral) and voice type (male, female). The whole microintonation analysis procedure is divided into four phases:

1. Determination of F0 values, definition of the voiced and unvoiced parts of the processed speech signal.
2. F0 contour analysis, microintonation extraction, calculation of zero crossing parameters, determination of pitch periods and jitter calculation (for comparison with microintonation values) in the voiced parts of the speech signal.

3. Microintonation and zero crossing statistical analysis of the concatenated signal (see Fig. 2 and Fig. 4).
4. Microintonation signal spectral analysis and 3-dB bandwidth (B_3) determination from the concatenated signal (see Fig. 3 and Fig. 5).

The introductory microintonation processing phase consists of the following steps:

1. Determination of the melody contours from the voiced parts of speech smoothed by a median filter.
2. Determination of $\mathrm{F0_{mean}}$ values and calculation of the *linear trend* (LT) by the mean square method.
3. Calculation of differential microintonation signal $\mathrm{F0_{DIFF}}$ by subtraction of these values from the corresponding F0 contours ($\mathrm{F0_{mean}}$ and LT removal)

$$F0_{DIFF}(n) = \left(F0(n) - F0_{Mean}\right) - LT(n)\ . \tag{1}$$

4. Calculation of the absolute jitter J_{Abs} values, as the average absolute difference between consecutive pitch periods L measured in samples [10]

$$J_{Abs} = \frac{1}{f_s\left(N_L - 1\right)} \sum_{n=1}^{N_L - 1} \left| L_n - L_{n+1} \right|, \tag{2}$$

   where f_{s} is the sampling frequency and N_L is the number of extracted pitch periods.
5. Detection of zero crossings, calculation of zero crossing periods L_{Z}.

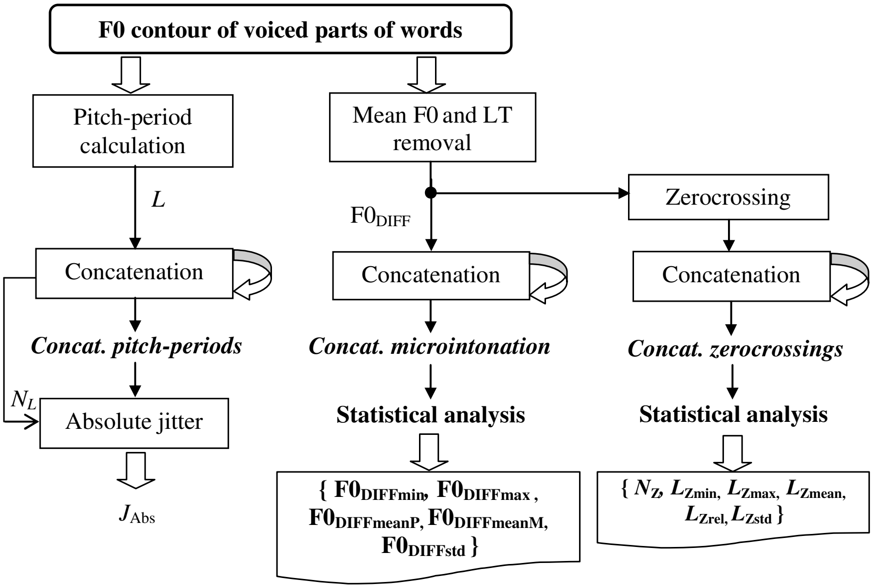

Fig. 2. Block diagrams of microintonation signal analysis: basic and zero crossing statistical analysis, absolute jitter calculation

Basic statistical analysis phase is performed in tree steps:

1. Statistical analysis of microintonation signal: minimum, maximum, and standard deviation (mean value of microintonation signal approaches to zero). For both positive and negative microintonation values the mean parameters are determined.
2. Statistical analysis of zero crossing periods: minimum, maximum, mean values, standard deviation, and a relative value defined as $L_{Zrel} = N_F / N_Z$ - where N_Z is the total number of zero crossings in each of the four emotions, and N_F is the total number of voiced frames.
3. Calculation and building of histograms from zero crossing periods L_Z for each of the emotion groups and both voices. Subjective evaluation by visual comparison of histograms and objective evaluation by hypothesis tests of distributions and analysis of variance (ANOVA) with multiple comparison of groups.

Spectral analysis of concatenated differential microintonation signal is also carried out for all emotions. This analysis phase is divided into three steps (see Fig. 3):

1. Calculation of the frequency parameters from the zero crossing periods $L_{Zx} = \{L_{Z\min}, L_{Z\max}, L_{Z\text{mean}}, L_{Z\text{rel}}, L_{Z\text{std}}\}$ as $F_{Zxl} = f_F / (2 \,.\, L_{Zx})$, where f_F is the frame frequency.
2. Microintonation signal spectral analysis by periodogram averaging using the Welch method.
3. Determination of B_3 values from these spectra for each of the emotion types.

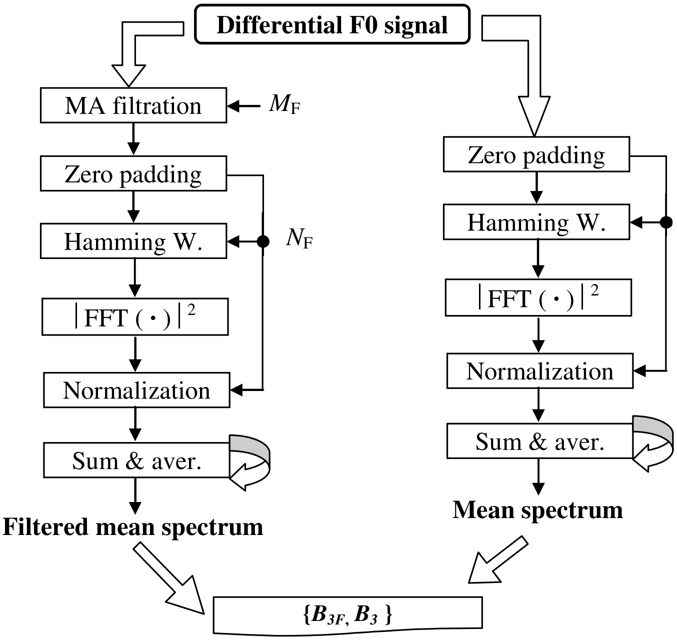

Fig. 3. Block diagrams of microintonation signal spectral analysis

To obtain spectrum of smoothed microintonation signal (see Fig. 5b), the concatenated differential F0 signal is filtered by a moving average filter of the length M_F (Voiced parts shorter than M_F+2 frames are not processed in further analysis).

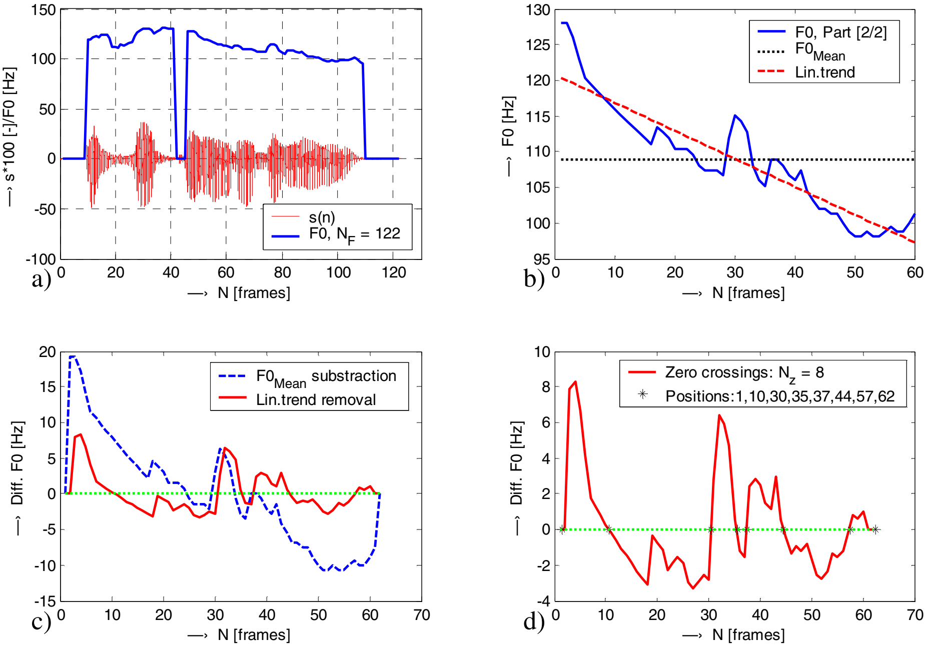

Fig. 4. Demonstration of microintonation analysis: speech signal with F0 contour (a), the second voiced part: original F0, mean F0, and LT (b), differential signal after $F0_{mean}$ and LT subtraction (c), zero crossings of differential F0 signal (d) – the sentence "Řekl Radomil" ("Radomil said") uttered in sad emotional style by a male Czech speaker.

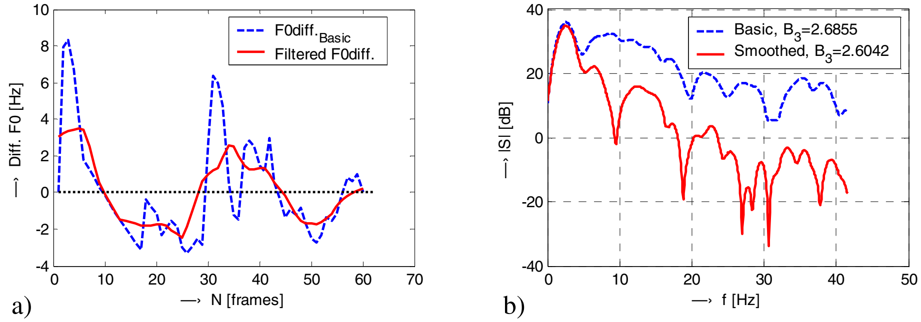

Fig. 5. Demonstration of microintonation smoothing and spectrum determination (obtained from the same sentence's second voiced part as in Fig. 4): basic differential F0 signal and the one filtered by moving average (a), corresponding spectra and their 3-dB bandwidths B3 (b).

3 Material, Experiments, and Results

Speech material for microintonation analysis was collected in two databases (male and female voices – 132 sentences, 8+8 speakers altogether) consisting of neutral and emotional sentences uttered by several speakers (extracted from the Czech and Slovak stories uttered by professional actors). Classification of emotional states was carried out manually, by subjective listening method with the help of distributed listening test program [21]. Each of the extracted sentences was evaluated by a small group of listeners. Every sentence having been evaluated as expressing the same emotional state by all listeners was added to our database.

The frame length depends on the mean pitch period of the processed signal. In our experiment, we had chosen overlapped 24-ms frames in 12-ms intervals for male voice, and 16-ms frames in 8-ms intervals for female voice. It corresponds to the frame frequency $f_F = 83.3$ Hz for males, and $f_F = 125$ Hz for females. The typical mean pitch-period length in samples of neutral emotional style, male voice is $L \approx 140$ which corresponds to 8.75 ms for $f_s = 16$ kHz, in the case of female voice it is $L \approx 80$ samples, corresponding to 5 ms.

Pitch contours were given with the help of the PRAAT [1] program [22]. The minimum length of the processed voiced parts was experimentally set to 10 frames and the corresponding filter length of $M_F = 8$ was chosen. Number of analyzed voiced parts / voiced frames) was:

- neutral: 112/2698, joy: 79/1927, sadness: 128/3642, anger: 104/ 2391 – Male.
- neutral: 86/2333, joy: 87/2541, sadness: 92/2203, anger: 91/2349 – Female.

Results of basic statistical microintonation analysis in comparison with absolute jitter values for all four emotional states are summarized in Tab. 1 (male voice) and Tab. 2 (female voice). Results of performed zero crossing analysis for male / female voices are shown in Tab. 3 / Tab. 4. Summary histograms of zero crossing periods L_Z for different emotions in dependence on the speaker gender are shown in Fig. 7 and Fig. 8.

For objective statistical comparison of zero crossing periods L_Z the Ansari-Bradley test [24] was performed. If is the test of the hypothesis that two independent samples come from the same distribution against the alternative that they come from distributions having the same median and shape but different variances. The result is $h = 0$ if the null hypothesis of identical distributions cannot be rejected at the 5% significance level, or $h = 1$ if the null hypothesis can be rejected at the 5% level. This test also returns the probability of observing the given result, or one more extreme by chance if the null hypothesis is true. Small values of this probability cast doubt on the validity of the null hypothesis.

The results of the hypothesis tests based on comparison of distributions are presented in Tab. 5 for the male voice group, in Tab. 6 for the female voice group, and in Tab. 7 for both voices. The second approach based on ANOVA was applied to zero

---

[1] The PRAAT internal settings for F0 values determination were experimentally chosen by visual comparison of testing sentences (one typical sentence from each of emotions and voice classes) as follows: cross-correlation analysis method [23], pitch-range 35÷250 Hz for male and 105÷350 Hz for female voices.

crossing periods L_Z together with multiple comparison test. Fig. 9 shows the graph with each group mean represented by a symbol and an interval around the symbol. Two means are significantly different if their intervals are disjoint, and are not significantly different if their intervals overlap.

Table 1. Summary results of microintonation basic statistical analysis (differential F0 parameters in [Hz]) together with absolute jitter values (in [ms]) – male voice

Emotion	$\mathbf{F0_{DIFFmin}}$	$\mathbf{F0_{DIFFmax}}$	$\mathbf{F0_{DIFFmeanP}}^{*)}$	$\mathbf{F0_{DIFFmeanN}}^{**)}$	$\mathbf{F0_{DIFFstd}}$	J_{Abs}
Neutral	-10.57	13.18	2.66	-3.07	3.92	0.2908
Joy	-31.76	37.62	7.27	-7.22	9.71	0.7057
Sadness	-32.97	25.71	4.02	-3.95	5.57	0.4485
Anger	-46.32	45.17	8.62	-8.89	10.23	0.6014

*) calculated from positive microintonation values.

**) calculated from negative microintonation values.

Table 2. Summary results of microintonation basic statistical analysis and jitter values – female voice

Emotion	$\mathbf{F0_{DIFFmin}}$	$\mathbf{F0_{DIFFmax}}$	$\mathbf{F0_{DIFFmeanP}}^{*)}$	$\mathbf{F0_{DIFFmeanN}}^{**)}$	$\mathbf{F0_{DIFFstd}}$	J_{Abs}
Neutral	-44.39	38.35	4.11	-5.46	7.06	0.1673
Joy	-40.87	43.55	11.85	-10.78	14.39	0.3329
Sadness	-54.28	50.66	7.05	-7.32	10.29	0.2354
Anger	-44.35	42.95	10.16	-10.16	13.07	0.3450

*) calculated from positive microintonation values.

**) calculated from negative microintonation values.

Table 3. Summary results of zero crossing analysis (zero crossing period L_Z parameters in [frames]) – male voice

Emotion	N_Z	$L_{Zmax}^{*)}$	L_{Zmean}	L_{Zstd}
Neutral	592	26	6.04	4.19
Joy	403	59	8.26	6.52
Sadness	681	57	6.82	5.69
Anger	521	23	6.74	4.57

*) $L_{Zmin} = 1$.

Table 4. Summary results of zero crossing analysis – female voice

Emotion	N_Z	$L_{Zmax}^{*)}$	L_{Zmean}	L_{Zstd}
Neutral	636	27	4.77	3.79
Joy	448	36	7.68	5.44
Sadness	680	40	5.41	4.81
Anger	381	23	5.83	3.66

*) $L_{Zmin} = 1$.

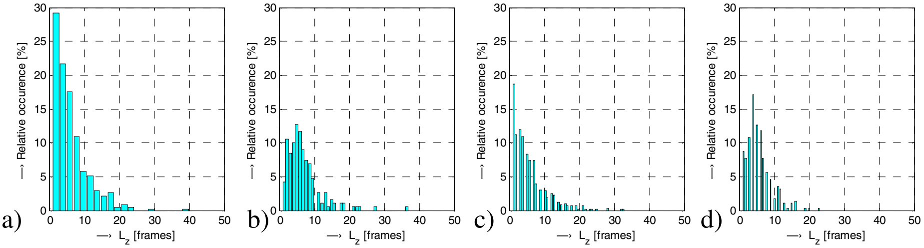

Fig. 7. Histograms of zero crossing periods L_Z calculated from differential F0 signal: "neutral" style (a), and emotions - "joy" (b), "sadness" (c), and "anger" (d) – male voice

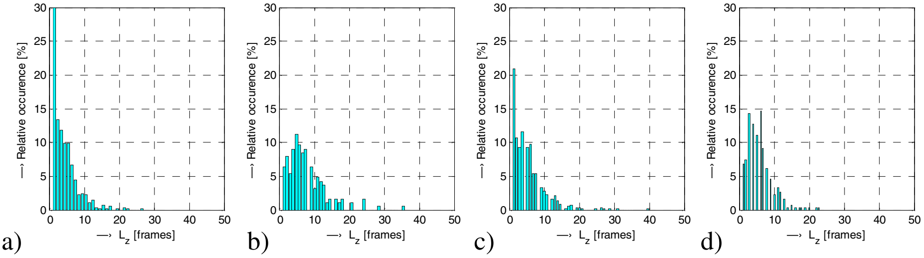

Fig. 8. Histograms of zero crossing periods L_Z calculated from differential F0 signal: "neutral" style (a), and emotions - "joy" (b), "sadness" (c), and "anger" (d) – female voice

Table 5. Partial results of zero crossing periods L_Z Ansari-Bradley hypothesis tests based on comparison of distribution – <u>male</u> voice group

*)	**Neutral**	**Joy**	**Sadness**	**Anger**
Neutral	**0/1**	1 / 3.37 10^{-7}	1 / 0.035	1 / 0.066
Joy		**0/1**	1 / 8.99 10^{-7}	1 / 0.006
Sadness			**0/1**	1 / 0.021
Anger				**0/1**

*) null hypothesis / probability values for 5% significance level.

Table 6. Partial results of zero crossing periods L_Z hypothesis tests – <u>female</u> voice group

*)	**Neutral**	**Joy**	**Sadness**	**Anger**
Neutral	**0/1**	1 / 4.8810^{-15}	1 / 0.006	1 / 0.017
Joy		**0/1**	1 / 0.002	1 / $4.01.10^{-8}$
Sadness			**0/1**	1 / 0.002
Anger				**0/1**

*) null hypothesis / probability values for 5% significance level.

Table 7. Summary results of zero crossing periods L_Z hypothesis tests (comparison <u>male</u> vs. <u>female</u> voice group) between particular emotions

	Neutral	**Joy**	**Sadness**	**Anger**
h / p*)	0 / 0.4397	0 / 0.8926	0 / 0.6953	0 / 0. 5773

*) null hypothesis / probability values for 5% significance level.

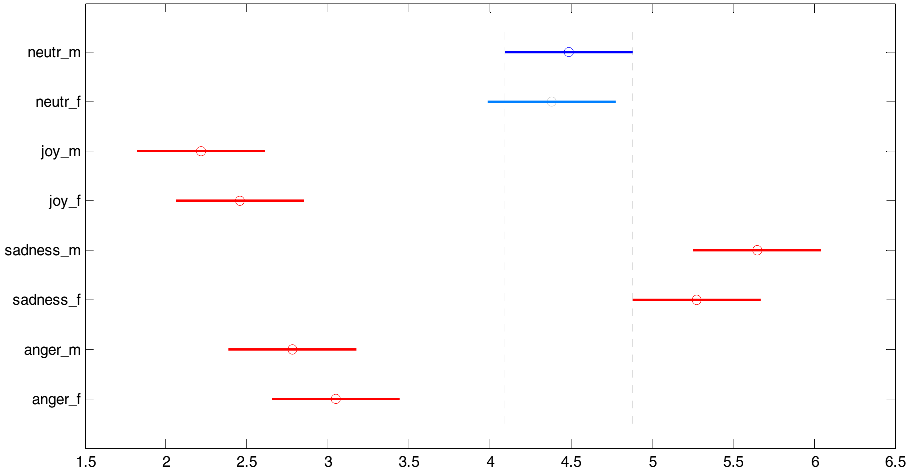

Fig. 9. Graphics results of zero crossing periods L_Z multiple comparison of ANOVA (male and female voice groups) for corresponding emotions

Zero crossing periods were subsequently used to calculate microintonation signal spectral analysis. Obtained results of spectral analysis including the 3-dB bandwidth values are shown in Tab. 8 for male voice, and in Tab. 9 for female voice. The average microintonation spectra (with and without smoothing by moving average) can be seen in Fig. 10 (male voice), and Fig. 11 (female voice).

Table 8. Summary results of spectral analysis (frequency parameters in [Hz] derived from concatenated differential F0 signal) – male voice.

Emotion	F_{Zmin} *)	F_{Zmean}	F_{Zrel}	F_{Zstd}	B_3	B_{3F} **)
Neutral	1.60	6.89	8.83	9.93	6.75	4.56
Joy	0.71	5.04	6.45	6.39	4.56	3.82
Sadness	0.73	6.11	7.78	7.23	4.39	2.69
Anger	1.81	6.18	8.00	9.12	5.37	4.07

*) $L_{Zmin} = 1 \Rightarrow F_{Zmax} = f_F / 2$.

**) 3-dB bandwidth corresponding to the signal smoothed by moving average filter with $M_F = 8$.

Table 9. Summary results of spectral analysis – female voice

Emotion	F_{Zmin} *)	F_{Zmean}	F_{Zrel}	F_{Zstd}	B_3	B_{3F} **)
Neutral	2.23	11.88	14.60	16.52	11.59	6.71
Joy	1.56	9.41	11.94	11.94	9.03	5.61
Sadness	1.56	9.33	11.66	11.52	7.20	3.17
Anger	2.08	9.88	12.59	14.11	10.74	5.86

*) $L_{Zmin} = 1 \Rightarrow F_{Zmax} = f_F / 2$.

**) 3-dB bandwidth corresponding to the signal smoothed by moving average filter with $M_F = 8$.

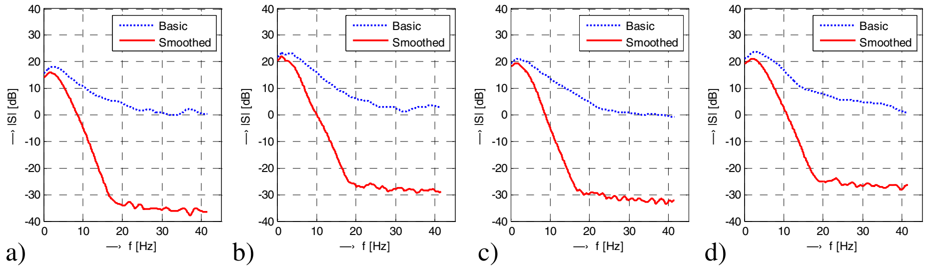

Fig. 10. Spectra of microintonation used for 3-dB bandwidth determination for emotions (with and without smoothing by moving average): “neutral” (a), “joy” (b), “sadness” (c), and “anger” (d) - male voice, f_F = 83.3 Hz

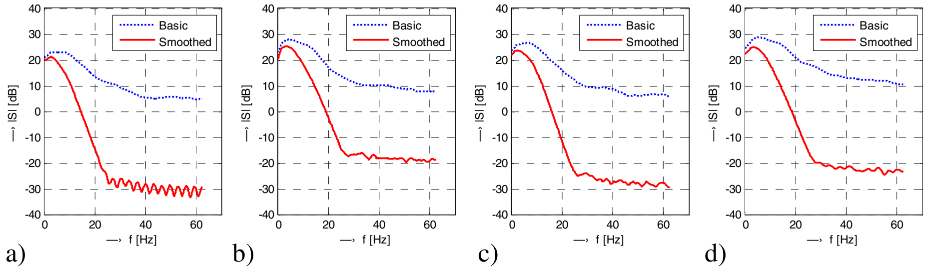

Fig. 11. Spectra of microintonation used for 3-dB bandwidth determination for emotions (with and without smoothing by moving average): “neutral” (a), “joy” (b), “sadness” (c), and “anger” (d) - female voice, f_F = 125 Hz

4 Conclusion

Statistical and spectral analysis of microintonation signal component of speech melody for several speakers and four emotional states (joy, sadness, anger, neutral state) was performed. From comparison of basic statistical microintonation analysis stored in Tab. 1 and Tab. 2 follows, that absolute jitter values are in accordance with the human vocal tract properties. Female shorter pitch periods are accompanied with shorter values of the absolute jitters, but higher relative changes in the frequency domain (mean $F0_{DIFF}$ values). The highest values of jitter correspond to “joy” and the lowest ones correspond to “sadness” for both voices. Similar results are shown in [25].

The same tendency can be observed for statistical results of zero crossing analysis. Although different frame lengths were used in microintonation frequency analysis for male and female voices, we can see matched similar values for all corresponding emotions. Visual comparison of histograms of zero crossing periods L_Z is not significant, but higher relative occurrence of low L_Z values can be noticed in “neutral” style for both voices. From objective statistical comparison of zero crossing periods L_Z by Ansari-Bradley hypothesis test follows, that the null hypotheses were rejected at the 5% significance level for each of the emotion types inside the gender group (see Tab. 5 and Tab. 6). On the other hand, the null hypotheses between the corresponding emotions of both types of voices are in all cases fulfilled at the same significance level (see Tab. 7). The result of final multiple comparison of ANOVA also confirms good correlation between particular emotions as shown Fig. 9.

Obtained results of spectral analysis (especially the B_3 and B_F values in Tab. 8 and Tab. 9) will be used to synthesize a digital filter for suppression of microintonation component of a speech signal. For microintonation removal from the F0 contour we can use FIR filters, e.g. an averaging filter or using Hann, Bartlett, or Hamming window. The first zero frequency of the transmission function of the averaging rectangular filter F_1 (approximate 1/2 of the main lobe width) will be given by $F_1 = f_F / M$, where M will the filter length. If we set this frequency F_1 equal to e.g. B_3 value, we obtain the rectangular filter length $M_R = f_F / F_1$. For the other three mentioned filters, the required filter length will be $M_X = 2 \, . \, M_R$. As regards visual comparison of average spectra in Fig. 10 and Fig. 11, the applied filtering by moving average was effective, the minimum signal suppression is about 25 dB for all emotions of both voices.

Out next aim will be to construct a special random noise generator for microintonation control and implement it to our TTS system with expressive speech production. This random generator will produce the samples (values for pitch-period length randomization) with different distribution in dependence of chosen type of emotion, as were analyzed in this paper.

Acknowledgments. The work has been done in the framework of the COST 2102 Action. It has also been supported by the Grant Agency of the Czech Republic (GA102/09/0989), by the Grant Agency of the Slovak Academy of Sciences (VEGA 2/0142/08) and the Ministry of Education of the Slovak Republic (VEGA 1/0693/08).

References

1. Iriondo, I., Planet, S., Socoró, J.-C., Martínez, E., Alías, F., Monzo, C.: Automatic Refinement of an Expressive Speech Corpus Assembling Subjective Perception and Automatic Classification. Speech Communication 51, 744–758 (2009)
2. Gobl, C., Ní Chasaide, A.: The Role of Voice Quality in Communicating Emotion, Mood and Attitude. Speech Communication 40, 189–212 (2003)
3. Keller, E.: The Analysis of Voice Quality in Speech Processing. In: Chollet, G., Esposito, A., Faúndez-Zanuy, M., Marinaro, M. (eds.) Nonlinear Speech Modeling and Applications. LNCS (LNAI), vol. 3445, pp. 54–73. Springer, Heidelberg (2005)
4. Campbell, N.: On the Use of Nonverbal Speech Sounds in Human Communication. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 117–128. Springer, Heidelberg (2007)
5. d'Alessandro, C., Darsinos, V., Yegnanarayana, B.: Effectiveness of a Periodic and Aperiodic Decomposition Method for Analysis of Voice Sources. IEEE Transactions on Speech and Audio Processing 6, 12–23 (1998)
6. Schoentgen, J.: Decomposition of Vocal Cycle Length Perturbations into Vocal Jitter and Vocal Microtremor, and Comparison of Their Size in Normophonic Speakers. Journal of Voice 17, 114–125 (2003)
7. Rank, E., Kubin, G.: An Oscillator-Plus-Noise Model for Speech Synthesis. Speech Communication 48, 775–801 (2006)
8. Hagmüller, M., Kubin, G.: Poincaré Pitch Marks. Speech Communication 48, 1650–1665 (2006)

9. Shahnaz, C., Zhu, W.-P., Ahmad, M.: A New Technique for the Estimation of Jitter and Shimmer of Voiced Speech Signal. In: Proceedings of the Canadian Conference on Electrical and Computer Engineering, CCECE 2006, Ottawa, Canada, pp. 2112–2115 (2006)
10. Farrús, M., Hernando, J., Ejarque, P.: Jitter and Shimmer Measurements for Speaker Recognition. In: Proceedings of the International Conference Interspeech 2007, Antwerp, Belgium, pp. 778–781 (2007)
11. Vasilakis, M., Stylianou, Y.: Spectral Jitter Modeling and Estimation. Biomedical Signal Processing and Control 4, 183–193 (2009)
12. Perrot, P., Aversano, G., Chollet, G.: Voice disguise and automatic detection: Review and perspectives. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) COST 277. LNCS, vol. 4391, pp. 101–117. Springer, Heidelberg (2007)
13. Murphy, P.: Source-Filter Comparison of Measurements of Fundamental Frequency Perturbation and Amplitude Perturbation for Synthesized Voice Signals. Journal of Voice 22, 125–137 (2008)
14. Juslin, P.N., Laukka, P.: Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code? Psychological Bulletin 129, 770–814 (2003)
15. Tao, J., Xin, L., Yin, P.: Realistic Visual Speech Synthesis Based on Hybrid Concatenation method. IEEE Transactions on Audio, Speech, and Language Processing 17, 469–477 (2009)
16. Duběda, T., Keller, E.: Microprosodic Aspects of Vowel Dynamics – An Acoustic Study of French, English and Czech. Journal of Phonetics 33, 447–464 (2005)
17. Low, P., Vaseghi, H.: Applying F0, Duration and Power Models with Microprosody Components in Text to Speech (TTS) Synthesis. In: Proceedings of the 47th International Symposium on Multimedia Systems and Applications, ELMAR 2005, Zadar, Croatia, pp. 229–232 (2005)
18. Navas, E., Hernáez, I., Luengo, I.: An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS. IEEE Transactions on Audio, Speech, and Language Processing 14, 1117–1127 (2006)
19. Přibil, J., Přibilová, A.: Application of Expressive Speech in TTS System with Cepstral Description. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 200–212. Springer, Heidelberg (2008)
20. Přibil, J., Přibilová, A.: Emotional Style Conversion in the TTS System with Cepstral Description. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 65–73. Springer, Heidelberg (2007)
21. Přibil, J., Přibilová, A.: Distributed Listening Test Program for Synthetic Speech Evaluation. In: Proceedings of the 34 Jahrestagung für Akustik DAGA 2008, Dresden, Germany, pp. 241–242 (2008)
22. Boersma, P., Weenink, D.: Praat: Doing Phonetics by Computer, Version 5.0.32 (Computer Program), `http://www.praat.org/` (Retrieved August 12, 2008)
23. Boersma, P., Weenink, D.: Praat - Tutorial, Intro 4. Pitch analysis, September 5, (2007), `http://www.fon.hum.uva.nl/praat/manual/Intro_4__Pitch_analysis.html`
24. Everitt, B.S.: The Cambridge Dictionary of Statistics, 3rd edn. Cambridge University Press, Cambridge (2006)
25. Tao, J., Kang, Y., Li, A.: Prosody Conversion from Neutral Speech to Emotional Speech. IEEE Transactions on Audio, Speech, and Language Processing 14, 1145–1154 (2006)