ČESKO-NĚMECKÝ FOND BUDOUCNOSTI
DEUTSCH-TSCHECHISCHER ZUKUNFTSFONDS

Institute of Phonetics

22ND CZECH-GERMAN WORKSHOP ON SPEECH COMMUNICATION

Prague, September 25-27, 2014

Book of Abstracts

http://fu.ff.cuni.cz/workshop2014/

Dear delegates,

It is our great pleasure to welcome you to the ***22nd Czech–German Workshop on Speech Communication*** at the Faculty of Arts in Prague!

We are happy that – judging from the number of registered participants at the Workshop – we can say the Workshop has been successfully revived and that we are able to continue the tradition dating back to the early 1990's of bridging boundaries not only between countries but also between disciplines. A cursory look at the Table of Contents reveals that the presentations cover a wide range of topics, as well as research methodologies: human speech is viewed in both the linguistic-phonetic and speech engineering framework, and analyzed using acoustic, articulatory, computational and neuroimaging methods. Several languages are analyzed, from the perspective of both native and second language, on the segmental and prosodic level.

The Workshop organization follows the traditional layout of the previous meetings, with two full days and a half dedicated to scientific presentations. These will also include two invited talks, by Prof. Klaus Kohler from Kiel University and by Prof. Petra Wagner from Bielefeld University.

Finally, we would like to gratefully acknowledge the support which the organization of the Workshop has received from the ***Czech–German Fund for the Future***.

We wish you a pleasant stay in Prague and to us all a successful scientific meeting!

Radek Skarnitzl & Oliver Niebuhr
and the local organizing team

Programme at a Glance

Thursday	Friday	Saturday
9:30 Opening		
9:45 Invited Talk 1 (*K. Kohler*)	9:45 Invited Talk 2 (P. Wagner)	10:00 Session 7
11:20 Session 1	11:20 Session 4	11:50 Session 8
14:30 Session 2	14:30 Session 5	15:00 Tour
16:20 Session 3	16:20 Session 6	
19:30 Musical Evening	18:30 Conference Dinner	

Programme

Invited talks *will take place in room no. 217, second floor*
Regular sessions *will be held in room no. 16, ground floor*

Thursday, September 25

9:30 Opening

9:45 ***Klaus Kohler: Elements of a network of communicative functions in speech interaction***

11:00 **coffee break**

SESSION 1 - chair: Petra Wagner

11:50 P. Šturm: Articulatory data on the syllable affiliation of Czech alveolar consonants

12:20 T. Valenta, L. Šmídl:
Commonly confused words in spontaneous Czech language transcription and recognition

LUNCH

SESSION 2 - chair: Radek Skarnitzl

14:30 P. Howson, P. J. Monahan:
An Acoustic Analysis of Czech Alveolar and Post-Alveolar Fricatives

15:00 A. Braun, J. Sommer, A. Jansen:
Speaker recognition experiments with fMRI: A feasibility study

15:30 I. Kraljevski, R. Jäckel, R. Kompe, G. Strecha, F. Kurnot, M. Rudolph, D. Hirschfeld, R. Hoffmann:
Speech Quality Assessment in a Pronunciation Trainer for Speech Disorder Therapy

16:00 **coffee break**

SESSION 3 - chair: Jan Volín

16:20 J. Veroňková, Y. Tolkunova:
Vowel-related glottalization in read speech of native and non-native speakers of Czech

16:50 P. Howson, E. Komova:
The Story of *ř*: An Ultrasound Examination

17:20 V. J. Podlipský, Š. Šimáčková:
Perception of prosodic accentedness by native and non-native listeners

19:30 Musical Evening

Friday, September 26

9:45 ***Petra Wagner: Research on prominence in phonetics and speech technology – about an intuitive concept, its usefulness and its methodological stepping stones***

11:00 **coffee break**

SESSION 4 - chair: Oliver Niebuhr

11:20 E. Churaňová: The relationship between production and perception of speech rhythm in controlled Czech words

11:50 J. Volín, H. Bartůňková:
Fundamental Frequency Parameters in Native and Foreign-Accented Speech

12:20 L. Weingartová, J. Volín: Amplitude Differences in Polysyllabic Words of Czech English

LUNCH

SESSION 5 - chair: Jonáš Podlipský

14:30 R. Landgraf, O. Niebuhr, G. Schmidt, T. John, C. Lüke, A. Theiß:
Simulating the Lombard-Effect in In-Car-Communication

15:00 P. Mizera, P. Pollák: Estimation of Articulatory Features for Czech language

15:30 O. Niebuhr: Stepping out of the line in intonation – New insights into the forms and functions of F0 stylization

16:00 **coffee break**

SESSION 6 - chair: Jindřich Matoušek

16:20 M. Jůzová: The Utilization of Contexts in Limited Domain Speech Synthesis

16:50 M. Jůzová, J. Vít: Artifact reduction in concatenation speech synthesis

17:20 R. Vích, J. Staněk: DCT based Voice Conversion with Multipoint Frequency Transformation

18:30 Conference Dinner

Saturday, September 27

SESSION 7 - chair: Petr Pollák

10:00 R. Skarnitzl, J. Vaňková, T. Bořil: Optimizing formant extraction in Praat and Snack: Comparison of manual and automatic measurements

10:30 M. Borský, P. Pollák: Recognition of Spectrally Distorted Speech after MP3 Compression

11:00 J. Heranová, T. Bořil, R. Skarnitzl:
Dynamic Harmonics-to-Noise Ratio in Segmentation Tasks

11:30 **coffee break**

SESSION 8 - chair: Zdena Palková

11:50 R. Hoffmann, L.-P. Löbe: Ernst Mach and Johannes Kessel in Prague 1871-1874

12:20 P. Šturm: Looking back at the 6th ICPhS in Prague, 1967: What does it tell us about phonetics?

15:00 Tour

TABLE OF CONTENTS

INVITED TALKS:

REGULAR TALKS:

Elements of a network of communicative functions in speech interaction

Klaus J. Kohler

Christian-Abrechts-Universität zu Kiel
kjk@ipds.uni-kiel.de

The study of prosody in general, and of tonal aspects in particular, has occupied the human mind for centuries to elucidate their contributions to meaning over and above that conveyed by lexical fields and syntactic structures. Systematic analysis of intonation has been carried out with increasing breadth and depth over the past century. In the resulting prosodic systems, meaning is subordinated to form. The question is not "What are the communicative functions in a network of human interactions, and how are they manifested by formal means - lexical, syntactic, prosodic - in speech acts in the languages of the world?" Rather, meaning is grafted onto formal linguistic structures, which Ladd (1996, 2008) termed 'The Linguist's Theory of Intonational Meaning'. The goal is to filter out all affect and attitude to show up the meaning of propositional linguistic structures. Such descriptive accounts of the prosodic phonology of a language are very useful, particularly when nothing or very little is known about the contrastive patterns in a language. But they do not give us a great deal of insight into how speech communication works in all its facets of meaning transmission, which speech scientists should now be interested in elucidating in those languages that have been thoroughly investigated from the formal angle.

This talk aims to redefine Ladd's intonational meaning as 'The Speech Scientist's Theory of Intonational Meaning', which puts a network of communicative functions first and then associates formal exponents with them, paying attention to lexical and syntactic form beside the main focus on prosody. The function-form paradigm is built on the axiomatic postulate that the network of communicative functions is inherent in human speech interaction, irrespective of any particular language. On the other hand, the association of form at various levels of description from lexicon and syntax to prosody varies between languages. So, such a functional network approach can provide a powerful tool in comparative prosodic studies. The power of this approach will be demonstrated with some elements from the global network of communicative functions, such as the functions of QUESTION and INTENSIFICATION.

References:

Ladd, D. R. (2008). *Intonational Phonology.* Cambridge: Cambridge University Press.

Research on Prominence in Phonetics and Speech Technology – Obstacles and Benefits Caused by its Terminological Vagueness

Petra Wagner

Universität Bielefeld
petra.wagner@uni-bielefeld.de

The term (prosodic/perceptual/phonetic) prominence is used to denote a series of related but nevertheless distinct phenomena: Sometimes it refers to phonological concepts such as the presence of lexical or sentence stress, sometimes to abstract phonetic events such as the presence of a pitch accent, sometimes to events on the phonetic signal level, e.g. the degree of pitch excursion, subglottal pressure or accentual lengthening. Central to this problem of vagueness in usage appears to be the lacking consensus on (i) a definition of prominence, (ii) its categorical or gradual nature, (iii) its physical manifestation, (iv) its language specificity or universality, and (v) the linguistic unit(s) it manifests itself on.

Despite the difficulties to pinpoint its exact meaning, the topic of prominence has received an increasing amount of attention in our community, probably due to its relevance in much applied research such as speech recognition, computer assisted pronunciation teaching, clinical phonetics, but also in theoretical linguistics and general phonetics. The obvious popularity of the term shows its general usefulness to treat and explore a plethora of related issues, which are under investigation across various research communities.

In my talk, I will give an overview about our current understanding of the term prominence, its diverging usages, its various cues and correlates, its measurement and its potential for applied research. I will also try to provide a set of methodological guidelines for a better understanding of its nature across disciplines.

Recognition of Spectrally Distorted Speech after MP3 Compression

Michal Borský & Petr Pollák

Czech Technical University in Prague, Department of Circuit Theory
borskmic@fel.cvut.cz

Research questions: The deployment of automatic speech recognition (ASR) systems into real-life are often met with difficulties of diverse acoustic conditions. This diversity is what forces the necessity to build the systems as robust to ensure their reliable performance regardless of the conditions. The usage of MP3 compression represents one of such conditions. when the property of lossy encoding degrades the quality of extracted features and therefore the recognition. The research of optimized settings for MP3 recognition has been conducted by various authors [Bar2001], [Bes2001], [Ng2004], [Pol2012], [Nou2013] and different solutions have been proposed. This work presents the analysis of optimized setup which was focused on blocks of feature extraction and acoustic modeling. The work summarizes the effects of methods proposed by the previously mentioned authors and tested to determine the potential contribution of each method separately as well as in unison.

Method: The main goal of the optimization was to find the proper segmentation. determine the importance of feature normalization and dithering and the application of acoustic model adaptation. The experiments were performed on signals of very good quality which were artificially compressed to simulate the effect of the spectral distortion. The PLP features were extracted and normalized using CMVN and various levels of noise were added. The main purpose was to reduce the effect of spectral distortion brought by compression. The context dependent AMs were trained for RAW data and 160kbit. 32kbit. 24kbit. 16kbit compression speeds. The final AMs were adapted by CMLLR and MAP techniques. The goal of adaptation was to further improve the AM quality and to test the model interchangeability. The recognition was done on LVCSR task of 1 hour with trigram LM.

Results: The first experiments were focused on finding the optimal setup for feature extraction. The window length of 25 [ms] and shift of 12.5 [ms] achieved the best results of all the commonly used ones. This setup performed consistently. the *WERs* for 160kbit. 32kbit. 24kbit. 16kbit AMs were 29%. 32%. 32\% and 36% respectively. The second set of experiments were focused on acoustic modeling. adaptations and model interchangeability. The usage of RAW AM on compressed data worked rather well and WERs were 30%. 32%. 34% and 41%. The application of dithering lowered this values to 23%. 26%. 29% and 35%. The application of AM adaptation lowered the *WER* to about 20% for CMLLR and 10% for MAP regardless of the bit-rate speed. The difference between bit-rate specific and RAW model was found to be negligible if the adaptation was used.

Conclusions: The recognition of MP3 compressed speech was analyzed. The results achieved showed that the application of mentioned methods can lead to significant improvement. The future research plans include another methods for acoustic modeling. such as speaker adaptive training. discriminative methods and alternative methods for unsupervised adaptation. not only for MP3 speech but also for other cases when the AM quality is degraded due to spectral distortion.

References:

C.-M. Liu. H.-W. Hsu. and W.-C. Lee (2008). Compression artifacts in perceptual audio coding. IEEE Transactions on Audio Speech. and Language Processing. pp. 681–695.

R. H. van Son (2005). A study of pitch, formant and spectral estimation errors introduced by three lossy speech compression algorithms. Acta Acustica united with Acustica. vol. 91. pp. 771–778.

C. Barras. L. Lamel. and J. Gauvain (2001). Automatic transcription of compressed broadcast audio. Proc.of IEEE International Conference on Acoustics. Speech. and Signal Processing, vol. 1. pp. 265–268.

L. Besacier. C. Bergamini. D. Vaufreydaz. and E. Castelli (2001). The effect of speech and audio compression on speech recognition performance. Proc. Of IEEE Multimedia Signal Processing Workshop.

P. S. Ng and I. Sanches (2004). The influence of audio compression on speech recognition systems. Proc. of Conference Speech and Computer.

P. Pollak and M. Borsky (2012). Small and large vocabulary speech recognition of MP3 data under realword conditions: Experimental study. E-Business and Telecommunications of Communications in Computer and Information Science. pp. 409–419.

J. Nouza. P. Cerva. and J. Silovsky (2013). Adding controlled amount of noise to improve recognition of compressed and spectrally distorted speech. Proc. of IEEE International Conference on Acoustics. Speech and Signal Processing. pp. 8046–8050.

J. Li. L. Deng. D. Yu. Y. Gong. and A. Acero (2009). A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions. Computer Speech & Language. pp. 389 – 405.

Lame encoder. Available at : http://lame.sourceforge.net/.

P. Fousek. P. Mizera. and P. Pollak. Ctucopy feature extraction tool. Available at. http://noel.feld.cvut.cz/speechlab/.

S. Young and et al.. The HTK Book. Version 3.4.1. Cambridge. 2009.

Speaker recognition experiments with fMRI: A feasibility study

Almut Braun[1], Jens Sommer[2] & Andreas Jansen[2]

[1]Department of Phonetics, University of Marburg, Germany
[2]Section of BrainImaging, Department of Psychiatry and Psychotherapy, University of Marburg, Germany
almut.braun@staff.uni-marburg.de
{jens.sommer|andreas.jansen}@med.uni-marburg.de

Research questions: The present study aims to investigate human speaker recognition ability while listeners are undergoing a functional MRI scan. Central questions are: To what extent - if at all - is it possible to do a more complex speaker recognition experiment within an MR scanner? If so, could different patterns of BOLD (blood-oxygen-level-dependent) activation and deactivation be linked to a listeners‘ performance in a speaker recognition task? Do familiar voices evoke BOLD activations in other areas of the brain than voices which have just been heard once before?

In previous studies, voice coding has been associated with the superior temporal sulci (STS) and the inferior frontal cortex (Andics et al. 2013), and voice recognition was associated with the middle and posterior STS, the right ventrolateral prefrontal regions and the insular cortex, the anterior temporal pole (Andics et al. 2010).

This is a feasibility study. It will be tested whether it is generally possible to do a speaker recognition experiment within the noisy environment of a 3-tesla MR scanner. Different settings of the scanner as well as different types of headphones (electroacoustic/pneumatic) have been tested to reduce the subjective noise level. It was reported that the best noise reduction could be obtained when the listener was wearing electroacoustic headphones (mr-confon). Further improvements could be achieved by wrapping the participant’s head with special foam material inside the head coil and by adjusting the scanning parameters (e.g. echo time, repetition time, field of view, matrix) to separate noise and voice frequencies.

If the feasibility study reveals no weaknesses in the local setup, a follow-up study with blind and sighted listeners will be carried out. Gougoux et al. 2009 found different activation patterns for blind and sighted listeners in a voice discrimination task (same/different speaker).

Method: The experiment consists of two parts. In the first part, a sound file with 15 spontaneous voice samples of different male speakers is played to the listeners while lying inside the MR scanner. Ten of these voice samples come from famous speakers which are supposed to be recognized easily, five samples are voices which the listeners have never heard before. Each voice sample (duration: 30 seconds) is followed by a silent interval of 10 seconds to ensure that a baseline can be established. The degree of familiarity of the selected speakers was priorly tested on later uninvolved participants. Famous striking

voices are included in the experiment to provoke extreme reactions to estimate a general effect size for the given task.

In a second part, the participants undergo a second fMRI scan and have to listen to another sound file. This sound file consists of voice samples of 3 non-famous speakers they had heard in the first part of the experiment and 6 new (unknown) voices. Participants are asked which of the voices they have heard before.

Results and Conclusions: Preliminary findings indicate that listeners are able to recognize familiar speakers even in the noisy environment of an MR scanner. The analysis of the fMRI data is still pending.

.

References:

Andics, A., McQueen, J.M., Petersson, K.M., Gal, V., Rudas, G. & Vidnyanszky, Z. (2010). Neural mechanisms for voice recognition. *Neuroimage, 52,* pp. 1528-1540.

Andics, A., McQueen, J.M. & Petersson, K.M. (2013). Mean-based neural coding of voices. *Neuroimage, 79,* pp. 351-360.

Gougoux, F., Belin, P., Voss, P., Lepore, F., Lassonde, M., & Zatorre, R.J. (2009). Voice perception in blind persons: a functional magnetic resonance imaging study. *Neuropsychologia,* 47 (13), pp. 2967-2974.

The relationship between production and perception of speech rhythm in controlled Czech words

Eliška Churaňová

Institute of Phonetics, Faculty of Arts, Charles University in Prague
eliska.churanova@ff.cuni.cz

Research questions: In previous research the concept of the *perceptual centres* [1, 2] which were supposed to be associated with the onset of a vowel in the stressed syllable [3] has been well established. The research paradigms were often based on adjusting stimuli to perceived isochrony [4, 5]; some investigations with speech-metronome synchronisation were carried out as well [6]. The present paper is one of the follow-up studies on this experimental design. The general aim was to discover whether the sensation of rhythmicity in Czech depends on the distance of the metronome beat (a representation of a p-centre, [6]) from the onset of a vowel in the stressed syllable rather than on the speaker's consistency in the articulation of given words along with periodic pulses.

Method: For the present perception experiment a part of the speech material was used from [7], in which native Czech speakers were asked to synchronize the first syllable of 2-syllable words with metronome pulses. They produced each word eight times; the first and the last repetitions are omitted from analyses. The material was processed in the Praat programme [8]. The average distance (within the six realizations) of a metronome beat from the beginning of the first vowel was calculated, as well as a standard deviation. Based on these two specifications, the items for the perception test were selected. The structure of the stressed syllables was taken into account as well. The perception test consisted of 42 items (14 speakers), each one contained six repetitions of certain word with audible beats. The 30 listeners then estimated on a three-point scale how rhythmical the items sounded.

Results: The average evaluation of each item was calculated. The groups of items, separated by the values of the above mentioned parameters, were compared with each other. It turned out that the average distance of pulses from the beginning of a vowel in the stressed syllable is not significant for the subjective evaluation of rhythmicity (Kruskal-Wallis ANOVA: $p > 0.1$), but the SD of the distance of a beat from the vowel proved to be significant for the sensation of rhythm – the items with smaller SD were considered more rhythmical (Kruskal-Wallis ANOVA: $H(2, n = 40) = 12.8; p < 0.01$). The items with the beat consistently located into the onset of the syllable had better evaluation than those with the beat fluctuating before, in or after the onset of a syllable (Kruskal-Wallis ANOVA: $H(1, n = 42) = 4.6; p < 0.05$). For the perception of rhythm the SD of the duration of pauses between the six repetitions of the words plays an important role as well (Kruskal-Wallis ANOVA: $H(2, n = 42) = 8.9; p < 0.05$).

Conclusions: The results of this study show that the speaker's consistency in the articulation of Czech words along with periodic pulses was more relevant than the exact location of the beats within the words. However, further research on natural speech is needed. The effect of words should be investigated as well.

References:

[1] Morton, J.; Marcus, S. & Frankish, C. (1976). Perceptual centres (P-centres). *Psychological Review*, *83*, pp. 405–408.

[2] Fowler, C. A. (1979). "Perceptual centers" in speech production and perception. *Perception & Psychophysics*, *25* (5), pp. 375–388.

[3] Allen, G. (1972). The location of rhythmic stress beats in English: An experimental study I. *Language and Speech*, *15*, pp. 72–100.

[4] Marcus, S. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception & Psychophysics*, *30*, pp. 247–256.

[5] Cooper, A. M., Whalen, D. H., & Fowler, C. A. (1986). P-centers are unaffected by phonetic categorization. *Perception & Psychophysics*, *39*, pp. 187–196.

[6] Barbosa, P. A.; Arantes, P.; Meireles, A. R. & Vieira, J. M. (2005). Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005)*, Lisbon, Portugal, pp. 1441–1444.

[7] Volín, J.; Churaňová, E. & Šturm, P. (2014). P-centre position in natural two-syllable Czech words. In *Proceedings of the 7th International Conference of Speech Prosody*, Dublin, Ireland, pp. 920–924.

[8] Boersma, P. & Weenink, D. (2013). *Praat: doing phonetics by computer* (Version 5.3.41). Retrieved from http://www.praat.org.

Dynamic Harmonics-to-Noise Ratio in Segmentation Tasks

Jana Heranová, Tomáš Bořil & Radek Skarnitzl

Institute of Phonetics, Faculty of Arts, Charles University in Prague
jana.heran@gmail.com, {tomas.boril|radek.skarnitzl}@ff.cuni.cz

Research questions: Segmented acoustic signal is a pre-requisite of any kind of acoustic analysis. HMM-based automatic segmentation (e.g., Pollák et al., 2007; van Niekerk & Barnard, 2007) is known to be inadequate for the purpose of phonetic analyses. The aim of the present study is to follow up on previous work (Heranová & Skarnitzl, 2011) in designing post-hoc adaptations of automatic forced alignment output and bringing it more in line with phonetically based segmentation criteria (Machač & Skarnitzl, 2009). Segment boundaries between voiceless obstruents (O) and neighbouring vowels (V) tend to be relatively straightforward for a manual labeller; however, automatically determined segment boundaries in these contexts turned out to be quite imprecise in comparison with manual segmentation. Heranová & Skarnitzl (2011) demonstrated that the dynamic course of harmonicity (HNR) in O–V and V–O contexts, as determined by a cross-correlation-based method in Praat (Boersma & Weenink, 2014; Boersma, 1993), could be used to obtain a more precise location of automatic forced alignment. This study compares our earlier results with the course of HNR values around the segment boundary obtained using a cepstrum-based method (de Krom, 1993) implemented in VoiceSauce (Shue, 2014).

Method: Automatic HMM-based and manual segmentation were performed on recordings of 12 Czech adult speakers. HNR values were retrieved with a time step of 5 ms for O–V and V–O contexts from –25 ms to +25 ms set around the manually determined boundary. HNR values were obtained using default settings for cross-correlation in Praat and using a cepstrum-based method in VoiceSauce with identical settings. Afterwards, a subset of contexts was selected based on the condition that the difference in placement of the automatic and manual boundaries ranged from 10 to 20 ms, yielding 459 occurrences (168 O–V and 291 V–O contexts) for analysis.

Results: Two types of HNR curves were identified based on the Praat output: standard curves containing a distinctive change between negative and positive HNR values around the segment boundary and non-standard curves where no such change took place. A new placement of segment boundaries can be proposed for the standard HNR courses (Figure 1); however, it is clear that non-standard curves would not be suitable for streamlining segmentation. The performance of Praat was superior to VoiceSauce in the standard cases, which yielded 103 non-standard combinations (22.4%). VoiceSauce seems to fill this gap as it provided segmentation clues in 89 of these 103 cases.

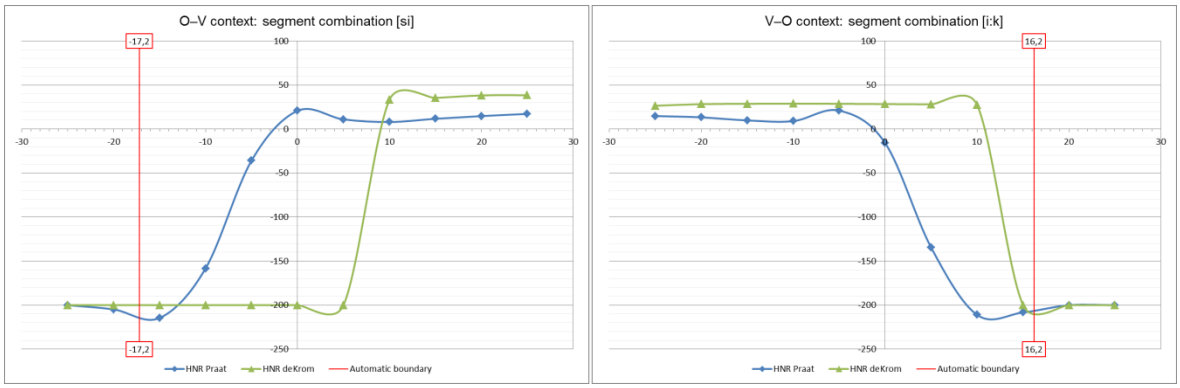

Figure 1: Example of boundary placement for standard HNR curves in O–V and V–O context. The blue curve represents the cross-correlation based method; the green curve stands for the cepstrum-based estimate of HNR around the manually set segment boundary (i.e. around the point zero). The red line marks the position of the automatic boundary.

Conclusions: In order to enhance automatic HMM-based segmentation, a post-hoc adjustments of boundary placement based on course of HNR is proposed: for standard HNR curves the Praat method suggests the local HNR maximum for O–V contexts and the +5 ms point after the local HNR maximum for V–O contexts. The cepstrum-based method should be used for non-standard HNR curves, with optimum boundary placement being 10 ms before the first positive HNR value in O–V contexts and 10 ms before the last positive HNR value for V–O contexts.

References:

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings*, 17, pp. 97-110.

Boersma, P. & Weenink, D. (2014). *Praat: doing phonetics by computer* (Version 5.3.77). Retrieved from http://www.praat.org.

de Krom, G. (1993). A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals. *Journal of Speech and Hearing Research*, 36, pp. 254-266.

Heranová, J. & Skarnitzl, R. (2011). Využití harmonicity při fonetické segmentaci řeči. *Akustické listy*, 17 (4), pp. 3-9.

Machač, P. & Skarnitzl, R. (2009). *Principles of Phonetic Segmentation*. Praha: Nakladatelství Epocha.

Pollák, P., Volín, J., Skarnitzl, R. (2007): HMM-Based Phonetic Segmentation in Praat Environment, Proc. of SPECOM 2007, p. 537-541, Moscow.

Shue, Y. (2014). *VoiceSauce: A program for voice analysis* [computer program], version 1.14, retrieved from: http://www.seas.ucla.edu/spapl/voicesauce/.

van Niekerk, D. R. & Barnard, E. (2007). Important factors in HMM-based phonetic segmentation. *Proc. of the 18th PRASA*, pp. 25-28. Pietermaritzburg, South Africa.

Ernst Mach and Johannes Kessel in Prague 1871-1874

Rüdiger Hoffmann[1] & Lutz-Peter Löbe[2]

[1]Institut für Akustik und Sprachkommunikation, Technische Universität Dresden
[2]Klinik Borkum-Riff, Borkum
ruediger.hoffmann@tu-dresden.de

Research questions: This paper summarizes a specific part of a project aimed to result in the scientific biography of the otologist Johannes Kessel (1839-1907), which does not yet exist. Kessel studied in Gießen and Würzburg and had a long postdoctoral phase in Vienna and Prague. Coming from Vienna, where he studied the histology of the ear at the institute of the renowned pathologist Salomon Stricker, Kessel turned to Prague for a working stay at the chair of the famous physicist Ernst Mach (1838-1916) in the years 1871-1874. This cooperation was important because a number of essential findings in the psychophysics of hearing were published by both authors. Following his habilitation, Kessel worked as an outside lecturer at the University of Graz, where he performed the first stapes mobilization (1875), followed by further new procedures in the surgery of the middle ear which may be characterized as early steps towards tympanoplasty. From 1886, he worked as a professor for otology at Jena University. Although his work was important for hearing acoustics, otology, and rehabilitation engineering as well, there is no more biographical material than a biographical sketch (Stelzig 1970) and a chapter in a PhD thesis on the history of the Jena ORL clinics (Pfeiffer 2005). This gap will be closed by our forthcoming book, and as far as Kessel's Prague period is concerned, by this paper.

Method: Whereas the Prague period of Mach is well documented (D. Hoffmann 1991), we knew nearly nothing about Kessel in Prague and the details of his cooperation with Mach. Therefore we analyzed (a) hitherto unpublished letters from both Mach and Kessel, (b) archive material, and (c) the diaries of Mach which are now deposited in the Deutsches Museum Munich.

Results: The cooperation of Mach and Kessel may be subdivided into three steps: In a common working period 1871/72, they performed investigations of anatomic dissections by means of Lissajous figures of the vibrations, first investigations of the behaviour of the living ear, and stroboscopic measurement of pitch. In 1873, they worked separately (because Mach had to serve as Dean). Kessel continued to investigate dissections of the eardrum and applied stroboscopic methods. Mach was mainly interested in experiments on the sensation of equilibrium and movement. In a final period (1874), the authors published a joint summarizing paper (Mach & Kessel 1874) in the proceedings of the Vienna academy. There they propose a coordinate system for the geometric description of the middle ear, publish examples from real measurements, and describe new stroboscopic analyses of the mechanics of the middle ear. The work was partially sponsored by a grant from the Vienna academy.

Conclusions: It is interesting to demonstrate that Prague is not only an important place for the history of phonetics and speech sciences but also for their psychophysical prerequisites.

Mach and Kessel performed an important common research project. From the medical point of view, this was pointed out already by Hybášek (2005a, b). We want to do this here for the community of speech researchers as well.

References:

Stelzig, G. (1970). Johannes Kessel - Vater der Stapes- und funktionellen Mittelohr-chirurgie. *Ztschr. f. Laryngologie, Rhinologie, Otologie und ihre Grenzgebiete, 49*, pp. 551-564.

Pfeiffer, W. (2005). *Entwicklung von Klinik und Lehrstuhl für Hals-Nasen-Ohrenheilkunde an der Universität Jena von 1884 bis 1957*. Med. Diss., Jena.

Hoffmann, D. (1991). *Ernst Mach in Prag*. In: D. Hoffmann & H. Laitko (Eds.), Ernst Mach. Studien und Dokumente zu Leben und Werk, pp. 141-178. Berlin: Deutscher Verlag der Wissenschaften.

Mach, E. & Kessel, J. (1874). Beiträge zur Topographie und Mechanik des Mittelohres. *Sitzungsberichte der mathematisch-naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften, LXIX, III. Abtheilung*, pp. 221-243.

Hybášek, I. (2005a). Z historie středoušní funkční chirurgie. *Otorinolaryngologie a foniatrie, 54*, pp. 109-110.

Hybášek, I. (2005b). Historie operačního mikroskopu a stroboskopie v otologii. *Otorinolaryngologie a foniatrie, 54*, pp. 169-170.

The story of *ř*: an ultrasound examination

Phil Howson[1] & Ekaterina Komova[2]

[1]Department of Linguistics, University of Toronto
[2]Department of East Asian Languages and Cultures, Columbia University
phil.howson@mail.utoronto.ca
ek2853@columbia.edu

Research questions: The phonetic inventory of Czech is frequently described as unique in having a typologically rare trill-fricative, represented orthographically as *ř*. The trill-fricative is typically considered to be a reflex of the Proto-Slavic *r^{j} (Carlton, 1991). Previous researchers (c.f. Iskarous & Kavitskaya, 2010) have postulated that the sound change from *r^{j} to /r̝/ resulted from the conflicting articulatory constraints on palatalization and trilling: palatalization requires fronting of the tongue dorsum (Ladefoged & Maddieson, 1996), while trilling requires backing of the tongue dorsum (Proctor, 2009). The purpose of this research is to compare the trill-fricative with the palatalized trill to determine what the conflict between palatalization and trilling is. Based on our results, we conclude that it is not the fronting of the tongue dorsum, but rather the raising of the tongue body that causes the instability of palatalized trills and led to the sound change from *r^{j} to /r̝/.

Method: The data of a single bilingual speaker of Russian and Czech was recorded in two sessions (one for each language). First, the participant produced the Czech phonemes /r̝, r/ in the environments #rV, VrV, Vr#, using the real words *řád* 'order', *pařát* 'talon', *tvář* 'face', *rád* 'glad', *paráda* 'finery' and *tvar* 'shape' as stimuli. The second session recorded the participant's production of the Russian phonemes /r^{j}, r/ in the same environments, using *r^{j}ad* 'row', *par^{j}at* '(they) soar', *tvar^{j}* 'beast', *rad* 'glad', *parad* 'parade' and *otvar* 'decoction'. Ten repetitions were recorded and the first two were discarded. Images at the point of maximum constriction - the maximum displacement of the articulator - were captured and Edgetrak was used to trace the tongue surface shapes. R Statistical Software (R Core Team, 2014) was used to generate and compare SSANOVAs of the tongue surface shapes for the Czech /r, r̝/ and the Russian /r̠, r^{j}/.

Results: The results indicate that compared to /r̝/, the tongue dorsum for /r^{j}/ was fronted. The tongue body for /r^{j}/ was also raised, implicating a more laminal articulation. These results suggest that the tongue body raising is responsible for the fronting of the tongue dorsum. These findings support Ladefoged & Maddieson's (1996) hypothesis that the raised tongue body impedes the airflow, making the precise aerodynamic conditions necessary to facilitate trilling harder to achieve. Conversely, the flat tongue contour for the unpalatalized trills acts like a funnel, allowing for the airflow to be channel towards the tongue tip. This produces a more favorable environment for trilling. Furthermore, having a more laminal articulation creates a larger vibrating mass during trilling, which reduces the lateral tongue bracing and requires more careful control to permit trilling.

Conclusions: The results suggest that the tongue dorsum fronting is a consequence of the raising of the tongue body. The raising of the tongue body makes the aerodynamic

conditions necessary for trilling harder to achieve, while laminal articulation reduces the stability of the tongue, which is paramount for the articulation of trills. These factors lead to the sound change from *r^j to /r̠/.

References:

Carlton, T.R. (1991). *Introduction to the phonological history of the Slavic languages.* Columbus: Slavica Publishers.

Iskarous, K. & Kavitskaya, D. (2010). The interaction between contrast, prosody, and coarticulation in structuring phonetic variability. *Journal of Phonetics* 38, 625-639.

Ladefoged, P. & Maddieson, I. (1996). *The Sounds of the World's Languages.* Oxford: Blackwell.

Proctor, M. 2009. *Gestural characterization of a phonological class: The liquids. Ph.D. dissertation,* Yale University.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

An Acoustic Analysis of Czech Alveolar and Post-Alveolar Fricatives

Phil Howson[1] & Philip J. Monahan[1,2]

[1]Department of Linguistics, University of Toronto
[2]Centre for French and Linguistics, University of Toronto Scarborough
phil.howson@mail.utoronto.ca
philip.monahan@utoronto.ca

Research Question: The aim of this study is to examine articulatory differences between voiced (i.e., /z, ɻ, ʒ/) and voiceless (i.e., /s, ʃ/) fricatives through acoustic measures. Liker & Gibbon (2013) suggested that the onset and offset of frication are different between voiced and voiceless fricatives as a result of the careful control necessary to facilitate or inhibit voicing and that this control is physiological. Our results suggest that there are also distinct supraglottal differences in the articulation of voiced and voiceless pairs.

Method: Four native speakers (2 female) of Czech recorded the following words: *záď* 'stern (ship)', *sát* 'suck', *řád* 'order', *žák* 'pupil (school)', *šál* 'scarf.' Six repitions of each word plus distractor tokens were produced in randomized order. We measured the spectral centre of gravity (COG) of the fricative and the first (F1) and second formants (F2) of the vowel with Praat (Boersma & Weenink, 2014). COG measurements were taken from a 20 ms interval at the mid-point of each fricative, and F1 and F2 measurements were taken at 5% of the following vowel. F0 was excluded from the COG analysis. For the analysis, a repeated measures ANOVA with factors Phoneme and Voice and post-hoc analyses were performed using R (R Core Team, 2013).

Results: We find a main effect of Voice for F1, with voiced fricatives showing a higher F1 than voiceless ones. This supports the findings that there is a physiological difference in the substantiation of voicing contrasts (Liker & Gibbon, 2013). To maintain voicelessness, the articulatory posture must be held longer, while voiced phonemes can quickly transition to the next gestural target. For F2, we observe a reliable difference for the factor Phoneme but not Voice. The voiced fricative /ʒ/ had a higher mean F2 than its voiceless counterpart /ʃ/. This follows Narayanan et al. (1995) in that voiced fricatives utilize a strategy of vocal cavity enlargement accomplished by tongue fronting. The tongue backing observed in /ʃ/ also supports the conclusion that a lateral lock between the tongue dorsum and the palate is required to help facilitate and control voicelessness (Liker & Gibbon, 2013). For the COG of the fricative, we find main effects for both Phoneme but not Voice. The alveolars had a higher mean COG than /ɻ/ and the post-alveolars. /ɻ/ also had a higher mean COG than the post-alveolars. Finally, we analyzed the power spectra density estimates and the spectral slopes (Żygis, Pape, & Jesus, 2012). The alveolars differed from the post-alveolars and /ɻ/ with respect to the front cavity resonance, and the fricatives differed in spectral slopes, which is generally thought to distinguish different fricatives.

Conclusions: The results of the current study suggest that there is a physiological difference between the mechanisms used to produce voicing contrasts, consistent with previous research. Careful control of the tongue dorsum is required to facilitate voicelessness, while

a larger supraglottal cavity helps facilitate voicing. This study also highlights the usefulness of acoustic measures in determining fine-grained differences in the articulation of consonant contrasts.

References:

Boersma, P. & Weenink, D. (2014). *Praat: doing phonetics by computer* [Computer Program]. Version 5.3.77. Retrieved from http://www.praat.org.

Liker, M. & Gibbon, F. (2013). Differences in EPG contact dynamics between voiced and voiceless lingual fricatives. *Journal of the International Phonetic Association* 43, 49-64.

Narayanan, S., Alwan, A., & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *Journal of the Acoustical Society of America* 98(3), 1325-1347.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Żygis, M., Pape, D., & Jesus, L. M. T. (2012). (Non-)retroflex slavic affricates and their motivation: evidence from Czech and Polish. *Journal of the International Phonetic Association* 42(3): 291-329.

The Utilization of Contexts in Limited Domain Speech Synthesis

Markéta Jůzová

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen
juzova@kky.zcu.cz

Research questions: While synthesizing a given sentence, parts of a real recorded speech (speech units) are concatenated together. When we prepare a text corpus for a general text-to-speech (TTS) system, we try to select texts containing as many speech units (e.g. diphones) as possible, regarding their prosodic and phonetic context. However, for the purpose of the limited-domain text-to-speech synthesis (LDTS), the corpus preparation is different, because a limited-domain (LD) corpus should ensure 100% coverage of the given domain. During the synthesis itself, longer speech units (like words or phrases) are concatenated. The concatenations are usually done in pauses, which the two concatenated units were expanded by, but it is unnatural and synthesized sentences do not sound fluently. In our research, we try to use word contexts in LDTS system to improve the synthesis and we compare its quality to the quality of a general TTS system.

Method: For the LD corpus building, we had large real data in disposal and we designed an algorithm [1] which selected a set of disjunctive text chunks (sequences of words) enhanced with their word contexts. Due to that, the context allows the system to find the optimal concatenation point of two chunks which naturally overlap in the synthesized sentence. During the corpus recording, we use a special approach of displaying selected chunks to the speaker to help him to read them in such a way to fit well into the original sentence. And even if they are a little different, the algorithm of speech synthesis has an overlap of chunks to find the best point of the concatenation [2].

Results: We carried out listening tests on selected sentences from a limited domain, for which the corpus containing contexts had been built and recorded. 57 listeners decided which of two sentence variants sounded better – one generated by the LDTS system and that by our general TTS system [3]. We got 280 answers which prefer the LDTS system to the general, 105 answers support the idea that the general TTS synthesis is better, and in 71 cases the two sentences were marked as of the same quality. The numbers show that the quality of LDTS system is higher. The validity of results was ensured using the statistical *sign test*.

Conclusions: Summing up the results, we confirmed the higher quality of the designed LDTS system working with the LD corpus containing contexts. It was not unexpected, because with concatenating of longer units, there are much fewer concatenation points in a synthesized sentence compared to the general synthesis and thus the probability of speech

artefacts is also lower. And due to the concatenating in overlaps caused by contexts, synthesized sentences sound very natural. On the other hand, in some cases the answers *"the general TTS system is better"* predominated. In our opinion, people prefer a (worse) constant quality synthesis to a natural-sounding sentence with only one little audible artefacts.

This work was supported by the European Regional Development Fund (ERDF), project ``New Technologies for Information Society'' (NTIS), European Center of Excellence, CZ.1.05/1.1.00/02.0090, and SGS-2013-032.

References:

[1] Jůzová, M. & Tihelka, D. (2014). Minimum Text Corpus Selection for Limited Domain Speech Synthesis. *Text, Speech and Dialogue 2014* (accepted).

[2] Jůzová, M. & Tihelka, D. (2014). Tuning Limited Domain Speech Synthesis Using General TTS System. *Text, Speech and Dialogue 2014* (accepted).

[3] Matoušek, J. & Tihelka, D & Ramportl, J. (2006). Current state of Czech text-to-speech system ARTIC. *Text, Speech and Dialogue 2006*, p. 439-446.

Artifact reduction in concatenation speech synthesis

Markéta Jůzová[1], Jakub Vít[1]
[1]University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
{juzova,vit89}@kky.zcu.cz

Research questions: This paper addresses problems of artifact reduction in concatenation speech synthesis. Unit selection is one of the techniques for generating synthetic speech. It is widely used and it is known for its ability to produce high-quality speech. However, it may suffer from sudden quality drops (so-called artifacts) at concatenation points. Although such quality problems can be reduced by using larger speech corpora, they can not be eliminated completely. Artifacts can be partially fixed by a signal modification method. Since every signal modification damages speech quality it must be done very carefully and preferably on as few signal segments as possible. Artifacts can have multiple causes so different signal modification is required for each of them. Finding positions of speech artifacts and a right way for fixing them is the question this paper discusses.

Method: The unit selection algorithm uses various criteria for selecting best sequence of units. One of them is a signal acoustic distance at a point of segment concatenation. Finding a sequence of units where all segments fit nicely to each other is not always possible. A higher acoustic distance should indicate a possible problem with the concatenation point. Unfortunately, this distance does not always reflect an actual perception by human listeners. To find out more features which can be used as indication of artifacts, a listening test experiment was carried out. Every listener was given a set of synthesized utterances. In each utterance, the listener should mark any phoneme-like segment that he/she found disturbing. Then a reference set of representative positive and negative artifacts was collected. In that set only those artifacts which were perceived by a majority of listeners were present. Every artifact sample was given set of features. These features included F0, energy, duration and spectral differences and some contextual features as the voiced/unvoiced characteristic, context and syllable characteristic. A standard two-class SVM classifier was trained based on the reference set using the described features. The classifier can now identify suspicious concatenation points for every unseen given utterance. Given the positions of the suspicious concatenation points, only the signal that corresponds to the suspicious concatenation points can be modified keeping other parts of signal unchanged. For example, the TD-PSOLA technique can now be used for re-synthesizing speech signal with F0, duration and energy contour smoothed at the concatenation points which were identified by the classifier.

Results: During the classifier training standard classification measures with cross-correlation were computed. F1 measure is over 0.8 which shows a very good ability to detect artifacts which will have an impact on the quality of a resulting synthetic utterance. Experiments with signal modification show that some of them could be fixed but more evaluation is needed to confirm this hypothesis.

Conclusions: A technique for coping with the presence of artifacts in synthetic speech was presented. This method uses a classifier which was trained from the data gathered to detect

audible artifacts and which identifies “suspicious” points in synthetic speech signal which may require a correction. In future work, signal modification will be subject of an experiment to find out a proper balance between a quality reduction caused by a signal modification and possible benefits from the artifact correction.

References:

Klabbers, E., Veldhuis, R (1998). *On the reduction of concatenation artefacts in diphone synthesis.* In: Proc. ICSLP, Sidney, Australia

Vít, J., Matoušek J. (2013). *Concatenation Artifact Detection Trained from Listeners Evaluations.* In *Text, Speech, and Dialogue 16th International Conference, TSD 2013, Pilsen*

Legát, M., Matošek, J. (2011). *Analysis of data collected in listening tests for the purpose of evaluation of concatenation cost functions*. In: Text, Speech and Dialogue. Volume 6836 of Lecture Notes in Computer Science. Springer, Berlin, Heidelberg

Chang, C.C., Lin, C.J.(2011). *LIBSVM: A library for support vector machines.*

Speech Quality Assessment in a Pronunciation Trainer for Speech Disorder Therapy

I. Kraljevski, R. Kompe, F. Kurnot, M. Rudolph, D. Hirschfeld[1], R. Jäckel, G. Strecha, R. Hoffmann[2]

[1]voiceINTERconnect GmbH, Dresden
[2]Institute of Acoustics and Speech Communication, TU Dresden
ivan.kraljevski@voiceinterconnect.de; rainer.jaeckel@tu-dresden.de

Usually, impaired speech is characterized by degraded intelligibility, deviating articulation rate, unprecisely and inconsistently pronounced segments, and declined voice quality. In this paper, we present a Computer-Aided Speech Therapy system for dysarthric speakers based on the concept of AzAR, a personal pronunciation trainer for foreign language learners. The system is re-implemented as a client-server platform, offering content sharing and exchange, supported by community based multimedia database. The users execute audiovisually presented exercises, such as uttering single words, word-pairs, phrases or texts passages, where pronunciation and prosody are compared against samples of reference speech. The intention is to emulate a training process which is usually conducted by a speech therapist by means of acoustic feedback.

The client-server architecture facilitates usage of PCs or mobile clients (tablets, smartphones or PDAs) for remote access to the exercises. The voice input is streamed to the server in real-time for the needs of audio processing, and the results are sent back to the client with the lowest possible latency. The requirement for real-time feedback indicates non-optimal estimation for some of the acoustic parameters (e.g. pitch and vowel formants). The LPC cepstrum is used in frame-based pitch determination and similarly LPC spectra peak picking for the formant tracks, along with median filter smoothing. For the phoneme segmentation, the whole utterance should be available in order to produce phoneme boundaries and likelihood scores. Trained acoustic models (HMMs) are used in phoneme scoring, where exercise specific finite-state grammars are employed for phoneme alignment.

For testing and evaluation of the speech processing and scoring concepts, a domain-specific speech corpus was created. The subjects are German native speakers divided into reference (2 healthy subjects), main (40 dysarthric subjects) and control (12 healthy subjects) group, recorded under comparable technical conditions, in surgery or studio environment. For articulation scoring, comparison of phoneme log-likelihood scores against the reference speech data was used. Pitch, short time energy, spectrogram and formants were normalized and mapped to the reference speech by DTW algorithm. The contribution of each measure to the overall score was statistically assessed over the available corpus. The Wilcoxon Mann-Whitney test showed that all measures significantly differ across the speakers groups (healthy and dysarthric). For phoneme articulation scores, the same statistical tests confirmed that there is a significant difference in phoneme articulation across the groups. In general, the control group achieved better averaged pronunciation scores than the test

group, which was expected. In order to get more conclusive results, each recorded sentence should be evaluated by human experts to assess the correlation with the automatically estimated quality scores.

References:

H. Strik, E. Sanders, M. Ruiter, and L. Beijer, "Automatic recognition of Dutch dysarthric speech: a pilot study," in Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '02), pp. 661–664, 2002.

Maier, A., Haderlein, T., et al., "Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer", EURASIP Journal on Audio, Speech, and Music Processing, Article ID 926951, 2010.

Donegan, M., (2000) Voice Recognition Technology in Education – Factors for Success, ACE Publications.

Vaquero, C., Saz, O., Rodrıguez, W. R., & Lleida, E. (2008). Human Language Technologies for speech therapy in Spanish language. Proceedings of the LangTech2008, 129-132.

Kitzing, P., Maier, A., & Åhlander, V. L. (2009). Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. Logopedics Phonatrics Vocology, 34(2), 91-96.

Skodda S., Visser W., Schlegel U.: Vowel articulation in Parkinson's disease. In: Journal of Voice, 2011, Vol. 25, No. 4, pp. 467-72.

Ackermann H., Ziegler W.: Die Dysarthrophonie des Parkinson-Syndroms. Fortschr. Neurol. Psychiatrie, 1989, 57: 149 – 160.

Forrest K., Weismer G., Turner G. S.: Kinematic, acoustic and perceptual analysis of connected speech produced by parkinsonian and normal geriatric adults. Journal of the Acoustic Society of America, 1989, 85: 2608 – 2622.

Simulating the Lombard-Effect in In-Car-Communication

Rabea Landgraf[1], Oliver Niebuhr[1], Gerhard Schmidt[2], Tina John[1], Christian Lüke[2], Anne Theiß[2]

[1]Dept. of General Linguistics, ISFAS, Kiel University, Germany
[2]Institute for digital signal processing and system theory, Kiel University, Germany
landgraf@isfas.uni-kiel.de

The Lombard effect has been subject to many different studies in the last decades. But, very little research was done on speech communication inside a car, and the modifications of speech production that are evoked by the noises of different driving speeds. To examine this kind of Lombard speech under controlled laboratory conditions, a unique acoustic ambiance simulation was designed at Kiel University (Kirat), in which speakers sit in a stationary car hearing the background noises of different driving situations (Lüke et al. 2013). The elaborate car-noise simulation is created through 8 loudspeakers and can be fully re-moved again from the speech signals after the recordings. Thus, no disturbing headphones are required to make high-quality recordings of Lombard speech that are suitable for all kinds of phonetic analyses. The aim of the present study was to compare our acoustic ambiance simulation to a real driving situation, and to investigate, if the speech in both situations is modified the same way as the noise level increases. That is, it was examined, if the Lombard effect that emerges in a driving car can be simulated in the laboratory.

A production experiment was conducted, with dialogue recordings made in the prepared car in the lab (acoustic ambiance simulation) and in a real driving situation. Each condition included driving speeds of 0 km/h, 50 km/h, 100 km/h and 150 km/h. A silent reference condition was recorded as well. Three pairs of speakers participated in the experiment. For eliciting separate spontaneous dialogues at each speed level, we used the video task paradigm developed by Peters (2001). Acoustic-phonetic analyses of the speech recordings were conducted covering a range of prosodic parameters.

The results confirm the assumed changes in speech production and interaction of the dialogue partners. Increases in driving speed/noise reduced the number of turns of the speakers, i.e. the interactivity of the conversation decreased and the turns becoming longer. Furthermore, speakers produced more hesitations and a larger amount of speech, but at a slower speaking rate. These qualitative changes in speech production emerged in both experimental conditions the acoustic ambiance simulation and the real driving situation. On the other hand, differences between the two conditions were found as well. In the lab condition, speakers produced less speech and spoke slower, despite more hesitations and more turns. Additionally, the data did not support the assumption that speakers habituate to the noise. Finally, individual differences were found concerning the order of magnitude of the Lombard effect.

The results show that the Lombard effect caused by the noise of a driving car can be simulated in our laboratory condition. We found similar modifications of speech production

and dialog interaction at different driving speed noises in the acoustic ambiance simulation and the real driving situation. There were only small differences concerning the quantity of the modifications between the two research situations. Moreover, while some of the found changes were known from the Lombard literature, we also revealed new Lombard speech parameters like the increase in the turn length.

References:

Lüke, Christian, Anne Theiß, Gerhard Schmidt, Oliver Niebuhr, Tina John. 2013. Creation of a Lombard speech database using an acoustic ambiance simulation with loudspeakers. *Proc. of the 6th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems*, S. 1-8. Seoul, Korea.

Peters, Benno. 2001. 'Video Task' oder 'Daily Soap Szenario' – Ein neues Verfahren zur kontrollierten Elizitation von Spontansprache. http://www.ipds.uni-kiel.de/pub_exx/bp2001_1/Linda21.html

Estimation of Articulatory Features for Czech language

Petr Mizera & Petr Pollák

Faculty of Electrical Engineering, Czech Technical University in Prague
mizerpet@fel.cvut.cz,pollak@fel.cvut.cz

Research questions: The issues of automatic speech recognition (ASR) aimed at the Czech language have been intensively studied in the past decades. The researches have successfully managed to develop several practical applications such as dictation programs [NOU09], automatic broadcast transcription (subtitling) [NOU06], [IRC01] and others. Accuracy of these ASR systems is generally satisfactorily high, however it is significantly lower in the case of spontaneous speech, implicit articulation or if the signal is corrupted, e.g. by high-level background noise or masking. These issues are still an obstacle for a wider usage of voice recognition technology under such conditions, because commonly achieved WER (Word Error Rate) of spontaneous speech recognition is above 50% in average. A possible solution to overcome this deficiency can be in the usage of speech production knowledge within ASR systems [KIN07]. Consequently, the speech production knowledge based on articulatory features (AFs) starts being used more often at feature level with the main purpose of improving the recognition of spontaneous or casual speech. The aim of our research is to analyse the possible contribution of articulatory features to the description of spontaneous or casual speech aimed for the Czech language.

Method & Results: Within the first steps of this research we focused on the estimation of AFs for the Czech language which is mostly carried out by Artificial Neural Networks (Multi-Layer Perceptron - MLP, Deep Neural Networks - DNN) [FRA07]. During our experiments we worked with the MLP-based AF classifiers trained and tested on speech signals from the Czech SPEECON database containing read speech [MIZ14]. The most important results of the first works can be summarized in the following points:

- The basic classes of AF features were defined in equivalent way as they were for English taking into account the specific peculiarities of phonetic standards for Czech.
- The analyses of MLP-based AF estimation accuracy for various input feature representations such as MFCC, PLP, TRAP were carried out. The optimum setup of MLPs for all particular AFs and for all analysed acoustic inputs was found.
- The best results achieved for particular Czech AF classes are as follows: voicing 95%, place consonant 86.5%, place vowel 89.4%, manner consonant 87.6%, manner vowel 88.3%, rounding 89.8, sonor 88.9%. The best frame level accuracy achieved across the articulatory feature classes was about 89,4%. Achieved results were compared to the results published by other authors for English.

Conclusions: The basic design of the articulatory features computation for the Czech language was performed. The further research will be focused on application of the articulatory features in tasks such as phoneme recognition (used frequently as the crucial

subpart of speaker and language recognition systems), phonetic segmentation [MIZ14], and for spontaneous and informal speech recognition (using Nijmegen Corpus of Casual Czech – NCCCz [ERN14]).

References:

[ERN14] Ernestus, M., Kockova-Amortova, L., Pollak, P.The Nijmegen Corpus of Casual Czech. *The 9th Language Resources and Evaluation Conference, Reykjavik, Iceland, May 2014.*

[FRA07] Frankel, J., et. al.: Articulatory feature classifiers trained on 2000 hours of telephone speech. *In proc. of Interspeech, Antwerp, Belgium, 2007.*

[IRC01] Ircing, P., et. al.: On large vocabulary continuous speech recognition of highly inflectional language - Czech. *In proc. of Interspeech 2001, Aalborg, pp. 487-490.*

[KIN07] King, S., et. al.: Speech production knowledge in automatic speech recognition. *Journal of the Acoustical Society of America, 121(2):723-742, February 2007.*

[NOU06] Nouza, J., Žďánský, J., Červa, P., Kolorenč, J.: Continual On-line Monitoring of Czech Spoken Broadcast Programs. *In: Proc. of Interspeech 2006, Pittsburgh, USA*

[NOU09] Nouza, J., Červa, P., Žďánský, J.: Very Large Vocabulary Voice Dictation for Mobile Devices, *In Proc. of Interspeech, 2009, Brighton, UK, pp. 995-998.*

[MIZ14] Mizera, P., Pollak, P.: Robust Neural Network Based Estimation of Articulatory Features for Czech. *SUBMITTED TO Neural Network World (accepted for the review, January 2014).*

[MIZ14] Mizera, P., Pollak, P.,Kolman, A., Ernestus, M.: Impact of Irregular Pronunciation on Phonetic Segmentation of Nijmegen Corpus of Casual Czech. *In Proc. of Text Speech and Dialogue (TSD), LNCS, 2014.*

Stepping out of the line in intonation – New insights into the forms and functions of F0 stylization

Oliver Niebuhr

Department of General Linguistics, ISFAS, Kiel University, Germany
niebuhr@isfas.uni-kiel.de

Research questions: Stylizing intonation means that F0 is flattened – typically at a medium or high level – over a number of syllables. Stylized intonation has, if at all, always played second fiddle in intonational phonetics and phonology. The most frequently studied stylization pattern is the so-called 'calling contour' (e.g., Abe 1962; Ladd 1978; Fagyal 1997; Gibbon 1998; Schaeffler & Biersack 2003; Day-O'Connell 2010, 2013; Niebuhr 2013). It consists of "a stepping down of one fairly steady level pitch to another" (Ladd 1978:517), each pitch level being associated with at least one syllable. So far, research has primarily focused on the functional part of the calling contour, while its intonational form is hardly questioned and associated "throughout the musicological and linguistic literature […] with a purportedly cross-cultural musical fingerprint: the interval of the minor third" (Day-O'Connell 2013:441). The present paper shows that this view is oversimplified both phonetically and phonologically. There are different types of calling contours; and since each of them has a different function, the paper also gives the functional debate about calling contours a new direction. More generally, the presented evidence clearly supports the position of Dombrowski (2013) that stylized intonations represent a separate inventory of forms and functions and are more frequent in speech communication than commonly thought.

Method: First, acoustic-phonetic analyses were conducted, based on enacted, quasi-spontaneous dialogues produced by two groups of eight native speakers of German. Calling contours were elicited without metalinguistic instruction just by the situational setting and the semantic-pragmatic context of the dialogues. The acoustic analyses focused on prosody and included syllable-related F0, duration, and intensity measurements. Then, perception experiments were performed, starting from naturally produced dialogue tokens whose calling contours were manipulated and resynthesized with respect to the results of the acoustic analyses. The stimuli were presented multiple times in randomized order and judged by two groups of overall 50 German listeners. Using the irony-judgments paradigm introduced by Landgraf (2014), different types of calling contours and their phonetic forms manifested themselves through matches/mismatches between contour function and utterance wording.

Results and Conclusions: The perception findings agree with the results of the acoustic analyses in revealing three different types of calling contours. They differ in the perceptual salience (higher/lower duration and intensity) of the two intonational plateaus, as well as in the plateaus' scaling relative to each other and to the preceding intonation contour. Moreover, contradicting the "minor-third rule", the step size down to the second plateau was for all three contour types positively correlated with the step size up to the first plateau. Functionally, the three contours are used by speakers to place special emphasis on the communication (channel), marking it as either well-functioning, impaired, or futile.

References:

Abe, I. (1962). Call contours. *Proc. 4th International Congress of Phonetic Sciences, Helsinki, Finland*, pp. 519-523.

Day-O'Connell, J. (2010). 'Minor third, who?': The intonation of the knock-knock joke. *Proc. 5th International Conference of Speech Prosody, Chicago, USA*, 1-4.

Day-O'Connell, J. (2013). Speech, Song, and the Minor Third: An Acoustic Study of the Stylized Interjection. *Music Perception, 30*, pp. 441-462.

Dombrowski, E. (2013). Semantic Features of 'Stepped' versus 'Continuous' Contours in German Intonation. *Phonetica, 70*, pp. 247-273.

Fagyal, Z. (1997). Chanting intonation in French. *Univ. of Pennsylvania Working Papers in Linguistics, 4*, pp. 77-90.

Gibbon, D. (1998). Intonation in German. In: D. Hirst & A. Di Cristo (Eds.), *Intonation systems: a survey of twenty languages*, pp. 78-95. Cambridge: Cambridge University Press.

Ladd, D.R. (1978). Stylized intonation. *Language, 54*, pp. 517-540.

Landgraf, R. (2014). Are you serious? Irony and the Perception of Emphatic Intensification. *Proc. 4th International Symposium on Tonal Aspects of Language, Nijmegen, The Netherlands*, 1-5.

Niebuhr, O. (2013). Resistance is futile – The intonation between continuation rise and calling contour in German. *Proc. 14th Interspeech Conference, Lyon, France*, 225-229.

Schaeffler, F. & Biersack, S. (2003). Aspects of the timing of fundamental frequency in German chanted call contours. *Phonum, 9*, pp. 129-132.

Perception of prosodic accentedness by native and non-native listeners

Václav Jonáš Podlipský & Šárka Šimáčková

Department of English and American Studies, Palacký University in Olomouc
vaclav.j.podlipsky@upol.cz, sarka.simackova@upol.cz

Research questions: Research on foreign accent has focused primarily on consonants and vowels although prosody substantially contributes to foreign-accentedness, e.g. [1, 2]. This study tests the perceptibility of prosodic accentedness in different groups of listeners. We ask whether (1) Czech learners of English can tell utterances produced by a native English speaker from their copies with imposed Czech-accented prosody (temporal dynamics: rhythm, segmental timing; intonation), whether (2) native English listeners with no knowledge of Czech or Czech-accented English will outperform the Czechs, and whether (3) native English listeners who share the dialect with the model speaker will in turn outperform those who do not.

Method: Twenty English sentences were elicited from a native speaker of British English and a Czech learner of English. Each original native-English sentence was resynthesized in Praat [3] to impose on it the Czech learner's temporal or tonal patterns. Listeners did not compare resynthesized sentences to the original ones. Instead, all sentences were resynthesized but with different proportions of native and non-native temporal or tonal properties. Two 'native'/non-native pairs were thus produced: (1) a sentence with 15% non-native durations (of each acoustically distinct portion, tempo-scaled, interpolated logarithmically) versus one with 100% non-native durations, and (2) a sentence with 35% versus one with 85% non-native intonation (interpolated in semitones). In addition, each sentence was mixed with speech-shaped noise at 7dB SNR. Twenty-four Czech advanced learners of English, four British, and four American listeners were presented with each pair of sentences, with a 750ms ISI, in both orders. The intonation and duration pairs were intermixed and their order was random for each listener. Listeners decided which version of the sentence sounded more English-like.

Results: Repeated-measures ANOVAs (and post-hoc Tukey HSD tests, where appropriate) revealed that the percentages of correct identification of the more native-like utterance (transformed into rationalized arcsine units [4]) were significantly higher for warped intonation than durations (both being above chance) and, in turn, the percentage of "can't decide" responses were lower for intonation than durations. Listeners' L1 alone did not have a significant effect on percent correct (which tended to be higher for British listeners) but it did affect percent incorrect (British listeners significantly lower than Czechs). Moreover, for correct responses Listeners' L1 interacted with the warped dimension: Czech and British listeners, but not American listeners, had more correct identifications for intonation than for durations, and Czechs had fewer correct identifications for intonation than British listeners. The effect of Listeners' L1 on reaction times approached significance: British listeners were faster than Czechs and Americans, who were comparable.

Conclusions: We showed that Czech learners of English could distinguish between native British and Czech prosody in English, despite their experience with Czech accent which could have decreased sensitivity to it. Secondly, the Czechs were outperformed by the British listeners (intonation being more perceptually salient than temporal dynamics for both groups). Finally, the British also outperformed the American listeners who had insufficient experience with both compared accents and did not gain an advantage by simply being native.

References:

[1] Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentals, Prosody, and Syllable Structure. *Language Learning*, 42(4), 529-555.

[2] Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38(2), 301-315.

[3] Boersma, P. & Weenink, D. (2014). Praat: doing phonetics by computer (Version 5.3.70). Retrieved from http://www.praat.org.

[4] Studebaker, G. A. (1985). A rationalized arcsine transform. *Journal of Speech, Language, and Hearing Research*, 28(3), 455-462.

How extra-linguistic factors affect pronunciation variation in different speaking styles

Barbara Schuppler

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
b.schuppler@tugraz.at

Research question: For the varieties spoken in Germany a lot of attention has been given to the study of pronunciation variation and reduction (e.g., Adda-Decker, M. & Lamel, L., 2000). For Austrian German, however, studies have been limited to prepared speech as spoken by trained radio speakers on the one hand and to very specific local dialects on the other hand. Only recently, we have collected read speech, elicited speech, and free conversations spoken by 38 speakers originating from the main cities of Austria (Schuppler et al., 2014). The aim of the current study is to investigate which extra-linguistic factors affect certain phonological- and reduction rules and whether their effect depends on the speaking style.

Materials and Method: Our preliminary results are based on 22 260 tokens from 12 speakers extracted from the Graz Corpus of Read and Spontaneous Speech (GRASS). For the final paper submission, this study will be based on in total more than 65 000 word tokens from 38 speakers. GRASS was manually transcribed on the orthographic level and automatically annotated with phonetic transcriptions by means of a forced alignment and a burst detector (Schuppler et al., 2014b). Based on these transcriptions, we analyze the distribution of 18 phonological- and reduction rules typical for standard Austrian German (e.g., vowel substitution/a:/ > /o:/, consonant lenition /b/ > /v/, etc.). For this purpose, we build mixed effects logistic regression models with speaker and word as random variables (Jaeger, 2008).

Preliminary Results: Based on 22 260 tokens from 12 speakers, we observed that pronunciation variation is in general more pervasive in read (33.1 %) than in conversational speech (63.2 %). This result is as expected. For certain rules, however, we observed no significant differences in frequency of occurrence between the speaking styles (e.g., devoicing of fricatives, vocalization and deletion of /r/). With respect to the extra-linguistic variables, we observed in general significant effects of age in all speaking styles and of social and regional background in the conversational speech. Our detailed analysis for each of the rules, however, showed a more complex picture of dependencies.

Conclusions: This study presents the first quantitative analysis of pronunciation variation in Austrian German and reports which extra-linguistic factors affect the kind of variation observed. The presented methods for modeling pronunciation variation will be incorporated into an Automatic Speech Recognition (ASR) system.

References:

Adda-Decker, M. & Lamel, L. (2000). Modeling reduced pronunciations in German. *Phonus 5*, Institute of Phonetics, University of the Saarland, p. 129-143.

Jaeger, T.F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, pp. 434–446.

Schuppler, B., Hagmueller, M., Morales-Cordovilla, J. A. & Pessentheiner, H. (2014). GRASS: The Graz Corpus of Read and Spontaneous Speech. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Iceland, pp. 1465-1470.

Schuppler, B., Grill, S., Menrath, A. & Morales-Cordovilla, J. A. (2014). Automatic phonetic transcription in two steps: forced alignment and burst detection. Accepted for presentation at 2nd *International Conference on Statistical Language and Speech Processing (SLSP'14)*, Grenoble, France.

Optimizing formant extraction in Praat and Snack: Comparison of manual and automatic measurements

Radek Skarnitzl, Jitka Vaňková & Tomáš Bořil

Institute of Phonetics, Faculty of Arts, Charles University in Prague
{radek.skarnitzl|tomas.boril}@ff.cuni.cz, jitka.vanka@gmail.com

Research questions: The analysis of vowel formants continues to be an important area of research in speech science, whether for purely linguistic-phonetic, sociophonetic, or forensic purposes. However, reliable extraction of formant values still remains a problem. Although software packages for phonetic analysis do recommend "default" settings for formant extraction (Sjölander, 2004; Boersma & Weenink, 2014), the empirical foundation of such settings is not straightforward. Any attempts to compare algorithms and their settings require a reference set of manually-labelled formant values, what has been called the "ground truth" by Deng et al. (2006). Deng and his colleagues, in what to our knowledge is the only reference set of this kind, unfortunately do not provide any detail on how the manual measurements were conducted. The first aim of our study therefore was to prepare a more basic database of hand-labelled vowel formant values in Czech but with more explicit guidelines for the labellers. The second aim was then to compare the "ground truth" values with automatically extracted formant values under different settings and, possibly, to suggest settings which outperform the default ones.

Method: Vowel formants (F1–F3) in 5 tokens of each of the 5 short Czech monophthongs from 10 speakers were manually determined based on, primarily, spectrogram information by four labellers (every vowel was labelled by two labellers and consensus was sought when formant values differed by more than 100 Hz). Finally, the "ground truth" values were obtained by averaging the pairs. Formant values were then automatically extracted using several settings in Praat (Boersma & Weenink, 2014) and Snack (Sjölander, 2004), and these values compared with the reference set.

Results: The following preliminary description is based only on the comparison of Praat extraction with manually obtained values. The different settings of formant extraction yielded very divergent differences from the manual measurements. For example, while the default setting for extracting formants of female voices (10^{th} LPC order, with maximum formant frequency being 5.5 kHz) leads to mean differences of F1–F3 of 53, 55 and 132 Hz, respectively, changing the maximum frequency to 5 kHz leads to mean differences of 49, 151 and 395 Hz.

Table 1 compares the default settings for females (10^{th} LPC order in the 0–5.5 kHz range) and for males (10^{th} LPC order in the 0–5 kHz range) with that setting which yielded the lowest differences from the manual measurements. These most successful settings – 9^{th} order LPC and extraction in the 0–5 kHz range for female voices and 11^{th} order LPC in the 0–5.5 kHz range for male voices – led to a slight improvement over the default settings. Some interesting formant-specific and vowel-specific differences may be observed.

	females	F1	F2	F3	males	F1	F2	F3
[ɪ]	*LPC 10, 5.5 kHz*	*34.1*	*46.2*	*101.0*	*LPC 10, 5 kHz*	*57.6*	*51.0*	*61.2*
	LPC 9, 5 kHz	**33.9**	**43.9**	**52.6**	**LPC 11, 5.5 kHz**	**56.9**	**45.1**	**60.0**
[ɛ]	*LPC 10, 5.5 kHz*	*92.0*	*82.5*	*110.4*	*LPC 10, 5 kHz*	*72.6*	*59.3*	*97.4*
	LPC 9, 5 kHz	**96.3**	**76.6**	**78.4**	**LPC 11, 5.5 kHz**	**66.4**	**43.0**	**68.9**
[a]	*LPC 10, 5.5 kHz*	*64.4*	*51.4*	*212.4*	*LPC 10, 5 kHz*	*38.8*	*25.9*	*86.7*
	LPC 9, 5 kHz	**61.6**	**46.7**	**158.8**	**LPC 11, 5.5 kHz**	**39.2**	**26.7**	**65.4**
[o]	*LPC 10, 5.5 kHz*	*34.3*	*38.8*	*162.9*	*LPC 10, 5 kHz*	*30.7*	*30.9*	*55.9*
	LPC 9, 5 kHz	**32.7**	**37.9**	**147.6**	**LPC 11, 5.5 kHz**	**30.6**	**33.4**	**51.7**
[u]	*LPC 10, 5.5 kHz*	*36.6*	*54.2*	*75.1*	*LPC 10, 5 kHz*	*49.4*	*27.4*	*103.8*
	LPC 9, 5 kHz	**36.9**	**41.8**	**80.5**	**LPC 11, 5.5 kHz**	**48.9**	**26.7**	**92.9**

Table 1: Comparison of default (italicized) and the most successful (boldfaced) formant extraction settings in Praat for female and male voices.

Conclusions: The presentation will also include results from the Snack algorithm and formant tracker embedded in Praat. In addition, we will discuss problems related to manual measurement uncertainty, such as the frequency resolution of the Fourier transform and inaccuracies caused by technical limitations.

References:

Boersma, P. & Weenink, D. (2014). *Praat: doing phonetics by computer* (Version 5.3.77). Retrieved from http://www.praat.org.

Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., Chen, Y. & Alwan, A. (2006). A database of vocal tract resonance trajectories for research in speech processing. In: Proceedings of ICASSP 2006, pp. 369-372.

Sjölander, K. (2004). *The Snack Sound Toolkit*. Retrieved from http://www.speech.kth.se/snack/

Articulatory data on the syllable affiliation of Czech alveolar consonants

Pavel Šturm

Institute of Phonetics, Charles University in Prague
pavel.sturm@ff.cuni.cz

Research questions: It is commonly assumed (e.g. [1: 271]) that syllable boundaries in Czech words follow the Maximum Onset Principle [2], especially with single consonants between vowels. The present article aims to verify or rectify this common assumption experimentally by investigating the articulatory correlates of syllable affiliation of four different alveolar consonants.

Method: EPG linguopalatal contact patterns were obtained from nine native speakers of Czech (5 female, 4 male, all employees or students at the Institute of Phonetics) using the EPG3 system [3]. The alveolars /t s n l/ were investigated in five prosodic contexts (the target sound – here /n/ – is underlined):

1. word-final before a pause	/ˈvɪdaːn̲/
2. word-final linked to a V-initial word (no [ʔ])	/ˈvɪdaːn̲‿ˈavɛrzɪ ˈfʃan͡ts/
3. word-initial & stress group-initial	/ˈnɛxaː ˈn̲avɛt͡ʃɛr ˈkus/
4. word-initial & stress group-medial	/ˈmaː n̲a ˈvɛlɪko͡u ˈpo͡uc/
5. word-medial	/ˈpaːn̲a ˈvɛsɛlɛ ˈkrɪl/

Subjects read the items grouped according to the contexts. Each item was repeated three times, and the whole list was read twice. The design resulted in 120 target realizations from each speaker (5 blocks × 4 consonants × 6 repetitions), in addition to the practice items that preceded each block. The degree of contact was measured in the alveolar area (rows 1 to 4) at 5 points during the target consonant production, and various data-reduction indexes, e.g., CA, CP, COG (see [4]), were computed.

Results: The experiment did not reveal any context-specific EPG patterns for the target sounds /t/, /s/ and /n/ in word-initial, -medial and -final positions (only for word-final before a pause). However, realizations of /l/ were highly position-sensitive. The word-initial position was associated, for instance, with greatest alveolar contact and greatest posteriority of the contact, while the word-final position showed lowest values. The word-medial articulation fell between the word-initial and the word-final in terms of all measures used. Intervocalic /l/ within a word thus seems to behave neither like an initial onset, nor like a final coda. In addition, contact patterns revealed (1) interesting results for the articulation of the lateral /l/ in general; (2) great variability between speakers (for all target sounds); (3) a significant interaction of context and speaker.

Conclusions: EPG investigation of the articulation of alveolar consonants in various positions yielded possibly syllable-related differences only for the lateral /l/. Compared to positions with straightforward syllable affiliation (word-initial and -final), word-medial /l/ was articulated in an intermediate fashion. This mixed evidence of its affiliation might be interpreted as an effect of ambisyllabicity or, alternatively, prosodic factors such as stress and boundary strength.

References:

[1] Palková, Z. (1994). *Fonetika a fonologie češtiny*. Praha: Karolinum.

[2] Blevins, J. (1996). The syllable in phonological theory. In: J. A. Goldsmith (Ed.), *The Handbook of Phonological Theory*, pp. 206-244. Oxford: Wiley-Blackwell.

[3] Wrench, A. (2006). *Artic Assist* (Version 1.14). Edinburgh: Queen Margaret University College.

[4] Fontdevila, J., Pallarès, M. D. & Recasens, D. (1994). The contact index method of electropalatographic data reduction. *Journal of Phonetics*, 22, pp. 141-154.

Looking back at the 6th ICPhS in Prague, 1967: What does it tell us about phonetics?

Pavel Šturm

Institute of Phonetics, Charles University in Prague
pavel.sturm@ff.cuni.cz

Purpose: The tradition of phonetic congresses dates as far back as 1932, when the first ICPhS was held in Amsterdam. Being a pivotal point in the organized life of phonetics, it gave rise to a regular series of creative gatherings of phoneticians and speech scientists from across various disciplines. The ultimate aim is to encourage cooperation among colleagues from different countries and fields and provide space for any well-founded phonetic research and, more importantly, for its fruitful discussion. In Peter Ladefoged's words, although we cannot expect to find all the answers to our questions, "every four years we can get together and pool our knowledge" [cited in 1: 28]. We may only speculate what the phonetic pioneers would think of the forthcoming 18th ICPhS; it will certainly be quite different from the path they trod. In an attempt to simulate this situation, the current paper presents a commented comparison of the 6th ICPhS held in Prague in 1967 with the most recent meeting, the 17th ICPhS in Hong Kong, 2011.

Method: Both congresses are compared in terms of participants, presented papers and topics [2, 3]. The course of the Prague congress was reconstructed and put into historical context. Most importantly, its participants that could still be contacted were asked to respond to the following questions:

> *1) How do you evaluate the congress in general? Can you compare it to some other congresses in terms of both organization and quality of speakers?*
>
> *2) What was the highlight for you personally? Was there anything particular, specific to the Prague congress?*
>
> *3) Were you acquainted with anyone from the Czech phonetic community before the congress? (and/or did you get acquainted with anyone during the congress?)*
>
> *4) What are your most memorable experiences from the boat trip along the river, the banquet or any other aspect of the social programme?*
>
> *5) How did you arrive in Prague? Did you have any difficulties with the authorities? How did you feel during the congress given the political situation at that time?*
>
> *6) What are your views on the subsequent development of phonetics or related disciplines?*

Results: Predictably, the 2011 congress scored higher as to number of participants and papers. Likewise, the range of topics has broadened and changed over time, with a much larger emphasis on prosody and new research areas unlatched by technological developments (including statistical methods). However, the respondents, long-time phoneticians, were not always happy with the new trends in phonetics and current publication practices, and spoke highly of the Prague congress and the general approach to phonetic work at that period.

Conclusions: International phonetic congresses evolve along with the field. The 6th congress was an honour to Prague, while the 17th congress acknowledged the increasing significance of Asia as a key player in the area of phonetics. The comparison proved to be instructive, and revealed interesting differences.

References:

[1] Ohala, J. (2000). Phonetics in the free market of scientific ideas and results. *Journal of the IPA*, 30, pp. 25-29.

[2] Hála, B., Romportl, M. & Janota, P. (Eds.) (1970). *Proceedings of the Sixth International Congress of Phonetic Sciences*. Prague: Academia.

[3] Lee, W.-S. & Zee, E. (Eds.) (2011). *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong: City University of Hong Kong.

Commonly confused words in spontaneous Czech language transcription and recognition

Tomáš Valenta & Luboš Šmídl

Department of Cybernetics, Faculty of Applied Sciences,
University of West Bohemia in Pilsen
valentat@kky.zcu.cz, smidl@kky.zcu.cz

Research questions: During our study of inter-annotator agreement of spontaneous Czech language (Valenta, Šmídl, Švec, & Soutner, 2014), a list of words that the annotators confused most frequently was created. A similar list was created for results of an automatic speech recognizer. In this study, we compare the contents of the lists and discuss the confusions.

Method: The inter-annotator agreement can be studied using the same methods as in the evaluation of the automatic speech recognition accuracy: at first, the best alignment between the transcriptions (or between the reference and the recognition hypothesis) with minimal Levenshtein editing distance (using insertion, deletion and substitution operations) is found (Levenshtein, 1966). Then the percentage of matching words is the agreement, or the accuracy. All pairs of substituted words were counted and just the ones that do not change the meaning when heard (although may change the meaning when read) were manually picked. Finally the confusion lists were made of them. The methods described above were applied on the Toll-free calls corpus consisting of telephone communication between two people. The people know each other very well so they use lots of local words and non-verbal sounds. They speak expressively and colloquially.

Results: The average inter-annotator agreement (IAA) on the corpus measured among three annotators was 86 %. The average accuracy of an automatic speech recognizer, taking progressively each annotation as a reference, was 49 %. If the confusion lists were not taken into account during the evaluation, the IAA and the accuracy were 5 % and 3 % lower respectively (Valenta et al., 2014). The annotations were made for acoustic modelling (as precise as possible), so the words confused by humans were basically homonyms (jsem–sem, let–led), different grammatical functions (byli–byly or bili–bily), incorrect pronunciations (říkám–říkam–řikám–řikam), colloquialisms (vždycky–dycky, být–bejt, osm–osum–vosum) or splits and merges (takže–tak že). The confusion list of words that could not change meaning, was similar, but shorter for the recognizer. The remaining items on the confusion lists could change the meaning (e.g. to–co, a–ale, si–se) and hence were the very matter of the inter-annotator disagreement or the recognition accuracy decrease. Generally, shorter words are confused more frequently both by humans and by the recognizer.

Conclusions: The inter-annotator agreement sets the upper bound of the speech recognition accuracy. Recognition accuracy higher than IAA would mean an overfitting problem of the recognizer. The mostly confused words by humans usually do not change the meaning when spoken, but people have problems to agree on their written form.

Acknowledgement

This research was supported by the Technology Agency of the Czech Republic, project No. TE01020197.

References:

Valenta, T., Šmídl, L., Švec, J., & Soutner, D. (2014). Inter-annotator Agreement on Spontaneous Czech Language, Limits of Automatic Speech Recognition Accuracy. In P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Text, Speech and Dialogue 2014*.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady,* 10, 707.

Vowel-related glottalization in read speech of native and non-native speakers of Czech

Jitka Veroňková & Yana Tolkunova

Institute of Phonetics, Charles University in Prague
jitka.veronkova@ff.cuni.cz

Research questions: Glottalization phenomena in the sense of a significant irregularity of glottal pulsing have been examined in numerous studies. Glottalization fulfils a number of linguistic functions and can occur in various contexts. Onset-related glottalization signals boundaries of words or morphemes beginning with vowels, for example in German [1], English [2], [3] and also in Czech. According to the Czech norm, the pronunciation of glottal stop is optional in most contexts, but it is recommended because it supports comprehensibility [4], [5]. Glottalization is also the object of second language acquisition, e.g. [6], [7], for Czech speakers [8], [9].

This contribution examined the rate of vowel-related glottalization of Russian speakers in Czech read speech in comparison with Czech native speakers. Czech belongs among the languages with relatively frequent glottalization [10]. According to [11] glottalization is quite common also in Russian, especially at boundaries of intonational phrases. The realization of vowel-related glottalization may include not only the canonical glottal stop, but also the other glottalization phenomena [12], [1], for Czech [13], for Russian [11].

The purpose of the study was to find out whether there are similarities or distinctions in the distribution of glottalization between native and non-native speakers of Czech and to examine the factors that could influence it [8], [11], [12].

Method: The short read text contained 14 potential positions where glottalization could occur in standard pronunciation. It covered several types of segmental contexts, and it took into account the structural position of word and phrasal unit. Two groups of non-professional speakers produced the texts: non-native speakers of Czech with Russian as a mother tongue (6 males and 6 females) and native speakers of Czech (4 males and 7 females). 322 tokens were analysed and glottalization was rated. The analysis included two main categories – glottal stop and creak – and was primarily based on perception: the significant salient impression of a glottal gesture had to be present.

Results: The rate of glottalization ranged from 71.4 to 100.0% (native group) and from 21.4 to 58.3% (non-native group). The significance was calculated using t-tests: the differences between native and non-native speakers are significant at the level $p < 0.05$, the differences between males and females are not significant. Glottalization of single tokens can be influenced by the segmental, prosodic, and lexical context.

Conclusions: The presented contribution was supposed to be a pilot study in the research of glottalization in the speech of non-native speakers of Czech. Three main directions for subsequent studies are considered: a) an acoustic analysis of the material, b) more detailed examination of the single types of segmental context and c) the examination of glottalization in (semi)-spontaneous speech. In the case of b) the examination of glottalization after non-syllabic prepositions is in process. The lexicon is taken into account as another factor.

References:

[1] Kohler, K. J. (1999). Plosive-related glottalization phenomena in read and spontaneous speech - A stød in German? *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 34*, pp. 281-321.

[2] Dilley, L. & Shattuck-Hufnagel, S. (1995). Variability in glottalization of word onset vowels in American English. In: *Proceedings of the ICPhS* Stockholm, pp. 586-589.

[3] Bissiri, M. P., Lecumberri, M. L., Cooke, M. & Volín, J. (2011). The Role of Word-Initial Glottal Stops in Recognizing English Words. In: *Proceedings of the 12th Annual Conference of ISCA Interspeech*, pp. 165-168. Florence: ISCA.

[4] Hála, B. (1967). *Výslovnost spisovné češtiny I. Výslovnost slov českých.* Praha: Academia, 2nd edition.

[5] Palková, Z. (1997). *Fonetika a fonologie češtiny*. Praha: Karolinum, 2nd edition.

[6] Balas, A. (2011). Glottal stops produced by Polish native speakers in Polish and in English. In: *Proceedings of the ICPhS XVII* Hong Kong, pp. 280-283.

[7] Volín, J., Uhrinová, M. & Skarnitzl, R. (2012). The effect of word-initial glottalization on word monitoring in Slovak speakers of English. *Research in Language, 10/2*, pp. 173-181.

[8] Bissiri, M.P. & Volín, J. (2010): Prosodic structure as a predictor of glottal stops before word-initial vowels in Czech English. In: R. Vích (Ed.), *20th Czech-German Workshop - Speech Processing*, Prague, pp. 23-28

[9] Skákal, L. (2011). *Užívání hlasivkového rázu u rodilých a nerodilých mluvčích francouzštiny*. Institute of Romance studies, Charles University in Prague. Unpublished bachelor thesis.

[10] Volín, J. (2012). Jak se v Čechách „rázuje“. *Naše řeč 95/1*, pp. 51-54.

[11] Krivnova, O. F. (2005). Lari[y]ngealization as a boundary marker in oral speech. In: *XVI Session of the Russian Acoustical Society*, pp. 546-549 [on-line 2014-04-15]

[12] Redi, L. & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics, 29*, pp. 407-429.

[13] Skarnitzl, R. (2004): Acoustic categories of nonmodal phonation in the context of the Czech conjunction “a”. In: Z. Palková & J. Veroňková (Eds.), *AUC Philologica 1/2004, Phonetica Pragensia X*, Praha: Karolinum, pp. 57-68

DCT based Voice Conversion with Multipoint Frequency Transformation

Robert Vích & Jan Staněk

Institute of Photonics and Electronics, AS CR, Prague
vich@ufe.cz

In the proposed paper a new voice conversion algorithm will be presented which transforms the utterance of a source speaker into the utterance of a target speaker.

In the past we used for voice conversion:

- Discrete Fourier Transform (DFT)
- Speech modeling based on the real cepstrum
- Speech modeling based on the complex cepstrum

Instead of the application of the mentioned approaches, we may use for voice conversion also another transform coding, the Discrete Cosine Transform (DCT).

The proposed new voice conversion approach is based on pitch synchronous speech analysis, DCT, multipoint unitary frequency transformation with linear interpolation and energy preservation, inverse DCT (IDCT) and pitch synchronous speech synthesis with overlap and add. The proposed voice conversion does not change the suprasegmental speech parameters. The prosody can be changed globally by statistical means or it can be modeled separately, e.g. as in emotional speech synthesis.

The DCT speech spectrum is, in contrary to the DFT speech spectrum, a real sequence but also contains the full information about the DFT speech spectrum, i.e. the magnitude and the phase. Moreover, the DCT speech model can be adaptively frame-wise compressed with the aim of memory reduction. For DCT voice conversion we use the unitary DCT type-2. The DCT computation will be illustrated on a simple example and the main properties of the DCT spectrum will be summarized.

For the change of the speaker type, e.g. a male into a female voice, i.e. of the speaker's segmental parameters, frequency transformation of the DCT source speech spectrum is used. The multipoint unitary frequency transformation uses several pairs of significant reference frequencies obtained from the smoothed spectra of the source and target speakers by means of Welch's averaged modified periodogram method of spectrum estimation. In another approach, the optimal frequency warping function is searched by Dynamic Frequency Warping, which is dual to Dynamic Time Warping.

The finite impulse response of the converted DCT speech model is obtained by the IDCT of the frequency transformed DCT speech model. It is of the mixed phase type.

The proposed voice conversion procedure is very fast and characterized by easy implementation with less computational requirements than the FFT-based conversion or the complex cepstrum implementation. It results in speech with high naturalness. Voice conversion examples of natural male and female speech and of the synthetic speech obtained by the Czech text-to-speech synthesis of the Latin text Gaudeamus Igitur will be presented.

The aim of this method is the generation of new voices in text-to-speech speech synthesis without the generation of new inventories for new speakers or for voice modification in different applications and in voice coding.

Fundamental Frequency Parameters in Native and Foreign-Accented Speech

Jan Volín & Hana Bartůňková

Institute of Phonetics, Faculty of Arts, Charles University in Prague
jan.volin@ff.cuni.cz

Research questions: Intonation is currently studied primarily through fundamental frequency contours (F0 tracks). Various aspects of speech melodies (e.g., forensic, affective, conversational) are often expressed in terms of the F0 mean and variance, although the search for other F0 descriptors has resulted in some more noteworthy suggestions (Gårding, 1983; Lieberman et al., 1985; Hermes, 2006; Adami, 2007; Lindh & Eriksson, 2007). The present study tests various fundamental frequency parameters to find out whether they can capture differences among contours produced in native Czech, native English and Czech-accented English. Apart from the overall tendency, the development across paragraphs is investigated together with the detailed analyses of individual units.

Method: Sixteen female non-professional speakers were asked to read out previously transcribed realistic news bulletins in Czech and English. Eight of the subjects were native speakers of English, eight were native speakers of Czech who, apart from the Czech bulletins, also read the English ones as they were also learners of English. The recordings were split into breath-groups, and F0 tracks were extracted in Praat (Boersma & Weenink, 2014) using the autocorrelation algorithm. The tracks were turned into *pitchtier* objects and manually corrected. Measures of central tendency (mean, median), variability (pitch range, percentile range, quartile range), and other descriptors (Lindh-Erikson base value, regression line capturing downtrends) were computed and compared across the three modes.

Results: The results are not as straightforward as various speech acquisition models suggest. Contrary to expectations based on generally held views, many of the Czech speakers displayed greater variation in F0 values. Some of the speakers raised their mean, presumably under the strain of foreign language usage. Thus, native English speakers' tracks differed from the Czech group's production of both their foreign-accented English or their native Czech, but this difference was not always in the direction one would expect based on currently held general views of Czech and English intonation.

Conclusions: Native speakers of English often describe Czech intonation as dull or signalling boredom. Our data show that this impression is not directly reflected in most of the F0 descriptors currently used. Weingartová et al. (2014) showed that prominence in Czech English is less contrastive than in native English. However, this does not seem to apply to the F0 contour as a whole. Nevertheless, a detailed, linguistically informed selective analysis of individual melodic events including some of the downtrends provides an optimistic view of the matter and shows that the research in the area is promising. Our present experiment has to be complemented by perception testing of liveliness, or the perceived degree of involvement of the speaker in the uttered propositions.

References:

Adami, A.G. (2007). Modelling prosodic differences for speaker recognition. *Speech Communication, 49*, 277–291.

Boersma, P. & Weenink, D. (2014). *Praat: doing phonetics by computer* (Version 5.3.62). Retrieved from http://www.praat.org.

Gårding, E. (1983). A generative model of intonation. In: A.Cutler & D.R.Ladd (Eds.) *Prosody: Models and Measurements*, pp. 11-25. Berlin: Springer-Verlag.

Hermes, D.J. (2006). Stylization of pitch contours. In: S. Sudhoff et al. (eds.), *Methods in empirical prosody research*. Berlin: Walter de Gruyter, 29–61.

Lieberman, P. et al (1985). Measures of the sentence intonation of read and spontaneous speech in American English. *JASA, 77/2*, pp. 649-657.

Lindh, J. & Eriksson, A. (2007). Robustness of long-time measures of fundamental frequency. In: *Proceedings of Interspeech 2007*, pp. 2025–2028. Antverp: ISCA.

Weingartová, L., Poesová, K. & Volín, J. (2014). Prominence Contrasts in Czech English as a Predictor of Learner's Proficiency. In: N. Campbell, D. Gibbon & D. Hirst (Eds.), *Proceedings of Speech Prosody 2014*, pp. 236–240. Dublin: TCD.

Amplitude Differences in Polysyllabic Words of Czech English

Lenka Weingartová & Jan Volín

Institute of Phonetics, Charles University Prague, Czech Republic
{lenka.weingartova| jan.volin}@ff.cuni.cz

Research questions: Prosodic variation in languages fulfills numerous roles, one of which is marking prominences. Prominence structures were shown to serve crucial functions in speech perception. One of the reasons for compromised intelligibility of foreign-accented speech is its atypical metrical patterning. Of the four dimensions of prominence (frequency, time, intensity, spectrum), intensity is the most disputed one (e.g., Fry, 1958; Havránek & Jedlička, 1966; Janota & Palková, 1974; Beckman, 1986; Kochanski et al., 2005; Andreeva & Barry, 2012). The present study investigates the amplitude differences between stressed and unstressed syllables in Czech English relative to native English and native Czech.

Method: Our previous study explored the global differences between pairs of stressed and adjacent unstressed syllables in longer texts without discriminating between the structural types and semantic classes of words (Weingartová et al., 2014). Currently, we focus on selected words and expand our attention to amplitude trajectories throughout the word and to comparable words in native Czech. Two- to four-syllable words were taken from 21 recordings of news bulletins in Czech, English and Czech English read by female non-professional speakers. The boundaries of phones and the placements of word stresses were manually labelled, and mean SPL (Sound Pressure Level) was measured in *Praat* (Boersma & Weenink, 2014) in the middle third of each vowel duration.

Results: Whereas the mean difference between the stressed and adjacent unstressed syllable measured globally throughout the text in the earlier study was by 2.8 dB greater for the native English speakers than for the Czech speakers of English, our refined material shows disparate behaviour for structurally different words. For instance, initially stressed three-syllable words produced an average difference of 5.7 dB between Czech English and native English. Detailed analysis revealed that this effect is caused specifically by the word *president*. Similarly, disyllabic words with stress on the second syllable yielded a mean difference of 2.3 dB, however, the word *reports* behaves differently from the word *against*. Other word-specific effects seem to be linked with the tense character of English fortis obstruents as opposed to the Czech.

Conclusions: In parallel with the findings of Volín (2005) and Volín and Poesová (2008) for temporal domain, Czech speakers of English exhibit different behaviour from the native speakers, but neither of the groups treats intensity in stressed syllables uniformly across the lexicon. The stress placement on the first syllable, which is canonical in the Czech language, does not lend any advantage to English learners for initially stressed English words. As for the Czech polysyllabic words, the amplitude differences between stressed and unstressed syllables are smaller than in English, although not negligible, and longer words display reversed tendencies relative to shorter words.

References:

Andreeva, B. & Barry, W. (2012): Fine phonetic detail in prosody. Cross-language differences need not inhibit communication. In: O. Niebuhr (Ed.), *Prosodies - context, function, and communication*, pp. 259–288, Berlin/New York: de Gruyter.

Beckman, M. E. (1986). *Stress and Non-Stress Accent.* Dordrecht: Foris.

Boersma, P. & Weenink, D. (2014). *Praat: doing phonetics by computer* (version 5.3.71). Retrieved from http://www.praat.org.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech, 1*, pp. 126–152.

Havránek, B. & Jedlička, A. (1966). *Stručná mluvnice česká.* Praha: Fortuna.

Janota, P. & Palková, Z. (1974). Auditory evaluation of stress under the influence of context. *AUC Philologica 2/1974, Phonetica Pragensia, 4*, pp. 29-59.

Kochanski, G., Grabe, E., Coleman, J. & Rosner, B. (2005). Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustical Society of America, 118 (2),* pp. 1038–1054.

Volín, J. & Poesová, K. (2008). Temporal and spectral reduction of vowels in English weak syllables. In: A. Grmelová, L. Dušková, M. Farrell & R. Pípalová (Eds.), *Plurality and Diversity in English Studies*, pp. 18–27, Praha: UK PedF.

Volín, J. (2005). Rhythmical properties of polysyllabic words in British and Czech English. In: J. Čermák, A. Klégr, M. Malá and P. Šaldová (Eds.), *Patterns: A Festschrift for Libuše Dušková,* pp. 183–194. Praha: Kruh moderních filologů.

Weingartová, L., Poesová, K. & Volín, J. (2014). Prominence Contrasts in Czech English as a Predictor of Learner's Proficiency. In: N. Campbell, D. Gibbon & D. Hirst (Eds.), *Proceedings of Speech Prosody 2014*, pp. 236–240. Dublin: TCD.