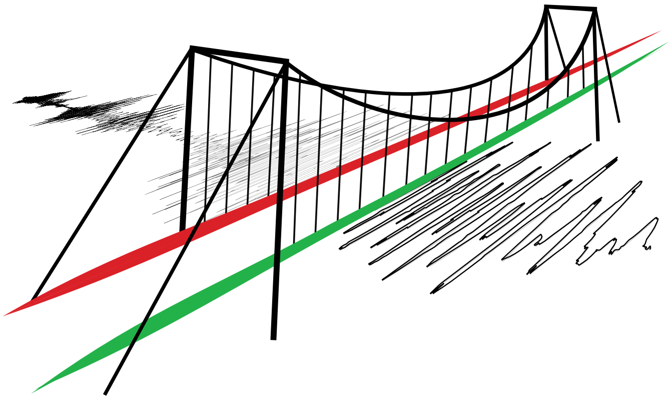


CAPSS2017

BOOK OF ABSTRACTS

Workshop on
Challenges in
Analysis and Processing of
Spontaneous Speech

Budapest, May 14–17, 2017

Contents

Program

Abstracts

Monday

Tuesday

Wednesday

CAPSS
Challenges in analysis and processing of spontaneous speech

In view of the rapid growth of various aspects in speech research, we will address the issue of challenges in the analysis and processing of spontaneous speech. Topics include a wide range of research from phonology through speech production and perception processing to speech technology. All submissions are expected to be related to events, processes and applications of spontaneous speech. The workshop will be a unique exchange forum for researchers working in all kinds of research fields focusing on relevant questions of spontaneous speech. Postgraduate students are particularly encouraged to attend the workshop. The organizers will offer the 'CAPSS Prize of the best young presenter'. This workshop is intended as the first event of a series of CAPSS workshops planned to take place in every two years in the future.

The first workshop will take place May 14–May 17 2017 in Budapest, Hungary. The event is organized by the Department of Phonetics of the Institute for Linguistics, Hungarian Academy of Sciences.

A special issue of *The Phonetician* containing selected papers of CAPSS is planned to be published in 2018.

Applications for oral papers, poster papers, and demonstrations are welcome. Two-page abtracts are to submitted. The topics of the workshop include - but are not limited to:

- Phoneme realizations in spontaneous speech
- Phonetic properties in spontaneous speech
- Coarticulation phenomena in spontaneous speech
- Prosodic structure of spontaneous speech
- Phonetic coherence in spontaneous speech
- Narratives vs dialogues
- Dysfluency phenomena in spontaneous speech
- Sound changes occurring in spontaneous speech
- Spontaneous speech across life span
- Development of spontaneous speech in L1
- Spontaneous speech in clinical population (aphasia, SLI, deafness, etc.)
- Analysis of paralinguistics in spontaneous speech
- Grammaticalization: Evidence from spontaneous speech
- Syntax of spontaneous speech
- Communication/speech accomodation in spontaneous speech
- Speech synthesis
- Dialogue act modelling
- Prosody modelling in spontaneous speech
- Speaker recognition forensic voice comparison
- Speech summarization in spontaneous speech
- Spontaneous speech recognition

The 'CAPSS Prize of the best young presenter'

The best talk on a piece of high quality research given by a presenter under 35 years of age is going to be rewarded in order to recognize and encourage their work.

Mária Gósy

Program Committee

Chairs:

Gósy, Mária (Department of Phonetics, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary)

Beke, András (Department of Phonetics, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary)

Gráczi, Tekla Etelka (Department of Phonetics, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary)

Program Committee:

Auszmann, Anita (Department of Phonetics, Institute for Linguistics, Hungarian Academy of Sciences)

Bartkova, Katarina (University of Lorraine, France)

Bóna, Judit (Department of Phonetics, Eötvös Loránd University, Hungary)

Deme, Andrea (Department of Phonetics, Eötvös Loránd University, Hungary)

Esposito, Anna (Seconda Università di Napoli, Department of Psychology & International Institute for Advanced Scientific Studies, Spain)

Fuchs, Susanne (Zentrum für Allgemeine Sprachwissenschaft, Germany)

Gocsál, Ákos (Faculty of Music and Visual Arts, University of Pécs, Hungary)

Gyarmathy, Dorottya (Department of Phonetics, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary)

Hazan, Valerie (Department of Speech, Hearing & Phonetic Sciences, UCL, UK)

Krepsz, Valéria (Department of Phonetics, Research Institute for Linguistics, Hungarian Academy of Science, Hungary)

Liker, Marko (University of Zagreb, Croatia)

Markó, Alexandra (Department of Phonetics, Eötvös Loránd University, Hungary)

Moosmüller, Sylvia (Acoustic Research Institute, Austrian Academy of Sciences, Austria)

Mycock, Louise J. (University of Oxford, UK)

Neuberger, Tilda (Department of Phonetics, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary)

Olaszy, Gábor (Department of Telecommunications and Media Informatics, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Hungary)

Recasens, Daniel (Universitat Autònoma de Barcelona, Spain)

Rosenhouse, Judith (Swantech Ltd., Israel)

Shriberg, Elisabeth (SRI International, USA)

Silber–Varod, Vered (The Open University of Israel, Israel)

Váradi, Tamás (Research Group for Language Technology, Department of Language Technology and Applied Linguistics, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary)

Vicsi, Klára (Department of Telecommunications and Media Informatics, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Hungary)

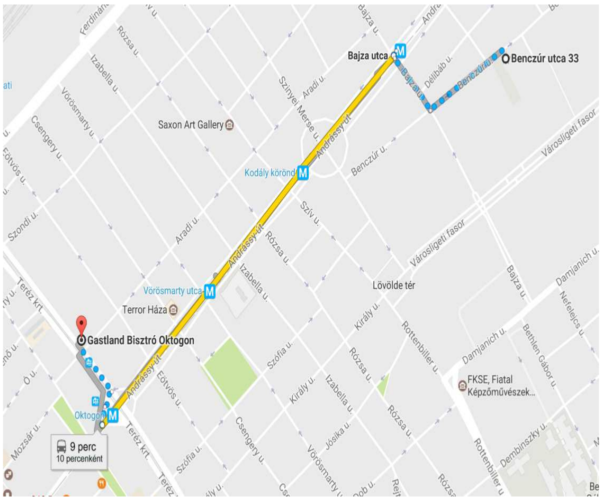
Technical details

Social Events

Reception

The reception will be at Gastland Bisztró Oktogon from 7 pm on Monday.

Address: H-1067, Budapest, Teréz krt. 23



Farewell Cakes and Drink Party

The Farewell Cakes and Drink Party will be at the Institute for Linguistics, Hungarian Academyof Sciences (conference venue) shortly after the last presentation (demo) on Thursday.

Publications

We plan to publish selected papers of the workshop in either the 2018 issue – as a special issue – of the journal The Phonetician or in an online book. The book is not planned to be a proceedings, rather an edited book on the specific topic of the workshop.

All papers will undergo double-blind peer-review by two experts of the topic.

The submission lasts from June 1st, 2017 to October 15th, 2017.

The exact details of the expected format will be given upon the authors' choice which possibility (Journal paper or book chapter) they decide for. Therefore, we will kindly ask you to tell us your preference during the workshop. You will be able to give your answer at the reception desk.

Program

Monday
May 15th, 2017

09,30	Opening	
10,00–10,45	**Plenary:** Valerie Hazan	*Spontaneous speech adaptations in challenging communicative conditions across the lifespan*
10,45–11,00	Questions	
11,00–11,30	Coffee break	
11,30–11,50	Vered Silber-Varod and Noam Amir	*"When two giants meet": The interaction between lexical stress and utterance-final prosody in spoken Hebrew*
11,50–12,10	Marko Liker	*Electropalatographic analysis of vowels in quasi-spontaneous speech*
12,10–12,30	Outi Tuomainen and Valerie Hazan	*Disfluencies in spontaneous speech in younger and older adults in easy and difficult communicative situations*
12,30–13,00	Questions	
13,00–14,00	LUNCH	
14,00–14,20	Alexandra Markó, Andrea Deme, Márton Bartók, Gergely Varjasi, Tekla Etelka Gráczi and Tamás Gábor Csapó	*Word-initial glottalization in the function of speech rate and vowel quality*
14,20–14,40	Sylvia Moosmüller, Hannah Leykum and Julia Brandstätter	*Is there a tendency to merge /e/ and /ɛ/ in Standard Austrian German? Data from read and spontaneous speech*
14,40–15,00	Krisztina Zajdó	*Building speech sounds through scaffolding: The case of motherese*
15,00–15,30	Questions	
15,30–16,00	Coffee break	
16,00–16,20	György Szaszák and Anna Moró	*Automatic punctuation recovery in read and spontaneous Hungarian using a recurrent neural network based sequential model for phonological phrases*
16,20–16,40	Judit Bóna	*Non-verbal vocalizations in spontaneous speech: The effect of age*
16,40–17,00	Dorottya Gyarmathy, Tilda Neuberger and Anita Auszmann	*The relationship between silent pause and breath-taking in spontaneous speech*
17,00–17,30	Questions	
19,00	**Welcome Party**	

Tuesday
May 16th, 2017

09,00–09,45	**Plenary:** Nick Campbell	*Towards interactive speech synthesis; an example of robot-human dialogues in a spontaneous environment*
09,45–10,00	Questions	
10,00–10,20	Katalin Mády, Uwe D. Reichel, Beáta Gyuris and Hans–Martin Gärtner	*The impact of syntax and pragmatics on the prosody of dialogue acts*
10,20–10,40	Tekla Etelka Gráczi, Alexandra Markó and Karolina Takács	*Word-initial glottalization in the function of articulation rate and word class*
10,40–11,00	Questions	
11,00–11,30	Coffee break	
11,30–11,50	Reinhold Greisbach, Anne Hermes and Alexandra Bückins	*Larynx movement in the production of Georgian ejective sounds*
11,50–12,10	Mária Gósy and Valéria Krepsz	*Phrase-final lengthening of phonemically short and long vowels in Hungarian spontaneous speech across ages*
12,10–12,30	Máté Ákos Tündik, Annamária Kovács, Miklós Gábriel Tulics, Anna Moró and Attila Gróf	*MagmaNet: Ensemble of 1D convolutional deep neural networks for speaker recognition in Hungarian*
12,30–13,00	Questions	
13,00–14,00	LUNCH	
14,00–14,45	**Plenary:** Ruth Huntley Bahr	*Variability in speech sound production: Covert contrasts in the speech of children with cochlear implants*
14,45–15,00	Questions	
15,00–15,20	Sarah Brandstetter	*Can you speak less dialect, please?*
15,20–15,40	Judit Bóna and Tímea Vakula	*Phonetic characteristics of disfluent word-repetitions: The effect of age and speech task*
15,40–16,00	Questions	
16,00–16,30	Coffee break	
16,30–16,50	György Szaszák and András Beke	*Exploiting prosodic and word embedding based features for automatic summarization of highly spontaneous Hungarian speech*
16,50–17,10	Davor Trošelj	*Vowel-formant frequencies of Hungarian–Croatian bilinguals and Hungarian monolinguals in spontaneous speech*
17,10–17,30	Questions	

Wednesday
May 17th, 2017

09,00–09,45	**Plenary:** Vesna Mildner	*Neurolinguistic aspects of speech processing*
09,45–10,00	Questions	
10,00–10,20	Tilda Neuberger and András Beke	*Effects of gemination on the duration and formant frequencies of adjacent vowels in Hungarian voiceless stops*
10,20–10,40	Ákos Gocsál	*Speaker age estimation by musicians and non-musicians*
10,40–11,00	Questions	
11,00–11,30	Coffee break	
11,30–11,50	Gordana Varosanec–Skaric, Zdravka Biocina and Gabrijela Kisicek	*Comparison of F0 measures for male speakers of Croatian, Serbian and Slovenian*
11,50–12,10	László Hunyadi	*On some linguistic properties of spoken Hungarian based on the HuComTech corpus*
12,10–12,30	Valéria Krepsz and Mária Gósy	*Stem and suffix durations in words of increasing length in children's spontaneous utterances*
12,30–13,00	Questions	
13,00–14,00	LUNCH	
14,00–14,20	István Szekrényes	*Challenges in automatic annotation and the perception of prosody in spontaneous speech*
14,20–14,40	Anita Auszmann	*A perceptual comparison: spontaneous speech of speakers' today and 40 years ago*
14,40–15,00	Mária Laczkó	*The temporal characteristics of teenagers in the various spontaneous speech genres*
15,00–15,30	Questions	
15,30–16,00	Coffee break	
16,00–16,30	Demonstration: Tamás Gábor Csapó, Andrea Deme, Tekla Etelka Gráczi, Alexandra Markó and Gergely Varjasi	*Synchronized speech, tongue ultrasound and lip movement video recordings with the "Micro" system*

17,00 **Farewell Cake and Drinks Party**

Abstracts

Monday

May 15th, 2017

Spontaneous speech adaptations in challenging communicative conditions across the lifespan

Plenary speech by

Valerie Hazan and Outi Tuomainen
Dept of Speech, Hearing and Phonetic Sciences, UCL, UK

Most of our knowledge about the acoustic-phonetic characteristics of speech come from speech production studies that have analysed controlled materials such as read sentences produced in isolation in a quiet environment. In typical communicative situations, the speech that we produce is likely to differ from such norms: it will be spontaneous, produced with true communicative intent, in less than ideal acoustic environments and quite often in a multi-tasking situation. In such situations, speech can be highly dynamic as we ongoingly adapt the level of clarity of our speech according to the demands of the communicative conditions, as suggested by Lindblom in his Hyper-Hypo model of speech production (Lindblom, 1990).

In our work, our aim has been to analyse the type of acoustic-phonetic adaptations made by speakers to counter the effects of adverse environments such as occur when communicating in noise, in the presence of other voices or with a hearing loss. We have recorded speech in laboratory conditions but have modelled natural communication by using a problem-based task that is carried out between two speakers. This picture-based 'spot the difference' task, called diapix (van Engen et al., 2010) involves the pair of speakers having to find 12 differences between their two pictures without seeing their conversational partner's picture. The degree of ease or difficulty with which speakers can communicate can be controlled by adding a communication barrier (e.g. babble noise, simulated hearing loss) affecting one or both of the speakers while they carry out the task together. This leads the person leading the interaction to naturally make adaptations to their spontaneous speech, producing a 'clear speaking style' in order to maintain effective communication, just as would happen in natural interactions. As speakers are recorded individually in connected sound-treated booths and communicate via headphones, a 'clean' and high-quality speech signal is recorded for each speaker.

In three consecutive large-scale studies, we have investigated such speech adaptations in 40 young adults (Hazan and Baker, 2011), 96 children aged 9 to 15 years (e.g., Hazan et al., 2016) and now 57 older adults aged 65 to 85 years, with 26 further younger adult controls (Tuomainen et al., 2016). A linked study has examined adaptations in hearing-impaired children while communicating with their hearing and hearing-impaired peers (Granlund et al., 2015). Each of these projects has led to the creation of large speech corpora (LUCID, kidLUCID and the forthcoming elderLUCID) containing many hours of spontaneous speech interactions. The lengthy processing of these corpora involves manual or automatic orthographic transcription, automatic alignment, manual checking of these alignments and the use of Praat scripts to obtain acoustic-phonetic measures. These measures include suprasegmental measures of articulation rate, fundamental frequency mean and range, relative intensity (representing spectral tilt) and segmental measures of vowel space.

In this talk, I will review our findings across these three studies spanning a broad age range; I will also discuss the challenges involved in the analysis of large spontaneous speech corpora. Our study with young adults showed that the adaptations that individual speakers made were, to an extent, dependent on the type of interference that was affecting their interlocutor (babble noise or vocoded speech), even though the speakers that we were analysing were not directly hearing the interference. This suggests that speakers used the direct or indirect feedback from their interlocutors during the interaction to attune their adaptations. Our study with children showed that they too made adaptations under similar conditions, although they had a tendency to use a strategy of increasing vocal effort (as shown by strong correlations between increases in fundamental frequency and decreases in spectral tilt) rather than using more varied strategies favoured by adults. Our ongoing study with older adults is showing a similar trend: older adults with age-related hearing loss tended to increase vocal effort to counter the effects of adverse conditions (again as shown by correlations between spectral tilt and fundamental frequency changes) while older adults with normal hearing thresholds and younger adults did not show this tendency.

In conclusion, investigating spontaneous speech in interaction in challenging communicative conditions can lead to a better understanding of the strategies used by speakers to maintain effective communication and of the impact of age and talker sex on such strategies. Despite the many challenges involved in the recording and analysis of spontaneous speech, such approaches will hopefully lead to a step forward in our knowledge of processes involved in speech communication.

References

Granlund, S., Hazan, V. L., & Mahon, M. (2015). Do children enhance phonetic contrasts in speech directed to a hearing-impaired peer? *Proceedings of the 18th International Congress of Phonetic Sciences*.

Hazan, V. L. & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *Journal of the Acoustical Society of America* 130(4), 2139–2152

Hazan, V., Tuomainen, O. & Pettinato, M. (2016). Suprasegmental characteristics of spontaneous speech produced in good and challenging communicative conditions by talkers aged 9 to 14 years old. *Journal of Speech, Hearing and Language Research* 59, S1596–S1607.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W.J. & Marchal, A. (eds) *Speech Production and Speech Modelling*, The Netherlands: Kluwer Academic, 403–439.

Tuomainen, O. & Hazan, V. (2016). Articulation rate in adverse listening conditions in younger and older adults. *Proceedings of Interspeech 2016*, 8–12 September 2016, San Francisco, USA.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of Native and Foreign-Accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech* 53, 510–540.

“When two giants meet”: The interaction between lexical stress and utterance-final prosody in spoken Hebrew

Vered Silber-Varod[1] and Noam Amir[2]

[1]The Research Center for Innovation in Learning Technologies, The Open University of Israel

[2]Department of Communication Disorders, Faculty of Medicine, Tel Aviv University

The present research investigates the interaction between the vowel duration at word level prosody and the phenomenon of phrase-final lengthening, as realized in spoken Israeli Hebrew. Experiments on the relations between the intonational phrase and lexical stress are numerous, and the phonological interaction of lexical and intonational features appears to be of typological interest (Ladd, 2001). Unlike experiments on how words in isolation are uttered, such studies strive to learn how phonological patterns are realized within natural environment, i.e., in an utterance or a sentence. For example, it is a common knowledge that syllables at breath group final position are longer than at its beginning (also known as *lengthening* and *anacrusis*, respectively (Cruttenden,1997)). Umeda (1975) studied such relations in large scale corpus of read speech. She found that the longer vowels were located at *pre-pausal* syllables, meaning, at stressed syllables by the end of utterance, paragraph, or sentence, to the rate of 1:2 or 1:3. The durational effects of phrase-final lengthening on the domain over which the speakers adjust it were studied in many languages, including Hebrew (Berkovits 1994). Turk and Shattuck-Hufnagel (2007) suggest two domains of the phrase-final lengthening in American English: the main stress syllable and the final syllable of the phrase-final word. They found that although most of the duration increase occurs in the phrase-final syllable rime, statistically significant lengthening of 7–18% also occurs in the main-stress syllable rime, when the main stress syllable *is not* the final syllable. Berkovits (1994) studied Hebrew stress manifestations of seven speakers, and found that utterance-final lengthening principally affected the final syllable regardless of stress. She also found evidence of progressive lengthening along the phrase-final syllable, which supports the suggestion that this phenomenon reflects the motor activity at utterance-final position, as the speech organs are slowing down. Silber-Varod and Kessous' (2008) study on weather broadcasts corpus, showed that stressed syllables in ultimate stress pattern located at intonation unit boundaries are longer than the preceding unstressed syllables, but stressed syllables in penultimate stress pattern are not always longer than the following unstressed utterance final syllables. The current database consists of 68 disyllabic target words in Hebrew, which differ phonemically only in their lexical stress pattern – final or penultimate. Target words were naturally embedded at the end of 68 carrier sentences (compared to six pairs of sentences and 12 disyllabic proper names in Berkovitz (1994)). Thirty subjects (13 Men and 17 Women) received the sentences in written form, and were instructed to read them aloud, pausing for at least two seconds between sentences. The research questions in the current study are: 1. How reliable is the contrast of the durational parameter between stressed and unstressed vowels of the same lexical word (not necessarily with identical vowel)?; 2. How reliable is the contrast of the durational parameter between stressed and unstressed identical vowels in the minimal pair words?

Results show that duration is an intrinsic indicator of stress, meaning the comparisons between stressed and unstressed vowels (p1* vs. p2, and u1 vs. u2* in Figure 1) of the same word showed significance differences (t(29)=9.446, −19.522 respectively, p<0.001 for both comparisons). As to the contrast between stressed and unstressed identical vowels (p1* vs. u1, and p2 vs. u2* in Figure 1), we found that duration is an extrinsic indicator of stress, i.e. the comparisons between stressed and unstressed identical vowels also showed significance differences (t(29)=26.253, 15.718 respectively, p<0.001 for both comparisons). Utterance-final lengthening affected the duration of the words and vowels. Words were ~26% longer at utterance-final position compared to nonfinal position. As to vowels, across words comparison showed that the effect on the final vowel was the largest, regardless of stress: Second vowels were lengthened to a larger extent (23% in penultimate stress and 18% in final stress) compared to first vowels (7% in penultimate stress and 4% in final stress). Within words comparisons showed a different effect on the two stress patterns: in penultimate words, the gap between stressed and unstressed vowels was reduced in utterance final position (35% in nonfinal; 17% in final position); in final stress words, the gap between stressed and unstressed vowels was *increased* in utterance final position (from 42% in nonfinal to 61% in final position). These findings suggest that although utterance final lengthening does not affect the relative dominancy in length of the stressed vowels, it lengthens the last vowels more than the first vowels, regardless of stress assignment.

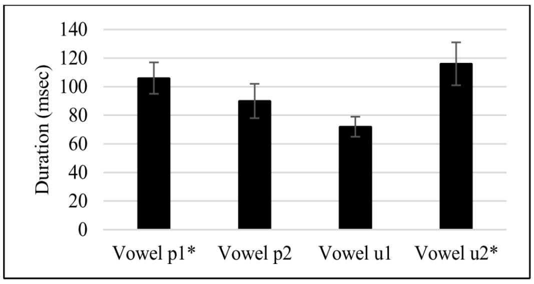


Figure 1. Mean vowel durations in utterance-final position, and 95% confidence intervals of four types of lexical stress conditions (p1*, p2, u1, and u2*). Asterisk [*] symbolizes the stress vowel.

Table 1. Means of durations (msec) for penultimate-stressed and final-stressed words in utterance-final position (current study, column B) compared to non-final position (Silber-Varod, Sagi, & Amir 2016, column A) and the study of Berokovits (1994) (columns D-F)

	A	B	C		D	E	F
	Non-final position (msec) (Silber-Varod, Sagi, & Amir 2016)	Utterance-final position (msec)	Difference		Non-final position (msec) (Berkovits, 1994)	Utterancefinal position (msec) (Berkovits, 1994)	Difference (Berkovits, 1994)
Penultimate stress							
Target word	337	424	26%		312	459	47%
Vowel p1*	99	106	7%		-	-	-
Vowel p2	73	90	23%		49**	77**	57%
Final stress							
Target word	336	427	27%		300	422	41%
Vowel u1	69	72	4%		-	-	-
Vowel u2*	98	116	18%		92**	127**	38%

* Stressed, ** In Berkovits (1994), target words had only [i] as the final vowel.

References

Berkovits, R. 1994. Durational effects in final lengthening, gapping and contrastive stress, *Language and Speech*, 37, 237–250.

Cruttenden, A. (1997). *Intonation.* Second edition. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Ladd, D. R. (2001). Intonational universals and intonational typology. In: M. Haspelmath (Ed.), *Language Typology and Language Universals: An International Handbook.* Vol. 2. De Gruyter, Berlin.

Silber-Varod, V., & Kessous, L. 2008. Prosodic boundary patterns in Hebrew: A case study of continuous intonation units in weather forecast. In P. A. Barbosa, S. Madureira,and C. Reis, (Eds.), *Proceedings of the Speech Prosody 2008 Conference*, Campinas, Brazil: Editora RG/CNPq. 265–268.

Silber-Varod, V., Sagi, H., & Amir, N. 2016. The acoustic correlates of lexical stress in Israeli Hebrew. *Journal of Phonetics*, 56, 1–14.

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. Journal of Phonetics, 35(4), 445–472.

Umeda, N. (1975). Vowel duration in American English. *The Journal of the Acoustical Society of America*, 58(2), 434–445.

Electropalatographic analysis of vowels in quasi-spontaneous speech

Marko Liker
University of Zagreb, Croatia

Instrumental data on speech production physiology are crucial for our understanding of speech motor control, coarticulation and connected speech processes (Stone, 2010). However, instrumental physiological techniques, such as electropalatography (EPG), are mostly limited to laboratories and laboratory speech is often described as different from spontaneous speech (Celata and Calamai, 2012). It is therefore quite challenging to use instrumental physiological techniques for spontaneous speech recording and analysis. Particularly demanding for instrumental physiological analysis in both laboratory and spontaneous speech are vowels (Howard & Haselwood, 2012). Articulatory analysis of vowels is considered demanding mainly because of difficulties with determining tongue shapes and positions during speech. No single instrumental physiological technique provides a complete insight into vowel articulation and EPG might seem particularly unsuitable (Howard & Haselwood, 2012). EPG is instrumental physiological technique for recording and analysis of tongue-to-palate contact patterns during speech (Gibbon & Nicolaidis, 1999). Since tongue-to-palate contact is relatively low or even non-existent during vowel productions (Hardcastle & Gibbon, 1997), EPG is not the first choice when it comes to analysing vowels. However, some research data suggest that high vowels and some diphthongs can be analysed via EPG quite successfully (Byrd, 1995; Recasens & Espinosa, 2005; Gibbon et al. 2010). The analysis of low vowels using EPG is still problematic as well as separate quantification of vertical (high-low) and horizontal (front-back) position of each vowel using EPG data only.

Motivated by the review above, there are two aims in this investigation. The first aim is to elicit and record quasi-spontaneous speech in the laboratory using EPG. The second aim is to explore the possibilities of using EPG to quantify three Croatian corner vowels on the vertical (high-low) and horizontal (front back) axis of the vowel chart, which is comparable to the vowel chart based on acoustic data (F1, F2), in quasi-spontaneous speech.

Data were extracted from the R-kor corpus of Croatian speech containing simultaneous acoustic and EPG data. Speech material from eight female speakers of Standard Croatian with no speech or hearing impairments was utilised. A dialogue situation was set up in the form of a map task. Each speaker was asked to describe the path through a maze and read signs at 15 check-points marked throughout the path. Each sign contained a two-syllable CVCV word with one of the corner vowels of the Standard Croatian (/i, a, u/) in stressed and in unstressed position (e.g. “rasa”, /’rasa/, English translation: race). During the recording session each speaker repeated each vowel five times in the stressed and unstressed position. Stressed vowels were analysed in this investigation. Six indices were calculated for each vowel: total contact, dental closure, alveolar closure, postalveolar closure, palatal closure and velar closure. Total contact was used to quantify the vertical (high-low) position of each vowel on the vowel chart. A newly developed measure used other five indices to quantify the horizontal (front-back) position of vowels. Articulate Assistant software (Wrench et al. 2002) was used for EPG analysis, while the statistical significance of differences was tested using two-way ANOVA with replication (alpha 0.05).

The clustering of the results for each vowel showed that on the basis of EPG indices it was possible to produce a vowel chart for each of the speakers. Two-way ANOVA with replication showed that differences between vowels were statistically significant both for vertical (F(7, 2) = 660.24, p<0.001) and horizontal (F(7, 2) = 122.78, p<0.001) axis of the vowel chart. The interaction analysis returned significant results (vertical: p<0.001, horizontal: p=0.01) showing that individual speakers’ productions differed substantially. A newly developed measure for the quantification of vowels along the horizontal (front-back) axis proved to be more discriminative than the traditional CoG-based measure. The analysis of the variability of each of the indices showed that vowel /a/ is most variable in the majority of the speakers, while /i/ is least variable.

This investigation demonstrates one possibility of eliciting quasi-spontaneous speech in the laboratory using EPG. The investigation shows that EPG is sensitive enough to quantify vowels along vertical and horizontal axes of the vowel chart. The results of vowel variability are discussed in terms of Degree of Articulatory Constraint theory (Recasens et al. 1997).

References

Byrd, D. (1995). Palatogram reading a phonetic skill: A short tutorial. *Journal of the International Phonetic Association*, 24, 21–34.

Celata, C., Calamai, S. (2012). Introduction. *Italian Journal of Linguistics*, 24(1), 43–64.

Gibbon, F. E., Lee, A., Yuen, I. (2010). Tongue palate contact during selected vowels in normal speech. *The Cleft Palate Craniofacial Journal*, 47(4), 405–412.

Gibbon, F. E., Nicolaidis, K. (1999). Palatography. In W. J. Hardcastle, N. Hewlet (eds.) *Coarticulation: theory, data and techniques*, Cambridge: CUP, 229–245.

Hardcastle, W. J., Gibbon, F. (1997). Electropalatography and its clinical applications. U: M. J. Ball, & C. Code (ur) *Instrumental Clinical Phonetics*. Whurr: London.

Howard, S., Haselwood, B. (2012). The contribution of phonetics to the study of vowels and vowel disorders. U: Ball, M. J., Damico, J. S., Gibbon, F. E. (ur) *Handbook of Vowels and Vowel Disorders*. Routledge: London, 61–112.

Recasens, D., Espinosa, A. (2005). Dispersion and variability of Catalan vowels. *Speech Communication*, 48(6), 645–666.

Recasens, D., Pallarès, M. D., Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *Journal of the Acoustical Society of America* 102 (1): 544–561.

Stone, M. (2010). Laboratory techniques for investigating speech articulation. In W. J. Hardcastle, J. Laver, F. E. Gibbon (eds.). *The Handbook of Phonetic Sciences*. Malden-Oxford-Chichester: WileyBlackwell. 9–38.

Wrench A. A., Gibbon, F. E., McNeill, A. M., Wood, S. E. (2002). *An EPG therapy protocol for remediation and assessment of articulation disorders*. In John H.L. Hansen; Brian L. Pellom (Eds.). *Proceedings of ICSLP-2002*: 965–968.

Disfluencies in spontaneous speech in younger and older adults in easy and difficult communicative situations

Outi Tuomainen and Valerie Hazan
Speech Hearing and Phonetic Sciences, University College London (UCL), UK

Spontaneous conversational speech is notoriously disfluent (Bortfeld et al., 2001): disfluencies (DFs) such as word and phrase repetitions and false starts may form even up to 5-10% of everyday conversations (Clark, 1994). Disfluencies in spontaneous speech are often associated with disruptions in word-finding or formulating sentences, with distractions (Yairi & Seery, 2011) or with an increase in cognitive load (Bortfield et al., 2001). It has also been shown that talkers become more disfluent when they are speaking in background noise (Jou & Harris, 1992, Southwood & Dagenais, 2001). There are well-documented changes in speech perception and production with increasing age. Older talkers have more difficulty retrieving words than do younger talkers (Burke et al., 1991) generating more disfluencies in discourse compared to younger adults (YAs) (Bortfeld et al., 2001). Older adult (OA) talkers also often report having particular difficulty communicating in challenging listening conditions, e.g. in noise or in the presence of other talkers.

When communication becomes effortful, talkers need to make various adjustments to their speech production to aid listener's understanding. For example, they may use a 'clear speaking style', which involves them speaking more slowly and reducing the complexity of their utterances amongst other characteristics. Adopting a more careful speaking style is likely to result in decrease in disfluencies in conversation, but it is not known whether OA talkers become less disfluent when speaking clearly than casually. Moreover, it is not known whether OA talkers become more disfluent than YA talkers when they are communicating in challenging listening conditions. The aim of the current study was to investigate disfluency rates in younger and older adults when they are speaking casually, when they are speaking clearly for the benefit of their interlocutor, and when they are speaking in background noise.

83 older adults aged 65 to 85 years (30 female) and 26 younger adults aged 18 to 35 years (15 female) were recorded while they completed a problem-solving spot-the-difference picture task (diapixUK; Baker & Hazan, 2011) with a young adult interlocutor (aged 18-33 years). The main participants (OA, YA) were told to take the lead in the interaction ('Talker A' role) while the young adult interlocutor had a more passive role ('Talker B'). Talker pairs completed the tasks in three different listening conditions: when no interference was present (NORM), when Talker B had a simulated severe-to-profound hearing loss (HLS), and when both talkers heard 8-talker babble noise (BAB2). It was expected that the NORM condition would elicit a casual speaking style in Talker A while the HLS and BAB2 would elicit a clear speaking style, as this would be necessary to communicate effectively despite the communication barrier. DFs were classified from Talker As speech using a system adapted from Shriberg's Disfluency Types (Shriberg, 2001) that has previously been used to analyse spontaneous speech (see Table 1). DF types and their position in an utterance (not reported here) were identified manually in Praat (version 6.0.19) and their frequencies were extracted using an in-house Praat script. Because the length of the speech samples differed between different speakers, the disfluency rate was calculated as a percentage of disfluent items relative to the total number of words produced in each listening condition. We predicted that all talkers would produce more disfluencies in a casual speaking style than in a clear style. We also predicted that OA talkers would produce more disfluencies overall than YA talkers both when communication was easy (NORM) and difficult (HLS and BAB2), and that they would be particularly affected by the background babble. In addition to the effects of age, some studies have shown that men produce more disfluencies than women (Shriberg, 1994; Bortfield, 2001), and we predicted that we would find these gender differences across all speaking styles.

Preliminary statistical results based on an analysis of a subset of talkers (N=20) across all disfluency types show that, as predicted, OA talkers produced more disfluencies (7.8%) than YA talkers (6.2%), and male talkers (8.3%) more than female talkers (5.5%) in the NORM condition. Furthermore, female talkers (both YA and OA) produced less disfluencies when they were talking clearly for the benefit of their interlocutor (HLS condition, see Table 2). However, male talkers did not show the same disfluency reduction in difficult communicative conditions. Against our predictions, when talking in background noise (BAB2), older adult male talkers produced less disfluencies than when communicating in good listening conditions (NORM). The other talker groups did not significantly increase or decrease the disfluencies in background noise (see Table 2). However, descriptive data (and preliminary statistical analyses) show that older adult female talkers produced marginally more disfluencies in the BAB2 condition than in the NORM condition indicating a potential difficulty communicating in background noise. Together these preliminary results indicate that there are age- and gender-related differences in disfluency rates in casual spontaneous speech. Furthermore, there are potential age and gender differences when talkers are either adapting their speech for the benefit of their interlocutor or communicating in noise. However, these results are based on a subset of the data and should be treated as showing possible statistical trends. Analysis from a larger set of talkers will be presented at the workshop. Analyses of other factors (such as speaking rate and hearing status) that might affect disfluency rates in these groups, along with analyses of different disfluency types (including within-speaker silent pauses) will also be presented.

Table 1. Disfluency types adapted from Shriberg, 2001.

Group of disfluency	**Type of disfluency**	**Example**
Filled pauses	Erm, err Other	Show flights from Boston on (erm) from Denver on Monday Show flights from (like) Boston
Repetitions	Word repetitions Phrase repetitions Part-word repetitions Insertions Articulation errors Substitutions	Show the – the morning flights Show the – show the morning flights Show the morn – morning flights Show the flights – show the morning flights Show me the flee – flights Show the morning – show the evening flights
Incomplete phrases	False starts	Show me the – which flights go to Boston

Table 2. Percentage of disfluencies out of all words produced in younger and older adult talkers (SD in brackets).

	NORM	HLS	BAB2
YA female	5.1 (2.7)	3.5 (1.7)	3.9 (3.5)
YA male	7.3 (1.9)	7.3 (2.9)	8.2 (1.8)
OA female	6.0 (1.4)	3.6 (2.4)	8.1 (6.4)
OA male	9.2 (1.7)	9.4 (1.4)	7.8 (2.0)

References

Baker, R. & Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43, 761–770.

Bortfeld, H., Leon, SD., Bloom, J., Schober, MF. & Brennan, S. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, 44, 123–147.

Burke, DM., MacKay, DG., Worthley, JS., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, 30, 542–579.

Clark, H. (1994). Managing problems in speaking. *Speech Communication*, 15, 243–250.

Jou, J. & Harris, J. (1992). The effect of divided attention on speech production. *Bulletin of the Psychonomic Society*, 30, 310-304. Shriberg, E. (2001). To 'errrr'is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31, 153–169.

Southwood, MH. & Dagenais, P. (2001). The role of attention in apraxic errors. *Clinical Linguistics and Phonetics*, 15, Yairi, E. & Seery, CH. (2011). *Stuttering: Foundations and Clinical Applications*. Boston: Pearson, 113-116.

Word-initial glottalization in the function of speech rate and vowel quality

Alexandra Markó[1,4], Andrea Deme[1,4], Márton Bartók[1,4], Gergely Varjasi[1,4],
Tekla Etelka Gráczi[2,4] and Tamás Gábor Csapó[3,4]
1Dept. of Phonetics, Eötvös Lorand University, 2Dept. of Phonetics, Research Institute for Linguistics, HAS,
3Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics,
4MTA-ELTE "Momentum" Lingual Articulation Research Group

Irregular phonation (a.k.a. glottalization) serve prosodic functions in typologically unrelated languages, e.g., American English (Dilley et al.1996), Czech, Spanish (Bissiri et al. 2011), German, Polish (Kohler 1994; Malisz et al. 2013), Hungarian (Markó 2013) and others. The occurrence of irregularity may be influenced by several factors, and it shows high inter- and intraspeaker variability (e.g., Dilley et al. 1996; Redi & Shattuck-Hufnagel 2001; Markó 2013). Kohler (2001: 282) defined four types of glottalization, as follows: (1) Vowel-related glottalization phenomena which signal the boundaries of words or morphemes. (2) Plosive-related glottalization phenomena which occur as reinforcement or even replacement of plosives. (3) Syllable-related glottalization phenomena which characterize syllable types along a scale from a glottal stop to glottalization (e.g., Danish stød). (4) Utterance-related glottalization phenomena which comprise (i) phrase-final relaxation of phonation and (ii) truncation glottalization, i.e., utterance-internal tensing of phonation at utterance breaks. Studies performed on Hungarian irregular phonation documented type (1) and type (4) both in read and spontaneous speech (Bőhm & Ujváry 2008; Markó 2013).

It is a prevalent view in the literature that irregular phonation is dependent on speech rate (Umeda 1978; Rodgers 1999). Studies in German and Polish read and spontaneous speech revealed that the appearance of word-initial irregular phonation is more frequent if the speech rate is slower and the word-initial vowel is back and open (Malisz et al. 2013; Lancia & Grawunder 2014).

In Hungarian a systematic analysis of speech rate and vowel quality has not been carried out so far. The lack of systematicity in earlier studies was partly due to unbalanced speech material (since previous studies analyzed mainly spontaneous speech which is inherently unbalanced in several aspects). Moreover, glottalization in Hungarian speech was analyzed only acoustically. In the present study we investigated two factors that may elicit glottalization, (i) speech rate, and (ii) vowel quality, by the analysis of the acoustic signal.

Based on previous results for other languages we plan to address the following questions. Is irregular phonation more frequent in the case of word-initial vowels if (i) the speech rate is slow (as opposed to fast); (ii) the vowel is back (as opposed to front); (iii) the vowel is open (as opposed to close or close-mid) also in Hungarian? We hypothesize that the frequency of occurrences of irregular phonation is higher in slow speech and in the case of back and open vowels, in accordance with results for other languages.

The test material consisted of trisyllabic non-sense words (V*tina*) preceded by an introductory phrase (*A szó:* 'The word is:') where the V represents one of the nine Hungarian vowel qualities /i eː ɛ y ø u o ɒ aː/. The stimuli were presented on a computer screen. Each trial consisted of two display screens: first the introductory phrase was showed to the participant, then the target item was displayed. In order to elicit speech rate differences between the conditions, the timing of the display screens was manipulated. In the "slow speech" condition, each display screen appeared for 1500 ms resulting in 3000 ms for one trial in total (including the introductory phrase and the target item). In the "fast speech" condition the timer was set to 300 ms, resulting in 600 ms for one trial in total. (In order to support the subjects' accommodation to the accelerating tempo throughout the experiment, one inbetween tempo was also used in filler blocks.) The participants' task was to read aloud the target items, but not the introductory phrase. However, as the timing of the introductory phrase reflected also the timing of the following (target) item, it enabled the subjects to prepare for the production of the following target item. The trials were ordered into blocks: within each block all the words including all the nine different vowels occurred in a randomized order once, and these blocks were repeated 5 times consecutively for each ("slow speech" and "fast speech") condition. In the case of every participant, the conditions were recorded in the order of tempo starting with the "slow speech" condition and ending with the "fast speech" condition (while the timing of the stimuli was increased gradually). In the present study 10 participants' speech samples were analyzed, and based on previous results (revealing that female speakers tend to glottalize more frequently than males, see e.g., Markó 2013) only female speakers were included.

Preliminary analysis of one speaker's data showed that the experimental design is suitable for the elicitation of different speech rates, as in the "fast speech" condition the duration of the target words' was on average 15.5±2.2% shorter than in the "slow speech" condition. In general, word-initial glottalization was 21.9% less frequent in "fast speech", than in "slow speech". In one speaker's data systematic differences cannot be observed either in terms of tongue height (close vs. open) or in terms of the horizontal tongue position (back vs. front).

References

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 Interspeech* 2011. Florence. 165–168.

Bőhm, T. & Ujváry, I. 2008. Az irreguláris fonáció mint egyéni hangjellemző a magyar beszédben [Irregular phonation as individual voice characteristic in Hungarian speech]. *Beszédkutatás* 2008: 108–120.

Dilley, L., Shattuck-Hufnagel, S. & Ostendorf, M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24: 423–444.

Lancia, L. & Grawunder, S. 2014. Tongue-larynx interactions in the production of word initial laryngealization over different prosodic contexts: a repeated speech experiment. In: Fuchs, S., Grice, M., Hermes, A., Lancia, L. & Mücke, D. eds. *Proceedings of the 10th International Seminar on Speech Production (ISSP)*. Cologne. 245–248.

Kohler, K. 1994. Glottal stops and glottalization in German. Data and theory of connected speech processes. *Phonetica* 51: 38–51.

Kohler, K. 2001. Plosive-related glottalization phenomena in read and spontaneous speech. A stød in German? In: Grønnum, N. & Rischel, J. ed. *To Honour Eli Fischer-Jørgensen.* Kopenhagen: Reitzel. 174–211.

Malisz, Z., Żygis, M. & Pompino-Marschall, B. 2013. Rhythmic structure effects on glottalisation: A study of different speech styles in Polish and German. *Laboratory Phonology* 4(1): 119–158.

Markó, A. 2013. *Az irreguláris zönge funkciói a magyar beszédben* [The functions of irregular phonation in Hungarian speech]. Budapest: ELTE Eötvös Kiadó.

Redi, L. & Shattuck-Hufnagel, S. 2001. Variation in the realization of glottalization in normal speakers. *Journal of Phonetics* 29: 407–429.

Rodgers, J. 1999. Three influences on glottalization in read and spontaneous German speech. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)* 34: 177–284.

Umeda, N. 1978. Occurrence of glottal stops in fluent speech. *Journal of the Acoustical Society of America* 64: 81–94.

Is there a tendency to merge /e/ and /ɛ/ in Standard Austrian German? Data from read and spontaneous speech

Sylvia Moosmüller, Hannah Leykum and Julia Brandstätter
Acoustics Research Institute, Austrian Academy of Sciences, Austria

Introduction

The influence of dialectal processes on Standard Austrian German (SAG) is obvious. E.g., the distance between tense and lax vowels is small to non-existent (Wiesinger, 2009; Harrington et al., 2012; Brandstätter and Moosmüller, 2015; Brandstätter et al., 2015); a distinct characteristic of Standard Austrian German, which points to the intermediate position of SAG between Standard German German (distinction between high tense and lax vowels) and the Bavarian dialects (no high lax vowels).

A similar, yet more complicated influence is the merger of /e/ and /ɛ/ in the Middle Bavarian dialects. The merger has its origin in the development of Old High German (OHG) *ë*, which, around 1200, merged with /e/ resulting from umlaut in the Bavarian dialects. In the Middle Bavarian dialects, to which Vienna belongs, only long *ë* merged with Umlaut /e/. Thus, [rɛːŋ] *Regen* „rain" started to rhyme with [leːŋ] *legen* „to lay", whereas [lɛkŋ] *lecken* „to lick" retained its quality (Kranzmayer, 1956). In these dialects, however, further inconsistences in the development of OHG ë emerged, in that in some specific instances either [ɛ] was maintained (e.g. in trisyllabics, which should have developed to [e]), or developed to [e] (e.g. derivations, which should have retained [ɛ]). Kranzmayer (1956) dubbed these developments E-confusion. Rowley (1990), however, succeeded in formulating a rule for the development of OHG *ë*.

In the 20th century, OHG *ë* was subjected to further developments; Vienna was affected most by these developments. According to Kranzmayer (1953), all e-vowels merged to [ɛː] or [ɛ] around 1920, the so-called *Viennese e-confusion*. Around 1960, Seidelmann (1971) observed a reversed tendency. Contrary to Kranzmayer, he observed a high variation in the realisation of the e-vowels, with a tendency to prefer the tense variant [e]. However, [e] and [ɛ] could be equally used. This tendency is still present in today's Viennese dialect (Moosmüller and Scheutz 2013).

Brandstätter and Moosmüller (2017) proved a statistically significant difference between /e/ and /ɛ/ in the task of reading logatomes. However, as concerns F2 and F3, the difference between [e] and [ɛ] was larger in the older age-group. A qualitative analysis revealed that occasionally, /e/ was merged with [ɛ] or reversely, /ɛ/ was merged with /e/. For this reason, we were interested whether /e/ and /ɛ/ were merged to a higher degree in spontaneous speech. We also compared the results with results from a reading task.

Method

The recording comprises semi-structured interviews and several reading tasks of 17 male and female speakers of SAG from Vienna, split into two age-groups. Formants (F1, F2, F3) were extracted over time (LPC, 46 ms window-length, 95% overlap) and duration measurements were performed.

Results

Both in the reading task as well as in spontaneous speech, a significant difference occurred with respect to F1 and F2, both in spontaneous and in read speech. The difference was more pronounced in the older age group as compared to the younger speakers.

The qualitative analysis, however, proved a merger of extension, especially in spontaneous speech. A cluster analysis revealed that 17 % of all vowels were assigned to the opposite cluster, of these, 25 % of phonological /e/ were pronounced as [ɛ] and 8,6 % of phonological /ɛ/ were realized as [e].

References

Brandstätter, J., Moosmüller, S. (2015): Neutralisierung der hohen ungerundeten Vokale in der Wiener Standardsprache, in: Lenz, A., Glauninger, M. (eds.), *Standarddeutsch im 21. Jahrhundert* (Wiener Arbeiten zur Linguistik). Wien: Vandenhoeck & Ruprecht, 183–203.

Brandstätter, J., Kasess, C., Moosmüller, S. (2015): Quality and Quantity in high vowels in Standard Austrian German, in: Leeman, A., Kolly, M., Dellwo, V., Schmid, S. (eds.), *Trends in phonetics and phonology in German speaking Europe*. Frankfurt am Main: Lang, 79–92.

Brandstätter, J., Moosmüller, S. (2017): Die Distinktion von standardsprachlich /e/ und /ɛ/ im Lichte der mittelbairischen E-Verwirrung, in: Lenz, A., et al. (eds.), *Bayerisch-österreichische Varietäten zu Beginn des 21. Jahrhunderts – Dynamik, Struktur, Funktion* (Zeitschrift für Dialektologie und Linguistik). Stuttgart: Steiner, 167–180.

Harrington, Jonathan, Hoole, Philip, & Reubold, Ulrich. (2012). A physiological analysis of high front, tense-lax vowel pairs in Standard Austrian and Standard German. *Italian Journal of Linguistics*, 24, 158–183.

Kranzmayer, Eberhard. (1953). Lautwandlungen und Lautverschiebungen im gegenwärtigen Wienerischen. *Zeitschrift für Mundartforschung* 21, 197–239.

Kranzmayer, Eberhard. (1956). *Historische Lautgeographie des gesamtbairischen Dialektraumes*. Wien: Böhlau.

Moosmüller, Sylvia, Scheutz, Hannes. (2013). Chain shifts revisited: The case of Monophthongisation and E-confusion in the city dialects of Salzburg and Vienna. In: P. Auer, J.C. Reina & G. Kaufmann (eds.), *Language Variation – European Perspectives IV. Selected Papers from the Sixth International Conference on Language Variation in Europe (ICLaVE)*, Freiburg, 2011, 173–186.

Rowley, Anthony R. (1990). Das "Kollmersche Gesetz" - Die Entwirrung von ahd. und mhd. A und E in den Dialekten des Bayerischen Waldes. *Zeitschrift für Dialektologie und Linguistik* 57, 54–59.

Seidelmann, Erich. (1971). Lautwandel und Systemwandel in der Wiener Stadtmundart. Ein strukturgeschichtlicher Abriss. *Zeitschrift für Dialektologie und Linguistik* 38:145–166.

Wiesinger, Peter. (2009). Die Standardaussprache in Österreich. In E. M. Krech, E. Stock, U. Hirschfeld & L. C . Anders (eds.), *Deutsches Ausspracheworterbuch*. Berlin/New York: de Gruyter. 229–258.

Building speech sounds through scaffolding: The case of motherese

Krisztina Zajdó
Széchenyi István Egyetem, Hungary

Sundberg (1998) hypothesized that the acoustic/phonetic properties of caregiverese change with the child's increased speech performance. Rather than producing hyperarticulated vowels, mothers may model vowels that teach the child about different aspects of vowel production as the child's level of speech performance increases. This process may make it challenging for children to build unambiguous phoneme categories.

To test Sundberg's hypothesis, a study was carried out examining the production of the Hungarian corner vowels /i:/, /u:/ and /a:/ in eight boys at the ages of 2;0, 2;6, 3;0, 3;6 and 4;0 years (n (children) = 40) and their mothers (n (mothers) = 40) as mothers were modeling pV(:)1p1V(:)1 structured tokens to their children and as the children used these words in conversation. A band filtering analysis of the vowel spectra at 50 randomly selected measurement points in each corner vowel category identified the space within the vowel triangle occupied. Mapping labeled vowel measures onto the reference plane suggests that children organize the acoustic space differently at different ages. In particular, while the positioning of the vowel /a:/ remains relatively constant in the speech of children at 2;0, 2;6, 3;0, 3;6 and 4;0 years, the other two corner vowels are positioned in different areas of the vowel space as the children get older. The vowel /u:/ is positioned as an increasingly higher back vowel, being fronted significantly in 4;0 years old children. The vowel /i:/ starts out by being positioned into a relatively high but centralized area of the vowel space at 2;0 years, followed by an increasingly higher and more frontal area of the acoustic vowel space in older children.

The positioning of the mothers' corner vowels within the acoustic vowel space also changes as mothers talk to older (as opposed to younger) children with more accurate speech sound production skills. In general, mothers position their corner vowels differently, depending the age of children they talk to. When modeling vowels to 2;0 years olds, mothers produce relatively low /a:/ vowels, high-central /i:/ vowels and /u:/ vowels that are back vowels but are not positioned very high. When talking to 2;6 and 3;0 years old children, mothers front their /i:/ vowels, thereby occupying a more compact area of the acoustic vowel space, and position their /u:/ vowels in a higher position. When talking to the oldest children, mothers produce the most fronted vowels, which positioning is closest to adult-like vowel qualities.

Overall, mothers guide their children towards the production of more accurate and more adult-like vowels, by taking into account the articulatory limitations (e.g., an initial inability to produce rounding in high vowels) their children are challenged with. However, these acoustical changes in phonological vowel categories in the input may challenge children as they are trying to build unambiguous phoneme representations. Potential explanations will be explored about the way children may take into consideration the relationship between their own acoustic vowel qualities as opposed to those produced by their mothers during the years towards developing more adult-like phoneme representations.

References

Cheour, M., Ceponiene, R, Lehtokoski, A., Allik, J., Alho, K., and Näätänen, R. 1998. Development of languagespecific phoneme representations in the infant brain. *Nature Neuroscience*, 1(5): 351–353.

Englund, K., and Behne, D. 2006. Changes in infant directed speech in the first six months. *Infant and Child Development*, 15: 139–160.

Fourakis, M. 1991. Tempo stress and vowel reduction in American English. *Journal of the Acoustical Society of America*, 90(4 Pt 1): 1816–1827.

Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., and Kuhl, P.K. 2006. Infant speech perception activates Broca's area: A developmental magnetoencephalography study. *NeuroReport*, 17(10): 957–962.

Johnson, K., and Martin, J. 2001. Acoustic vowel reduction in Creek: Effects of distinctive length and position in the word. *Phonetica*, 58, 81–102,

Kuhl, P.K. 2000. A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97(22): 11850–11857.

Kuhl, P.K., Andruski, J.E., Chistovich, I.A., Chistovich, L.A., Kozhevnikova, E., and Ryskina, V.L., 1997. Crosslanguage analysis of phonetic units in language addressed to infants. *Science*, 277: 684–686.

Näätänen, R. 2001. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, 38(1): 1–21.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., and Iivonen, A., 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, 385: 432–434.

Padgett, J., and Tabain, M. 2005. Adaptive dispersion theory and phonological vowel reduction in Russian. *Phonetica*, 62: 14–54.

Sundberg, U.1998. *Mother tongue - Phonetic aspects of infant directed speech*. Doctoral dissertation, PERILUS, Stockholm.

Van der Stelt, J.M., Zajdó, K., and Wempe, T.G. 2005. Investigating the acoustic vowel space in two-year-old children: Results for Dutch and Hungarian. *Speech Communication*, 47: 143–159.

Winkler, I., Lehtokoski, A., Alku, P., Vainio, M., Czigler, I., and Csépe, V. 1999. Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations. *Cognitive Brain Research*, 7: 357–369.

Ylinen, S., Shestakova, A., Huotilainen, M., Alku, P., and Näätänen, R. 2006. Mismatch negativity (MMN) elicited by changes in phoneme length: A cross-linguistic study. Brain Research, 1072: 175–185.

Zajdó, K. 2002. *The acquisition of vowels in Hungarian-speaking children aged two to four years: A cross-sectional study*. Unpublished doctoral dissertation, University of Washington, Seattle, USA.

Zajdó, K., and Stoel-Gammon, C. 2003. The acquisition of vowels in Hungarian: Developmental data. In *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, vol. 3, pages 2229-2232. Barcelona: Universitat Autònoma de Barcelona.

Zajdó, K., Van der Stelt, J.M., Wempe, T.G., and Pols, L.C.W. 2005. Cross-linguistic comparison of two-year-old children's acoustic vowel spaces: Contrasting Hungarian with Dutch. In *Proceedings of "InterSpeech 2005"*, Lisbon, Portugal, September 4-8, Bonn, Germany: ISCA, 1173–1176.

Automatic punctuation recovery in read and spontaneous Hungarian using a recurrent neural network based sequential model for phonological phrases

György Szaszák and Anna Moró
Budapest University of Technology and Economics

Despite the advances in automatic speech recognition technology, the automatic placement of punctuation marks is still an open issue; in most dictation systems, users have to explicitly dictate commas, sentence terminal punctuation marks, etc. In some applications – such as office or medical dictation (Vicsi et al., 2006), – this is an unnatural, but possible way of adding punctuations to recognized texts. On the other hand, in several tasks based on automatic speech transcription – such as close captioning / automatic subtitling (Varga et al., 2015), voice mining in audio archives, etc. – missing punctuation marks make reading and interpretation difficult as they require an increased mental effort.

Text analysis tools (POS taggers, dependency parsers, etc.) also highly rely on punctuation marks, which may be missing from texts obtained via speech recognition.

Basically two approaches exist in punctuation recovery: (i) prosody based (Christensen et al., 2001) and (ii) language modelling (Batista et al., 2008; Gravano et al., 2009) or recently, sequence modelling (Tilk & Alumäe, 2015) based approaches. The two can be combined into hybrids, as well. Prosody based approaches work fast, whereas language modelling based approaches are usually more resource demanding, which makes their application difficult in online (real-time) automatic speech recognition tasks, which already require much computation for the speech recognition. Moreover, language modelling based approaches show higher language dependency and may be used less easily for spontaneous speech, as they suffer from problems caused by ungrammatical words or non-verbal feedback expressions (Markó–Gósy–Neuberger, 2014) so characteristic in spontaneous speech.

Our interest is to explore punctuation recovery for Hungarian based on speech prosody, keeping in mind the above mentioned considerations, i. e. we would like to propose a fast and efficient approach in terms of both computational requirements and adaptability for spontaneous speech. Prosody based approaches typically focus on some prosodic markers related to punctuations, i.e. look for acoustic markers in duration, fundamental frequency or energy (c.f. Christensen et al., 2001). Unlike most of the studies in the field, which use direct acoustic markers, we intend to incorporate an abstract level of prosodic modelling beside using the acoustic features which relate directly to speech prosody. We do this lead by the conviction that prosody should be considered not only at the layer of different acoustic markers, which relate to different events (i.e a stress or a word boundary), but also as a coherent structure imposed onto the utterance. In Vicsi-Szaszák (2010), a framework was proposed to recover intonational and even phonological phrases directly from speech. This approach relies on modelling phonological phrases with a Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) hybrid, and uses a Viterbi alignment to match the most likely phonological phrase sequence against a speech utterance, using not more than 7 different models (including silence) and fundamental frequency and energy as input acoustic features. This approach reaches ~90% precision and recall in Hungarian phonological phrase detection task with a time accuracy within a length of a syllable (Szaszák et al., 2016).

In our contribution, we prove the hypothesis that the phonological phrase sequence shows characteristic patterns for different punctuation marks, especially regarding within sentence (mostly commas, but also semicolons or dashes) and sentence terminal (period, question or exclamation mark) punctuation marks. The next step is an attempt to model these sequential characteristics with Recurrent Neural Networks (RNN) using Long-Short Term Memory (LSTM) cells, and predict the probability of punctuation marks in a sequence labelling approach. By the implementation of the RNN, we keep a simple structure in order to allow for fast operation. For read speech, this approach is expected to yield the punctuation marks (or probability scores for these punctuation marks) required within a sentence, whereas in spontaneous speech, the approach is a candidate for automatically detecting virtual sentences (Gósy, 2008), and also to identify modality (declarative or interrogative).

We intend to evaluate the RNN in terms of punctuation recovery (precision and recall) both in read speech tasks and in spontaneous speech task. In the latter, a subjective evaluation test is required to analyse the proposed „segmentation" for virtual sentences.

References

Batista, F., Caseiro, D., Mamede, N., & Trancoso, I. (2008). Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news. *Speech Communication*, 50(10), 847–862.

Christensen, Heidi, Yoshihiko Gotoh, & Steve Renals (2001). "*Punctuation annotation using statistical prosody models*." ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding.

Gósy, M. & Kovács, M. 2008. Virtual Sentences of Spontaneous Speech: Boundary Effects of Syntactic-Semantic-Prosodic Properties. In *Human Factors and Voice Interactive Systems*. Springer US, 361–379.

Gravano, A., Jansche, M., & Bacchiani, M. (2009). Restoring punctuation and capitalization in transcribed speech. In *Acoustics, Speech and Signal Processing, 2009*. ICASSP 2009. IEEE International Conference on, 4741–4744.

Markó, A., Gósy, M. & Neuberger, T. (2014). "Prosody patterns of feedback expressions in Hungarian spontaneous speech." *Speech Prosody 2014*, Dublin.

Szaszák, G., Tündik, M.Á., Gerazov, B., Gjoreski, A. (2016). Combining atom decomposition of the F0 track and HMM-based phonological phrase modelling for robust stress detection in speech. Lecture Notes In Computer Science 9811: pp. 165–173.

Tilk, O., & Alumäe, T. (2015). LSTM for punctuation restoration in speech transcripts. In *Interspeech*, 683–687.

Varga, A., Tarján, B., Tobler, Z., Szaszák, G., Fegyó, T, Bordás, C., Mihajlik, P. (2015) Automatic Close Captioning for Live Hungarian Television Broadcast Speech: A Fast and Resource-Efficient Approach. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE 9319. 105–112.

Vicsi, K., Szaszák, G. (2010) Using prosody to improve automatic speech recognition. *Speech Communication* 52:(5) pp. 413–426.

Vicsi, K., Velkei, Sz., Szaszák, G., Borostyán, G., Gordos, G. (2006) Folyamatos, középszótáras beszédfelismerő rendszer fejlesztési tapasztalatai: kórházi leletező beszédfelismerő. *Hiradástechnika* 61:(3) 14–21.

Non-verbal vocalizations in spontaneous speech: The effect of age

Judit Bóna
Dept. of Phonetics, Eötvös Loránd University, Hungary

It is well known that spontaneous speech contains quite a few non-verbal vocal elements in addition to the verbal content (Trouvain 2014). We can express our mood, our emotions, our opinion with them but we can also reflect our speech partner's message. There are also non-verbal vocal elements without meaning (unintentional body sounds, physiological re-flexes). The vocal elements which refer to our emotional state or physical condition and those body sounds which occur as a natural attribute to articulation are unintentional ele-ments of speech, while other gestures, sounds, hummings are created intentionally (Vicsi et al. 2011; Neuberger 2012).

Research on non-verbal vocalizations is quite underrepresented in speech sciences, how-ever, it is quite important from the aspect of many practical applications in addition to their role in linguistics (e.g. speech technology, forensic phonetics, judgements of speakers' at-tributions; Li et al. 2008; Mohammadi et al. 2010; Prylipko et al. 2012; Neuberger & Beke 2013; Sárosi et al. 2014). The frequency and duration of non-verbal vocal elements depend on several different factors like individual characteristics of the speaker, their age, physical and emotional state, the relationship between the speakers or the speech task. This presenta-tion deals with the effect of age.

The most frequent types of non-verbal vocalizations are audible breathing and laughters (Trouvain & Truong 2012), and tongue clicks are relatively frequent, too (Bóna 2015). Pre-vious studies were carried out mostly with young and middle-aged adults. There are no data about the speech of other age groups. The aim of this presentation is to analyse how occur-rences of audible breathing and tongue clicks change depending on the speakers' age.

The main questions of this presentation are the following: 1) Is there any difference in the frequency of occurrences of audible breathing and tongue clicks in the speech of children, young adults and the elderly? 2) What is the duration of the realization of these elements in the three different age groups?

In the presentation these questions will be answered by the analysis of the speech of 60 speakers: 20 9-year-old, 20 young (20-30-year-old), and 20 elderly (70+) speakers. The speech samples were selected from two speech databases: GABI (Bóna et al. 2014) and BEA (Gósy 2012). 5-minute-long spontaneous narratives from each speaker (altogether 300 minutes of speech) were analysed. Non-verbal vocal elements were annotated by Praat soft-ware. The frequency, duration and place of occurrence were analysed.

Results show that the examined phenomena occurred most frequently in children's speech. There was significant difference in the duration of breathing between the age groups, too. The duration of tongue clicks did not differ significantly between the groups.

Our research carries an overall importance for practice. The further analysis of non-verbal vocalizations could contribute to a more precise profiling of speakers, to define the speak-ers' age based on the acoustic structure of speech.

Acknowledgement

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. K-120234.

References

Bóna, Judit – Imre, Angéla – Markó, Alexandra – Váradi, Viola – Gósy, Mária 2014. GABI – Gyermeknyelvi Beszédadatbázis és Információtár, *Beszédkutatás* 2014, 246–252.

Bóna, Judit 2015. Nonverbális hangjelenségek fiatalok és idősek spontán beszédében. *Beszédkutatás 2015.* 106–119.

Gósy, Mária 2012. BEA – A multifunctional Hungarian spoken language database, *Phonetician* 105–106.: 50–61. http://www.isphs.org/Phonetician/Phonetician_105_106.pdf

Li, Yanxiong – He, Qianhua – Li, Tao – Wang, Weining 2008. A detection method of lip-smack in spontaneous speech. In: *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on.* IEEE. 292–297.

Mohammadi, Gelareh – Vinciarelli, Alessandro – Mortillaro, Marcello 2010. The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In: *Proceedings of the 2nd international workshop on Social signal processing*. ACM, 17–20.

Neuberger, Tilda 2012. Nonverbális hangjelenségek a spontán beszédben. In Gósy, Mária (szerk.): *Beszéd, adatbázis, kutatások.* Akadémiai Kiadó, Budapest, 215–235.

Neuberger, Tilda – Beke, András 2013. Automatic Laughter Detection in Spontaneous Speech Using GMM–SVM Method. In *Text, Speech, and Dialogue*. Springer, Berlin–Heidelberg, 113–120.

Prylipko, Dmytro – Vlasenko, Bogdan – Stolcke, Andreas – Wendemuth, Andreas 2012. Language Modeling of Nonverbal Vocalizations in Spontaneous Speech. In *Text, Speech and Dialogue*. Springer, Berlin–Heidelberg, 488–495.

Sárosi, Gellért – Tarján, Balázs – Fegyó, Tibor – Mihajlik, Péter 2014. Automated transcription of conversational Call Center speech–with respect to non-verbal acoustic events. *Intelligent Decision Technologies* 8(4). 265–275.

Trouvain, Jürgen 2014. Laughing, Breathing, Clicking-The Prosody of Nonverbal Vocalisations. In *Proc. Speech Prosody*. 598–602.

Trouvain, Jürgen – Truong, Khiet P. 2012. Comparing non-verbal vocalisations in conversational speech corpora. In: *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals,* Istanbul, 36–39.

Vicsi Klára – Sztahó Dávid – Kiss Gábor 2011. Nem verbális hangjelenségek spontán társalgásban. *Beszédkutatás 2011.* 134–147.

The relationship between silent pause and breathtaking in spontaneous speech

Dorottya Gyarmathy, Tilda Neuberger and Anita Auszmann
Dept. of Phonetics, Research Institute for Linguistics, HAS

Our speech is occasionally interrupted by pauses. They are an essential part of human speech. Until now it is not clarified what can be actually considered as pause, how differently pauses can appear in speech, what are their acceptable minimum and maximum durations, and what functions they may have in speech.

Until the first half of the 20th century researchers examined pauses mainly from rhetorical aspects, they analysed the relationship between punctuation of written texts and their spoken realisations (Mátray 1861; Hevesi 1908; Simonyi 1903; Lindroth 1933). In 1877, Sweet was the first scientist who mentioned pauses as parts of the language system. He connected pauses of speech with breathing, and called the section of speech uttered with one act of breathing out as "breath group". Following Sweet's idea, various researchers discussed the problem of pauses of speech, such as Hungarian Balassa (1886), German Viëtor (1894), and Danish Jespersen (1904). Previous studies connected pauses of speech with either breath-taking, or punctuation (Weiske 1838; Bieling 1880).

From the second half of the 20th century empirical researches showed that silent pause is the most common phenomenon in the spontaneous speech, and carries many different functions in speech (Boomer 1965; Goldman-Eisler 1958; Hargreaves–Starkweather 1959; Levin et al. 1967; Tannenbaum et al. 1967; Verzeano–Finesinger 1949; Misono–Kiritani 1990; Gósy 2000, 2003; Menyhárt 2003; Markó 2005; Bóna 2007; 2013; Neuberger 2014). Speakers need pauses to breath, plan what they are going to say, or negotiate turn-taking. Esposito et al. (2007) conclude that pauses in speech are typically multi-determined phenomena; they have socio-psychological, communicative, linguistic and cognitive reasons. It is known from previous research that silent pauses caused by speech planning difficulties differ from those occurring at syntactic boundaries (Boomer 1965; Lounsbury 1965; Szende 1976). A recent study confirmed statistically significant difference between syntactical silent pauses and editing phase silent pauses (Gyarmathy 2017).

There are still many controversies about the relationship between silent pause and breath-taking. The aim of the present study is to investigate whether breath-taking is subordinated to thinking and is not performed by biological functions in spontaneous speech. Our study focuses on the analysis of the temporal structure of silent pauses and breaths in Hungarian spontaneous speech. We hypothesized that there would be proportional and/or durational differences among pause categories depending on breath-taking.

Our research is based on narratives of the BEA Hungarian spontaneous speech database (Gósy 2012). We used the recordings of 10 adult subjects (5 males and 5 females, their mean age was 27.4 years), and analysed all occurrences of their silent pauses. The spontaneous speech material we analysed was 71 minutes long. Each occurrence of silent pauses was annotated by Praat, version 5.4.21 (Boersma & Weenink 2013). We defined their duration manually. For statistical analysis, General Linear Mixed Model (GLMM) was used (SPSS 20.0).

Results provide detailed information about the relationship between silent pauses and breath-taking. In addition, results can be used in various areas of speech technology, or speech therapy.

References

Balassa J. 1886. *A phonetika elemei, különös tekintettel a magyar nyelvre*. [The Elements of Phonetics]. Magyar Tudományos Akadémia, Budapest.

Bieling, A. 1880. *Das Prinzip der deutschen Interpunktion nebst einer übersichtlichen Darstellung ihrer Geschichte*. Berlin.

Boersma, Paul – Weenink, David 2013. *Praat: doing phonetics by computer* (Version 5.4.21) [Computer program]. http://www.praat.org

Bóna J. 2007. *A felgyorsult beszéd produkciós és percepciós sajátosságai.* PhD disszertáció. Budapest

Bóna J. 2013. *A spontán beszéd sajátosságai az időskorban*. ELTE Eötvös Kiadó, Budapest.

Boomer, D. S. 1965. Hesitation and grammatical encoding. *Language and Speech* 8. 148–158.

Esposito, A. – Stejskal, V. – Smékal, Z. – Bourbakis, N. 2007. The significance of empty speech pauses: Cognitive and algorithmic issues. *Advances in Brain, Vision, and Artificial Intelligence*. Springer Berlin Heidelberg. 542–554.

Goldman-Eisler, F. 1958. Speech production and the predictability of words in context. Quarterly. *Journal of Experimental Psychology* 10. 96–106.

Gósy M. 2000. A beszédszünetek kettős funkciója. *Beszédkutatás 2000.* 1–14.

Gósy M. 2003. A spontán beszédben előforduló megakadásjelenségek gyakorisága és összefüggései. *Magyar Nyelvőr* 127/3. 257–277.

Gósy, M 2012. BEA – A multifunctional Hungarian spoken language database *Phonetician* 105/106: 50–61.

Gyarmathy, D 2017. A néma szünetek funkciói a spontán beszédben. [The function of silent pauses in spontaneous speech] *Beszédkutatás 2017.* (in press)

Hargreaves, W. A. – Starkweather, J. A. 1959. Collection of temporal data with the duration tabulator. *Journal of the Experimental Analysis of Behavior* 2. 179.

Hevesi S. 1908. *Az előadás művészete* [The Art of Delivery]. Stampfel, Budapest.

Jespersen, O. 1904. *Lehrbuch der Phonetik.* Leipzig – Berlin.

Levin, H. – Silverman, I. – Ford, B. 1967. Hesitations in children's speech during explanation and description. *Journal of Verbal Learning and Verbal Behavior* 6. 560–564.

Lindroth, H. 1933. *Sprachpsychologie und Interpunktion.* Archives Néerlandaises de Phonetique Experimentale VIII – IX.

Lounsbury, Floyd G. 1965. Transitional probability, linguistic structure and system of habit-family hierarchies. In Osgood, Charles E. – Sebeok, Thomas A. (eds.): *Psycholinguistics. A survey of theory and research problems*. Indiana University Press, Bloomington–London, 93–101.

Markó A. 2005. A temporális szerkezet jellegzetességei eltérő kommunikációs helyzetekben. *Beszédkutatás 2005*. 63–77.

Mátray, G. 1861. *A rendszeres szavalattan alaprajza*. [Outline of Systematic Elocution]. Trattner–Károlyi, Pest.

Menyhárt K. 2003. A spontán beszéd megakadásjelenségei az életkor függvényében. In Hunyadi László (szerk.): *Kísérleti fonetika – laboratóriumi fonológia a gyakorlatban*. Debreceni Egyetem Kossuth Egyetemi Kiadója. Debrecen. 125–138.

Misono, Y. – Kiritani, S. 1990. The distribution pattern of pauses in lecture-style speech. *Logopedics and Phoniatrics* 2. 110–113.

Neuberger T. 2014. *A spontán beszéd sajátosságai gyermekkorban*. ELTE Eötvös Kiadó, Budapest.

Simonyi Zs. 1903. *Iskolai helyesírás*. [School Orthography]. Budapest.

Szende T. 1976. *A beszédfolyamat alaptényezői*. Akadémiai Kiadó, Budapest.

Sweet, H. 1877. *A Handbook of Phonetics*. Clarendon Press, Oxford.

Tannenbaum, P. H. – Williams, F. – Wood, B. S. 1967. Hesitation phenomena and related encoding characteristics in speech and typewriting. *Language and Speech* 10. 203–215.

Verzeano, M. – Finesinger, J. E. 1949. An automatic analyzer for the study of speech in interaction and in free association. *Science* 110. 45.

Viëtor, W. 1894. *Elemente der Phonetik*. Leipzig.

Weiske, J. 1838. *Theorie der Interpunktion aus der Idee des Satzes*. Leipzig.

Tuesday

May 16th, 2017

Towards interactive speech synthesis; an example of robot-human dialogues in a spontaneous environment

Plenary speech by

Nick Campbell
Trinity College, Dublin

Speech synthesis is (by definition) *NOT* spontaneous. However, speech synthesis is increasingly being used in situations where spontaneous speech is common.

In the past, the main challenges for speech synthesis have been voice quality and prosody prediction, but we argue that nowadays the most important goal for synthesis technology research is for the system to know WHEN to speak, and to be able to parse the reaction of any listener(s) present. The second priority perhaps is to know WHAT to speak; in the sense that someimes it is necessary to repeat or paraphrase an utterance to facilitate smoother communication.

In order to explore the possibilities of Interactive Speech Synthesis, we are developoing a sentient dialogue system (Cara) which is able to monitor the cognitive states of its partner through sensing of vocal and physical dynamics throughout the conversation.

Work with the JST/ESP Expressive Speech Corpus has shown us that tone-of-voice is a key factor in displaying cognitive state changes and interpersonal dynamics. Multimodal signal processing as tested in the TableTalk and D64 data collections allows us to monitor similar reactive changes in body-posture and gestural dynamics.

Together with the HMI Research Group at UTwente, we have developed a sensitive Receptionist Robot that is able to manage short task-based conversations and to be aware of and cope with third parties that may intrude on the dialogue.

With DFKI and colleagues in the Metalogue Project we developed a computer dialogue system that sensed MetaCognitive processes in public speakers, and in the Joker Project with LIMSI and other European partners we are developing a joking conversational robot for elderly or socially deprived people.

Current work at the Speech Comunication Lab in Dublin includes extending the Herme robot for interactive short social dialogues and testing autonomous dialogue sensing mechanisms for timing and content control in potential entertainment or customer-care applications.

This invited talk will present the findings of our recent research into Interactive Speech Synthesis for Spontaneous Interactions and will describe our thinking behind future developments and uses of this exciting new technology.

The impact of syntax and pragmatics on the prosody of dialogue acts

Katalin Mády, Uwe D. Reichel, Beáta Gyuris and Hans–Martin Gärtner
Research Institute for Linguistics, HAS, Hungary

Introduction

Task-oriented spoken dialogues have several advantages: (1) Since speakers are involved in a non-linguistic task, they tend to concentrate less on the fact that they are being recorded, (2) various settings allow for the elicitation of repetitions of certain elements (words, names, etc.). (3) Since these tasks create a specific setting, intentions of speakers are more easy to control in terms of information structure, e.g. whether a certain element is given, new, contrastive etc.

In this paper the dialogue structure coding scheme by [1] was used in order to test whether dialogue acts can be classified based on their prosodic features developed by [4]. Dialogue acts (DA) were investigated under two aspects: (1) they belonged to different sentence types such as yes/no questions vs. wh-questions or to DA pairs, e.g. yes/no questions and positive or negative responses to them, alternatively, (2) they differed in their informational weight within the same sentence type, e.g. explaining new information vs. assuring that previous information was understood correctly in declaratives. The goal was to find out whether syntactic and pragmatic categories can be distinguished by different prosodic features.

Materials and methods

Data were taken from the Hungarian version of the object game of the Columbia Games Corpus [2] based on a computer-aided game with two participants and two laptops. Players see objects on their screens that are identical except for one object. The first player describes the position of the blinking object in relation to the other objects. The second player is supposed to place the object in exactly the same position. Participants get a score after each turn (altogether 14 in each game) on a 0 to 100 scale. (See [3] for more detail).

Annotation: The signal was manually segmented into inter-pausal chunks, text-transcribed annotated for dialog acts. F0 was extracted by Praat and preprocessed as described in [4]. Within each chunk prosodic phrases were extracted, and within each dialog act the most prominent syllable. Details on these unsupervised automatic annotation methods are given in [4].

Feature extraction: From this annotation we extracted temporal, energy and f0 features (cf. Table 1) on the entire dialog level (*glob*) and in an analysis window of length 0.3s around the most prominent syllable (*loc*). One part of the syllable-related features refers to its *Gestalt* properties, i.e. to what extent its f0 register is distinct from the underlying intonation phrase. For this purpose a base-, mid- and a topline were fitted both to the syllable as well as to the related prosodic phrase, and for each line pair the RMS was calculated within the syllable analysis window. The second local feature set describes the shape of the f0 contour in terms of the coefficient values of a third order polynomial. All features were extracted within the *CoPaSul* intonation stylization framework by a freely available toolkit [4].

Table 1: Temporal, f0, and energy features for each dialog act. *glob*-features were extracted on the dialog act level, *loc*-features in an analysis window centered on the nucleus of the most prominent syllable within the dialog act

Dialog act level temporal (glob_temp), f0 (glob_f0), and energy (glob_en)	
dur	duration
syl_rate	number of syllables per second
f0_max,med,sd	f0 maximum, median, standard deviation
en_max,med,rms,sd	energy maximum, median, RMS, standard deviation
Syllable level f0 *Gestalt* (loc_gst) and shape (loc_shape)	
loc_bl,ml,tl_rms	RMS between syllable and IP baseline, midline, topline
loc_c0–3	polynomial coefficients 0–3

Results

Mann-Whitney and Kruskal-Wallis tests were used due to the lack of normal distribution in all samples. Significance level was set to $p < 0.05$. 1 First, two pairs of DAs belonging to different sentence types were compared: (1) yes/no questions (QY) and the positive response (RY) given to them, (2) yes/no questions (QY) and *wh*-questions (QW). The above sentence types were distinguished mostly by local features: QY had higher *Gestalt* values than RY, presumably due to the obligatory low accent in yes/no questions, and QY and QW were differentiated by accent shape, supposedly being linked to a low accent in the first and a falling one in the second question type.

Two types of declaratives, the general category *EXPLAIN* (EX) containing new information and *CLARIFY* (CL) used for reassuring that the speaker has understood previous information properly, i.e. all-given information, were compared. EX was characterised by higher overall energy and longer duration than CL. Two other declaratives, COMMENT (CO) and READY (RE) were also compared to EX. These latter categories do not contain information relevant for the task itself, but either comments such as ’well, this is all I can tell you’ or a transition to the next turn ’o.k., so I press the button’. Thus again, their informational weight is lower than that of EX. Again, global features connected to duration, energy, and syllable rate show higher values for EX, while interestingly, *Gestalt* values showed a higher emphasis on the most prominent syllable for comments that often expressed emotions.

Based on [1]’s scheme, DAs were divided into three categories based on their position within a turn during the game. Initiations are mostly questions, responses consist of declaratives, while the category preparation signalises that a speaker is ready for a new turn. Initiations were realised with higher values for most global categories, i.e. duration, f0, energy and syllable rate. Preparation and response were best distinguished by accent shape.

Discussion

In this paper, a first attempt was made to test whether dialogue acts suggested by [1] can be characterised by stylised prosodic parameters. Comparisons were either based on syntactic or on pragmatic categories. While DAs that express grammatical categories such as various question types seem to be connected to different prosodic categories based on local features, DAs that belong to the same sentence type but carry different pragmatic meaning tend to be distinguished along global prosodic parameters. The findings of the present study will be extended by more pragmatic categories. A mid-term goal is to predict DAs simply on the basis of automatic prosodic feature extraction.

References

[1] J. Carletta, A. Isard, S. Isard, J.C. Kowtko, G. Doherty-Sneddon, and A.H. Anderson 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

[2] Agustín Gravano, Štefan Beňuš, Héktor Chávez, Julia Hirschberg, and Lauren Wilcox 2007. On the role of context and prosody in the interpretation of `okay'. In *Proc. 45th Annual Meeting of Association of Computacional Linguistics*, Prague, 800–807.

[3] Katalin Mády and Uwe D. Reichel. 2016. How to distinguish between self- and other-directed wh-questions? In Proc. *Phonetik und Phonologie im deutschsprachigen Raum*, Munich, Germany.

[4] U.D. Reichel 2017. *CoPaSul Manual – Contour-based parametric and superpositional intonation stylization.* RIL, MTA, Budapest, Hungary. https://arxiv.org/abs/1612.04765.

Acknowledgements

This work was funded by the Alexander von Humboldt Foundation and OTKA K 115922.

Word-initial glottalization in the function of articulation rate and word class

Tekla Etelka Gráczi[1,3], Alexandra Markó[2,3] and Karolina Takács[2]
[1]Dept. of Phonetics, Research Institute for Linguistics, HAS, [2]Dept. of Phonetics, Eötvös Lorand University,
[3]MTA-ELTE "Momentum" Lingual Articulation Research Group

Irregular phonation (laryngealization, glottalization) is a phonation type characterized by the irregular vibration of the vocal folds. It corresponds to regions of voiced speech with substantial, abrupt, cycle-to-cycle changes in either the spacing or the amplitude of the glottal impulses or both. In such cases, the deviation from periodicity exceeds the usual jitter and shimmer values that are present in regular phonation (see Surana & Slifka 2006).

According to the international literature, irregular phonation is a multifunctional phenomenon. Regarding the aims of the present study the most relevant role of irregular phonation is boundary marking. Irregular phonation often occurs before a word-initial vowel in English (Dilley et al. 1996), in German (Kohler 1994), and in Hungarian (Markó 2013). Comparisons in German revealed that the appearance of word-initial irregular phonation is more frequent if the speech rate is slower (Pompino-Marschall & Żygis 2010; 2011). In Polish phrase-initial irregular phonation is more probable if the lexeme is a content word (as opposed to when it is a function word) (Malisz, Żygis & Pompino-Marschall 2013).

Some of the glottalization's functions were also shown for Hungarian speech (Markó 2013), however a systematic analysis of the articulation rate and word class has not been carried out so far. Therefore, in the present study we plan to answer the following questions: (1) Is irregular phonation significantly more frequent in the case of word-initial vowels if the articulation rate is slow (as opposed to fast)? (2) If it is so, can this effect be shown in spontaneous and/or read speech? (3) Does the type of the word (content word vs. function word) have an effect on glottalization?

Our hypotheses are the following. H1: Irregular phonation is significantly more frequent in the case of word-initial vowels if the articulation rate is slow, but the frequency of occurrence of glottalization doesn't change hand in hand with the articulation rate. H2: In read utterances the articulation rate effect is larger than in spontaneous speech. H3: Word-initial vowels are glottalized to a similar extent in case of content words and function words (due to the forms of the Hungarian definite articles *a*/*az*).

In order to investigate these questions, spontaneous and read utterances of 12 speakers (6 females, 6 males, aged between 20 and 45 years) were selected from BEA Hungarian spoken language database (Gósy 2012). The read material consisted of 25 utterances, while from the spontaneous subcorpus approximately 2 minutes of speech per subject in an interview situation were chosen. The annotation of glottalized realizations was performed in accordance with the methodology of previous studies (e.g. Dilley et al. 1996; Bőhm & Ujváry 2008; Markó 2013), combining visual and auditive information, in Praat (Boersma & Weenink 2015).

In the present study all of the word-initial vowels that occurred in the material will be analyzed. We are planning to measure the duration, the articulation rate and the average syllable duration of the entire speech interval concerned, as well as the duration of the word-initial vowel and the syllable. Relative frequency of occurrences and possible correlations between the data of duration and rate and occurrences of irregular phonation will be analyzed.

References

Boersma, P., Weenink, D. 2015. *Praat: doing phonetics by computer* [Computer program]. http://www.praat.org.

Bőhm, T., Ujváry, I. 2008. Az irreguláris fonáció mint egyéni hangjellemző a magyar beszédben [Irregular phonation as an individual speaker's characteristic in Hungarian speech]. *Beszédkutatás 2008*, 108–120.

Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24, 423–444.

Gósy, M. 2012. BEA: A multifunctional Hungarian spoken language database. *Phonetician* 105–106, 51–62.

Kohler, K. J. 1994. Glottal stops and glottalization in German. *Phonetica* 51: 38–51.

Malisz, Z., Żygis, M. & Pompino-Marschall, B. 2013. Rhythmic structure effects on glottalisation: A study of different speech styles in Polish and German. *Laboratory Phonology* 4(1): 119–158.

Markó, A. 2013. *Az irreguláris zönge funkciói a magyar beszédben* [The functions of irregular phonation in Hungarian speech]. Budapest: ELTE Eötvös Kiadó.

Pompino-Marschall, B. & Żygis, M. 2010. Glottal marking of vowel-initial words in German. In: Weirich, M. & Jannedy, S. eds. *Papers from the Linguistics Laboratory. ZASPiL* 52: 1–17.

Pompino-Marschall, B. & Żygis, M. 2011. Glottal marking of vowel-initial words in German. In: Lee, W-S. & Zee, E. eds. *Proceedings of the XVIIth International Congress of Phonetic Sciences*. Hong Kong. 1626–1629.

Surana, K. & Slifka, J. 2006. Acoustic cues for the classification of regular and irregular phonation. In: *Proceedings of Interspeech* 2006. 693–696.

Larynx movement in the production of Georgian ejective sounds

Reinhold Greisbach, Anne Hermes and Alexandra Bückins
IfL Phonetik, University of Cologne, Germany

In this study, we present a new non-invasive method for investigating laryngeal movement in the production of ejective sounds. Being non-invasive this method can be used easily in the study of spontaneous speech.

Ejective sounds are relatively rare in European languages. The production of ejectives involves a non-pulmonal airstream mechanism. The airstream is invoked by raising of the closed larynx. At the same time there has to be a constriction (plosive, fricative) taking place in the supraglottal space, namely in the mouth. The raising of the closed glottis leads to an increase in pressure in the space behind the constriction. Due to the greater pressure drop, ejectives sound more prominent compared to pulmonal sounds (Ladefoged & Maddieson, 1996:78).

However, although the raising of the larynx is clearly visible from the outside especially in male speakers, the production mechanism of ejective sounds is not very well understood. In a pilot study, we used Electromagnetic Articulography (EMA) to investigate the larynx movement during the production of ejective sounds in Georgian.

Typically EMA is used to monitor tongue and lip movements in speech production. In this study, we placed 4 sensor coils on the outside of the skin just above the larynx in the area of the cricoid cartilage of a male speaker.

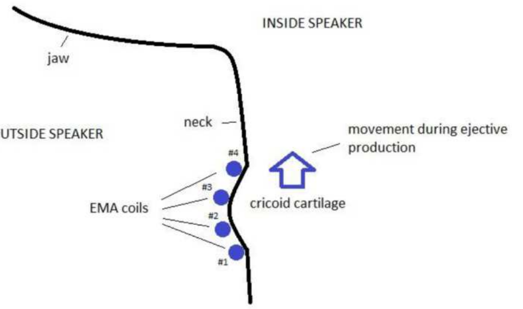


Figure 1: Position of the EMA coils on the outside of the neck just above the larynx in midsagittal plane

For this pilot study, one speaker was recorded, producing various Georgian words which contain ejective plosives, e.g. ტოტი [t’ot’i]. Thereafter, the same words were produced with aspirated voiceless plosives, e.g. თოთი [toti]. Thus, we could analyze a couple of pseudo minimal pairs, contrasting real words containing ejectives with nonsense words having aspirated plosives at the same spot in the word.

A preliminary analysis reveals that there is considerably greater movement of the coils during the production of ejectives as compared to the pulmonal sounds. Contrary to the expectations based on the traditional articulatory description of these sounds the magnitude of the movement in vertical direction (i.e. along the neck surface, up - down) is almost the same as in horizontal direction (i.e. perpendicular to the neck surface, forward - backward). Moreover the movement of the EMA coils depends on the manner and place of articulation of the ejectives. Affricate ejectives and apical articulation imply stronger movements than plosive ejectives and labial and velar articulation.

We will present a detailed analysis of the movements of the EMA coils and show how this method can help to understand ejective production mechanism.

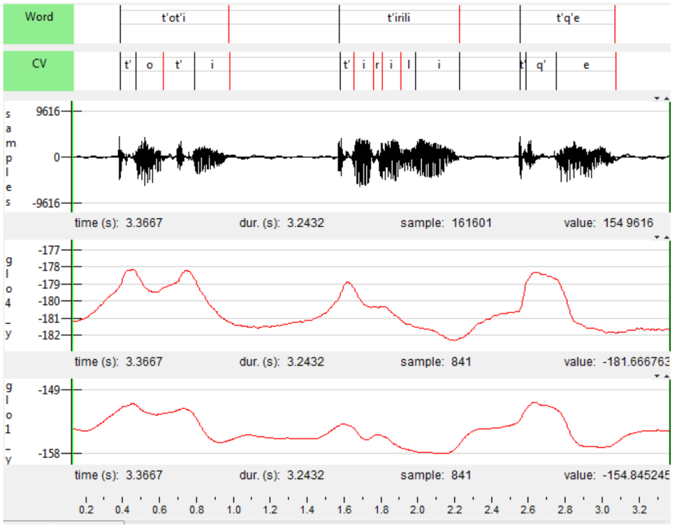


Figure 2: Oscillogram and tracks of coils 1 and 4 coping movement perpendicular to the neck surface pronouncing the Georgian words ტოტი [t’ot’i], ტირილი[t’irili], ტყე [t’q’e].

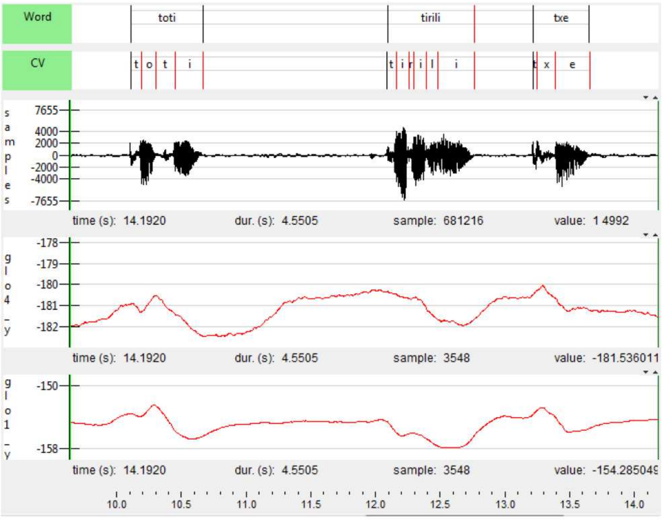


Figure 3: Oscillogram and tracks of coils 1 and 4 coping movement perpendicular to the neck surface pronouncing the Georgian nonsense words თოთი [toti], თირილი[tirili], თხე [txe]

References

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

Phrase-final lengthening of phonemically short and long vowels in Hungarian spontaneous speech across ages

Mária Gósy[1] and Valéria Krepsz[1,2]
[1]Dept. of Phonetics, Research Institute for Linguistics, HAS, [1,2]Dept. of Phonetics,
Eötvös Loránd University

Introduction

Phrase-final lengthening is a phenomenon that has been known in phonetics for several decades (e.g., Lindblom 1968). For definition, the last syllable of the word is lengthened in a phrase-final position, at a prosodic boundary or before a phrase-final pause resulting in longer duration than that of a segmentally identical phrase-medial syllable. Several factors are suggested that might trigger the lengthening of the vowels like subglottal pressure, decreasing articulation activity, linguistic, phonological, and higher-level factors, as well as syntactic structures, stress patterns, etc. (e.g., Den 2015). Lengthening might concern the vowels or the consonants of the phrase-final syllable, or the whole syllable (Dimitrova & Turk 2012). The phenomenon has been reported, on the basis of controlled experiments, to exist in various languages, irrespective of their typological and prosodic patterns, but also to show specific differences across languages (Cho 2016). Vowel quantity as a phonemic distinction was also shown to interact with phrase-final lengthening (Nakai et al. 2009). The questions arise whether Hungarian speakers regulate utterance-final lengthening to preserve the phonemically different quantity of vowels in spontaneous speech, and whether this regulation is dependent on age. In addition, word lengths differences might also affect the temporal patterns of phrase-final lengthening. Four hypotheses were formulated. (i) Phrase-final lengthening will preserve the phonemic quantity differences of the target vowels irrespective of the speakers' age, (ii) target vowels will not show durational differences in phrase-initial and phrase-medial positions in old speakers' speech, (iii) the number of syllables of the words will influence the durations of the target vowels occurring in phrase-final positions, and (iv) the length of words will have a greater effect on vowel durations as produced by old speakers than those of young speakers.

Methodology

Spontaneous narratives (more than 5 hours' material) of 10 young subjects (aged between 20 and 30) and 10 old ones (aged between 70 and 80) were randomly selected from the BEA Spontaneous Speech Database of Hungarian (Gósy 2012). Each group consisted of an equal number of females and males. A phonemic pair of short and long vowels ([ɔ, a:]) was selected as target vowels (they are, however, different in tongue height and lip rounding). More than 2,400 vowel tokens were identified in words with various numbers of syllables (from 2 to 6) occurring in phrase-initial, phrase-medial and phrase-final positions in the last syllables of the words. All syllables containing the target vowels were unaccented. Durations of the vowels were taken by measuring the interval between the onset and offset of the second formants of the vowels based on annotated files in Praat software (Boersma & Weenink 2012). A specific script was written for obtaining the values automatically. Data were normalized across speakers using the z-normalization method. Durations were analyzed according to (i) vowel quality, (ii) word length, (iii) word position in the phrase, and (iv) speakers' age. To test statistical significance, General Linear Mixed Model analyses were carried out to test the effects of the fixed factors 'position', 'vowel quality', 'word length', 'age' and their interactions on durations of the vowels (dependent factors). The confidence level was set at the conventional 95%.

Results

Both phonemically short and long vowels were significantly longer in phrase-final positions than in phrase-initial and phrase-medial positions in both age groups. Durations of vowels in phrase-medial positions were significantly shorter than those occurring in phrase-initial positions in young speakers' speech while there were no differences in their durations between the two positions in old speakers' speech (Fig. 1).

Phonemically long vowels were significantly longer than phonemically short vowels in all positions irrespective of age. Short vowels produced by old speakers in phrase-final positions were significantly longer than those produced by young speakers. The opposite tendency was found in the case of long vowels. Long vowels produced by old speakers in the same positions were significantly shorter than those produced by young speakers.

Word length had a significant effect on vowel durations that were the least variable in phrase-medial positions in both age groups. The largest ranges of the durations were found in vowels produced in phrase-final positions across various word lengths in both age groups. Vowel durations in words containing more than three syllables were significantly shorter in old speakers' speech than in young speakers' speech.

Conclusions

Our results confirm that utterance-final lengthening does exist in Hungarian spontaneous speech irrespective of age; they also exhibit the phenomenon that phonemic vowel quantity contrasts are preserved. The data supported our hypotheses. The strong distinction of short and long vowels also in phrase-final positions suggests that speakers avoid violating the phonemic patterns of the vowel system. The similar durations of the target vowels in phrase-initial and phrase-medial positions as well as the decrease of durations along with the increase of word length in old speakers' speech is assumed to be the consequences of both their breathing and cognitive processing (Hooper–Cralidis 2009). Discussion of the results will concern the interrelations of vowels' phonemic quantity, their positions and the length of the words, appearing in part differently in young and old speakers' speech.

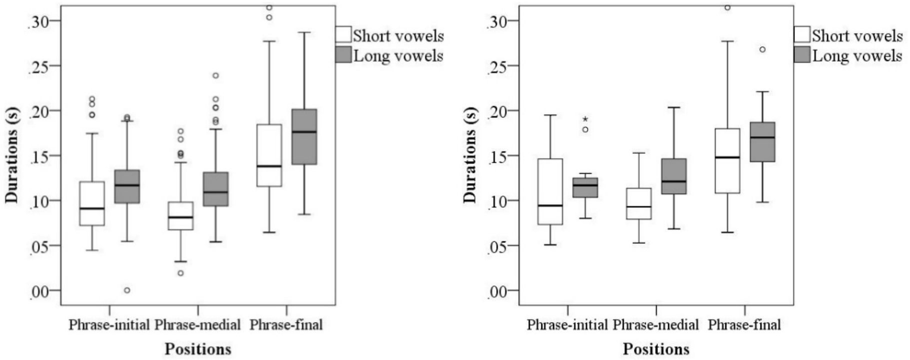


Figure 1: Durations of vowels in various phrase positions (left: young speakers, right: old speakers)

References

Boersma, P. & Weenink, D. 2015. *Praat: doing phonetics by computer*. http://www.praat.org

Cho, T. 2016. Prosodic boundary strengthening in the phonetics-prosody interface. Language and Linguistics Compass 10, 120–141.

Den, Y. 2015. Some phonological, syntactic, and cognitive factors behind phrase-final lengthening in spontaneous Japanese: A corpus-based study. *Lab. Phonology* 6, 337–379.

Dimitrova, S. & Turk, A. 2012. Patterns of accentual lengthening in English four-syllable words. *Journal of Phonetics* 40, 403–418.

Gósy, M. 2012. BEA – A multifunctional Hungarian spoken language database. *The Phonetician* 105/106, 50–61.

Hooper, C. R. & Cralidis, A. 2009. Normal changes in the speech of older adults: You've still got what it takes; it just takes a little longer! *Perspectives on Gerontology* 14, 47–56.

Lindblom, B. 1968. Temporal organization of syllable production. *Speech Transmission Laboratory Quarterly Progress* 9, 1–6. Stockholm: Royal Institute of Technology.

Nakai, S., Kunnari, S., Turk, A., Suomi, K. & Ylitalo, R. 2009. Utterance-final lengthening and quantity in Northern Finnish. *Journal of Phonetics* 39, 29–45.

MagmaNet: Ensemble of 1D Convolutional Deep Neural Networks for Speaker Recognition in Hungarian

Attila Gróf[1], Annamária Kovács[2,3], Anna Moró[1], Miklós Gábriel Tulics[2] and Máté Ákos Tündik[2]

[1]Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics,
[2]Budapest University of Technology and Economics, [3]Research Centre for Natural Sciences, Hungarian Academy of Sciences, Hungary

Speaker recognition is one of the classical and widely documented tasks in speech technology. In short it can be said that speaker recognition is the process of automatically recognizing the person behind the voice, on the basis of information obtained from his or her speech.

Building speaker recognition systems has a large literature, especially from GMM/HMM era, which were the first pioneers in this field, but Neural Network-based solutions are gaining ground in this area as well.

Zaki and his colleagues used cascade neural networks [1] for speaker classification, others, such as Zakariya et al. used i-vectors and ANN for text-independent speaker classification [2].To achieve higher performance using different ANNs committee neural networks were proposed by Narender et al. [9]: several ANNs were trained to achieve excellent recognition, but in the final step, only the five best performing networks were fed into the (final) committee network. This way, the final decision was determined based on the majority voting of the member networks.

The authors of [11] used a deep 2D-Convolutional Neural Network (CNN), applied to the raw spectrograms. The main idea was to create convolutions that sample raw information across both frequency and time makes the spectrograms suitable for CNN analysis. The input was an image of size 513 by 107, a spectrogram of a 20ms segment of an utterance. They used a pre-trained network, AlexNet for the classification. They reached 17,1% Equal Error Rate (EER). An i-vector based system was also trained and tested on the same data, resulting in 39,7% EER.

The state-of-the-art approaches for text-independent speaker recognition are using Joint Factor Analysis (JFA) or i-vector based modeling. JFA is a powerful and widely used technique for compensating the variability caused by different channels and sessions. The total variability i-vector modeling has gained significant attention in speaker recognition due to its excellent performance, low complexity and resulting small model sizes. The authors of [10] reached 2,53% EER.

Although several ANN/DNN implementation exist in the speaker recognition field, one-dimensional convolution was not applied specifically to this problem. However, not only the architecture of the neural network is important, it is substantial to choose the proper speech features for this task as well. Much depends on the appropriate feature selection. There are well-established speech features for these tasks: Melfrequency cepstral coefficients (MFCCs) [3], linear-prediction coefficients (LPC) [4], mean Hilbert envelopes [5], and also hybrid feature-sets [6] [7].

In this research we showed that using a low number of features is sufficient to train a DNN-based system to perform this classification task with a high performance. For this reason, we used a small database containing a total of 161 sound recordings from 11 native Hungarian speakers, reading “The NorthWind and the Sun” folk-tale. We used five speakers to differentiate them, the other speakers were used as an impostor model.

MFCCs, LPCs and Linear Predictive Cepstral Coefficients (LPCC) were extracted in 30 ms windows with 10 ms overlaps, these features were used as input vectors. We implemented four neural network models: a Multi-layer Perceptron (MLP), a 1D ConvNet and a 1D Dilated ConvNet with an LSTM Layer, and an Ensemble one called ’MagmaNet’1. Our 1D Convolutional - based Neural Network Architectures can be seen on Figure 1.

The impact on the classification accuracy of the acoustic parameters was examined. In case of MLP the best classification results was obtained using only MFCC features (47%). As one of the simplest DNN architectures, the MLP classification accuracy can be considered as a baseline. Convolutional Networks yield better accuracies when MFCC and LPCC features were combined together. Both networks reached an accuracy of 76%. Manual and automatic hyperparameter optimization was performed for each network.

In hope of getting further improvement in classification accuracy, in our first attempt the two convolutional networks were merged and were simultaneously trained. We performed the same processing steps on the Ensemble model like on the other networks. First, we selected the best feature set (MFCC) for classification, then we performed manual hyperparameter optimization obtaining accuracy results of 74% on the validation set and 68% on the test set. Our second attempt was using pre-trained 1D Convolution models, storing their weights, and applying them in a transfer learning approach.

This ’MagmaNet’ Ensemble model reached 78% accuracy on the test set, which is a quite good achievement for this classification task, on this limited database. We can show comparable results in accuracy with [12], who investigated speaker identification in TV Broadcast data. Although the data is different, the task is similar.

We have some ideas for future improvement: we would like to examine various number of features, so take more effort to feature selection. We are also aware of the limitation of our research, thus we are going to switch to a public database containing more recordings. All in all, our results confirmed the plausibility of using 1D Convolution based DNN as a means for implementing a valuable speaker recognition solution.

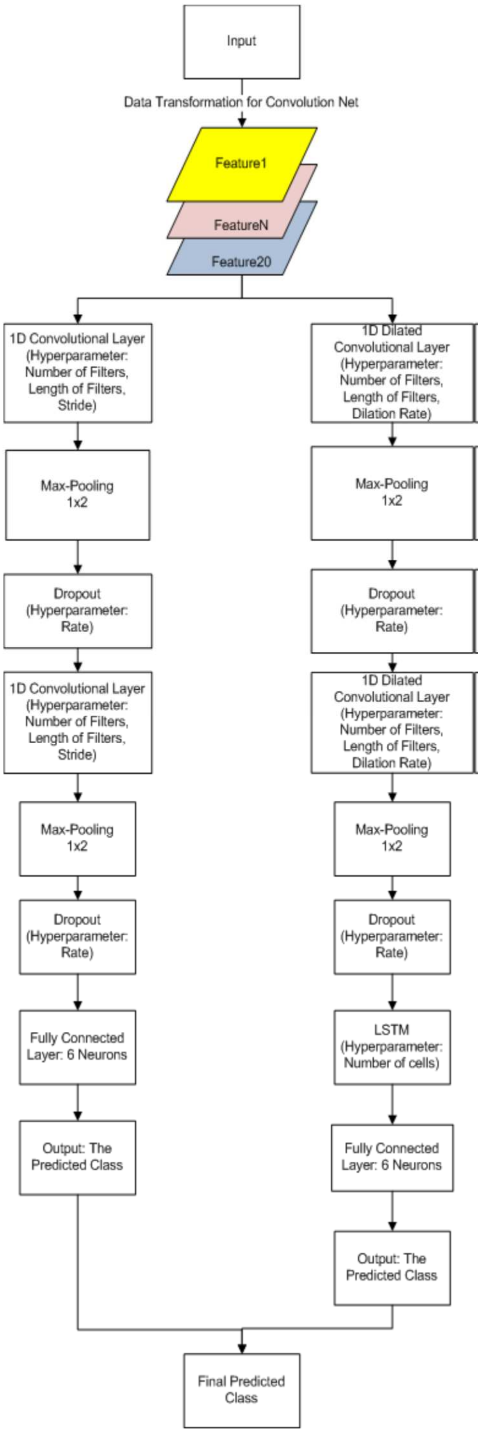

Input
Data Transformation for Convolution Net
Feature1
FeatureN
Feature20
1D Convolutional Layer (Hyperparameter: Number of Filters, Length of Filters, Stride)
1D Dilated Convolutional Layer (Hyperparameter: Number of Filters, Length of Filters, Dilation Rate)
Max-Pooling 1x2
Max-Pooling 1x2
Dropout (Hyperparameter: Rate)
Dropout (Hyperparameter: Rate)
1D Convolutional Layer (Hyperparameter: Number of Filters, Length of Filters, Stride)
1D Dilated Convolutional Layer (Hyperparameter: Number of Filters, Length of Filters, Dilation Rate)
Max-Pooling 1x2
Max-Pooling 1x2
Dropout (Hyperparameter: Rate)
Dropout (Hyperparameter: Rate)
Fully Connected Layer: 6 Neurons
LSTM (Hyperparameter: Number of cells)
Output: The Predicted Class
Fully Connected Layer: 6 Neurons
Output: The Predicted Class
Final Predicted Class

Figure 1. The MagmaNet components: 1. 1D ConvNet:on the left, 2. 1D Dilated ConvNet with LSTM: on the right. Ensemble Model: together

References

[1] M. Zaki, A. Ghalwash, A. Elkouny, 1996. Speaker recognition system using a cascade neural network, Int. J. Neural Syst. 7. 203–212.

[2] Zakariya Qawaqneh, Arafat Abu Mallouh, Buket D. Barkana 2017. Deep neural network framework and transformed MFCCs for speaker's age and gender classification, Knowledge-Based Systems, Volume 115.

[3] P. Mermelstein 1976. ”Distance measures for speech recognition, psychological and instrumental,” in Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed., pp. 374-388. Academic, New York.

[4] S.B. Davis, and P. Mermelstein, 1980. ”Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” in IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357–366.

[5] Seyed Omid Sadjadi, John H.L. Hansen 2015. Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification, Speech Communication, Volume 72. 138–148.

[6] Leandro D. Vignolo, S.R. Mahadeva Prasanna, Samarendra Dandapat, H. Leonardo Rufiner, Diego H. Milone 2016. Feature optimisation for stress recognition in speech, Pattern Recognition Letters, Volume 84. 1–7.

[7] Ji-Won Cho, Hyung-Min Park 2016. Independent vector analysis followed by HMM-based feature enhancement for robust speech recognition, Signal Processing, Volume 120. 200–208.

[8] S. K. Singh, Supervisor: Prof P. C. Pandey: Features and Techniques for speaker recognition

[9] Reddy N.P., Buch, O., 2003. Speaker verification using committee neural networks, Computer Meth. Programming in Biomed., Vol. 72. 109–115.

[10] Ming Li, Andreas Tsiartas, Maarten Van Segbroeck and Shrikanth S. Narayanan 2014. Speaker verification using simplified and supervised i-vector modeling.

[11] Lior Uzan, Lior Wolf: I Know That Voice: Identifying the Voice Actor Behind the Voice

[12] Mateusz Budnik, Laurent Besacier, Ali Khodabakhsh, Cenk Demiroglu. 2016. Deep complementary features for speaker identification in TV broadcast data. Odyssey Workshop 2016, Jun 2016, Bilbao, Spain. Odyssey.

Variability in Speech Sound Production: Covert Contrasts in the Speech of Children with Cochlear Implants

Plenary speech by

Ruth Huntley Bahr
University of South Florida, USA

Covert contrast is often used as evidence against the traditional view that pronunciation shifts in children are caused by solely phonological changes (Scobbie, et al., 1996). As such, covert contrasts represent intermediate productions that are their own stage of learning, allowing us broader insight into how children acquire a phonological system (Hewlett & Waters, 2004). Additional support for this idea comes from the fact that children with speech sound disorders who produce covert contrasts have much better prognoses than those who do not (Byun, et al., 2015). The fact that a child produces a covert contrast between two phonemes suggests that he/she can perceive some difference between them (Byun, et al., 2015). The perception of subtle differences in speech sounds is essential for individuals who may receive a distorted or diminished speech signal, such as children who use cochlear implants (CIs).

This presentation will describe two projects in the area of covert contrast with children who use CIs. The first study focuses on the accuracy of speech sound productions in young children with CIs when compared to their speech age-matched normal hearing (NH) peers. The second project is a pilot investigation that considers the utility of different scales in the assessment of covert contrast.

Study1. Nine children (mean age = 57 months) who were implanted by age 3 and had used their CIs for one year were matched by articulation age to 9 typically developing NH peers (Gonzalez, 2013). Speech sound testing revealed that phonetic inventory size and intelligibility were comparable across groups. Covert contrast analysis focused on the production of VCV non-words containing /t, d, tʃ/ taken from the OlimSpac (Boothroyd et al., 2006). Two different listening experiments were conducted that focused /t-d/ and /t-tʃ/ contrasts. Thirty-three graduate students in speech-language pathology rated the phonetic accuracy of the C produced in the VCV syllable using a visual analogue scale (VAS).

A confusion matrix of the children's productions on the OlimSpac indicated that /t/ productions in children with CIs were frequently distorted, despite phoneme mastery. The results of a 3-way repeated measures ANOVA revealed that children with NH showed a typical pattern of speech sound acquisition. The earlier developing /d/ had a large, well-developed contrast, however the later developing /tʃ/ showed little contrast with /t/. Children with CIs demonstrated the opposite trend. The t/d substitutions were much more /t/-like, indicating that they were not making sufficient covert contrast and that /t/ was less developed than in children with NH. Children with CIs displayed a larger contrast for /t/ and /tʃ/ than the children with NH. This finding suggested that children with CIs struggled more with voicing than affrication. Children with CIs also showed a larger contrast for /tʃ/ than children with NH, supporting previous findings revealing earlier emergence of /tʃ/.

Taken together, these results suggest that children with CIs approach the speech acquisition task differently than children with NH. They may be attending to acoustic cues (such as aspiration) in an idiosyncratic way, or weighting these cues inappropriately. These perceptions may be a learned behavior, acquired from several years of aural habilitation, or a tendency that all children with CIs have due to the nature of the sound-processing in CI technology.

Study 2. This pilot study was designed to test the utility of different scales in measuring covert contrast. One scale was similar to a traditional VAS; it was a single line with “r” and “l” at the end points and “w” at the center. The other scale was triangular, with “r”, “l”, and “w” at each corner. Given the acoustic similarity of these phonemes and the common substitution patterns among them, the utility of a traditional VAS versus the new triangular rating scale was investigated. The goal was to evaluate the sensitivity of these scales in identifying covert contrast across participant group and phoneme.

Productions of /r, l, w/ were extracted from single words produced by the same participants in Study 1. These phones were arranged in listening experiments, each testing a different rating scale (VAS versus triangular). Students in speech-language pathology listened to the phones and rated the quality of production. Differences in scale utility were determined across speaker group and phonemes.

Results indicated that both rating scales were sensitive to subtle acoustic differences in speech sound production. However, listeners preferred the triangular scale for the /r, l, w/ contrast because it provided greater response sensitivity. Results also indicated more variability in sound production within CI users.

Can you speak less dialect, please? Phonetic modifications enabling understanding between members and non-members of a dialect community

Sarah Brandstetter
University of Vienna, Austria

There is a broad variety of dialects spoken in Austria which fall into two bigger categories: Bavarian and Alemannic dialects. Both groups pose a variety of difficulties for non-native speakers of German as dialects are widely used in daily life. When speaking to non-dialect speakers, dialect-speakers are likely to try to use a more standard variety of German in order to be understood. On the segmental level, these changes concern e.g. less laxing of the vowel and less diphthongization of the laxed vowel (Wiese 1996), another change that could be reversed is that in the Austrian standard variety the intervocalic /b, d, g/ is lenited to fricatives [β, ð, ɣ] (Moosmüller, 2007). The current study investigates segmental changes that occur in speakers changing their style from spontaneous dialect speech to learner-directed speech, focusing on the typical elements of the dialect e.g. the vocalization of liquids, the reduction of fortis consonants, the omission of lenis consonants, a change in the roundedness of vowels, a reduction or omission of vowels in unstressed syllables, an unclear pronunciation of <a> and additional diphthongs.

Study

Communication difficulties of speakers of different languages or dialects are a common phenomenon in foreign language acquisition. Dialect speakers in Austria tend to try to speak less dialect when talking to non-dialect or non-native speakers, moving on the continuum from dialect to standard German in order to make themselves better understood. But they will not resort to "full" standard German, switching continuously back and forth. This is also one of the phenomena Berend and Frick (2016) found when investigating which elements of their dialect varied when members of the German minority in Russia who moved back to Germany were talking to members of their group as opposed to non-members. In the current study I hypothesize that the speakers will modify their articulation only if they themselves deem certain words "difficult" to understand whereas they tend to not change anything in the – as they seem to think – more "basic" lexicon. Furthermore, I expect that they change their articulation regardless of the fact if the word actually exists in the standard German lexicon.

Based on these two assumptions, the present study explores which phonetic features of their dialect the speakers drop or substitute in order to reach an understanding and which ones will remain nonetheless. How far do speakers of a specific Bavarian dialect in rural Austria change their speech when talking to non-dialect speakers who do not live in the same area?

Since dialect is used primarily in informal situations, it is difficult to elicit authentic dialect in experimental settings, or controlled communicative situations. To reach this goal two native speakers of Austrian German and two non-native speakers who speak German at a very advanced level visited Wankham, a small village in Upper Austria, during February 2017. Conversations with members of two families who have been living there for a couple of decades and most of the time only talk to other inhabitants of the village or the surrounding area have been recorded. The dialogues cover topics that are part of the daily lives of the people living in Wankham (e.g. work, gardening).

In a pilot study, major phonetic differences were found between the spoken standard variety of (Austrian) German and the dialect in question. The phonetic profile of the dialect has been established evaluating recorded spontaneous speech. The result consists of 7 main and a series of minor features (as described above). One very typical feature of the dialect can be seen below in Figure 1.

The main research part consists of recorded dialogues which are analyzed in-depth with regard to phonetic changes, e.g. in word codas. Modifications were collected and systemized in order to find strategies underlying the modifications. Phonetic analyses were made using Praat. Sample audio files underline the differences between the usual dialect the speakers use and the "modified" dialect that is used with people who are not members of their dialect community.

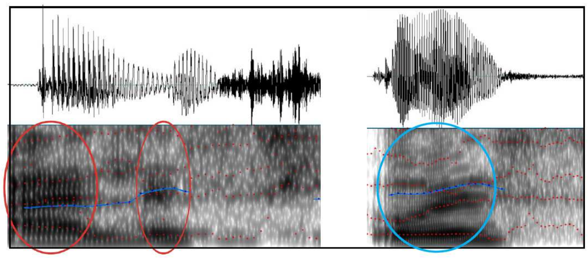


Figure 1: The utterance /alles/ (engl: "everything") in the Austrian standard variety (left) compared to the dialect /ois/ (right).
In this particular dialect, liquid consonants tend to be vocalized

Relevance of the topic

The dialects of German spoken in Austria are often a difficulty that people who learn German as a foreign or second language encounter during their stay in Austria. Learners struggle to understand what dialect speakers talk about, and this often makes it difficult for them to get in touch with Austrian native speakers. Yet there is not much that is done to address this issue in German as a foreign language (GFL) classes in Austria. The aim of my research is to collect experimental evidence of spontaneous speech data in order to reveal changes that occur when switching from dialect spoken to other dialect speakers to a variety that contains less elements of dialect but can also not be considered the standard variety. This allows learners of GFL to study the more persistent elements of the dialect in question which will empower them with respect to their receptive capabilities of spoken dialect.

References

Auer, P. – Hinskens, F. – Kerswill, P. (ed.) 2005. Dialect Change. Convergence and Divergence in European Languages. New York: Cambridge University Press.

Berend, N. – Frick, E. 2016. Dialektwandel und Veränderung der individuellen Varietätenrepertoires: Ergebnisse und Materialien einer empirischen Untersuchung zur Standard – Dialekt-Variation bei russlanddeutschen Aussiedlern in Deutschland. Mannheim: Institut für Deutsche Sprache.

Ender, A. – Kaiser, I. 2009. Zum Stellenwert von Dialekt und Standard im österreichischen und Schweizer Alltag. in: Zeitschrift für germanistische Linguistik. 37:2. S.266–295.

Hornung, M. – Roitinger, F. 1950. Unsere Mundarten. Eine dialektkundliche Wanderung durch Österreich. Wien: Österreichischer Bundesverlag.

Moosmüller, S. 2007. Vowels in Standard Austrian German. An acoustic-phonetic and phonological analysis. Habilitationsschrift. Wien.

Wiese, R. 1996. The Phonology of German, Oxford: Oxford University Press.

Wiesinger, P. 2014. Das österreichische Deutsch in Gegenwart und Geschichte. Wien: Lit Verlag

Phonetic characteristics of disfluent word-repetitions: The effect of age and speech task

Judit Bóna and Tímea Vakula
Department of Phonetics, Eötvös Loránd University, Hungary

In spontaneous speech, one of the most frequent disfluencies is word-repetition (Shriberg 1995) which can indicate word-finding problem, difficulty in conceptual planning, or covert self-monitoring (Plauché–Shriberg 1999). Repetitions consist of several parts which are the following: the original utterance, the first instance of the repeated word, the second instance of the repeated word and the continuation of the utterance (Plauché–Shriberg 1999). Optional pauses may also occur next to the main parts (Plauché–Shriberg 1999). The duration of the first and second instances of the repeated words and the occurrence of pauses are related to the function of the repetitions. Plauché and Shriberg (1999) found three main types of functions: canonical repetition, covert self-repairs, and stalling repetition.

The characteristics of repetitions (like other disfluencies) are influenced by several factors. The aim of this study is to analyse temporal features and functions of the repeated words in diverse age groups and speech tasks. The hypotheses of the research are: 1) there will be difference in the temporal characteristics and functions of repetitions between the age groups in both speech tasks; 2) there will be significant difference between the speech tasks in each age group. The different cognitive efforts of planning the different speech tasks will be shown by the differences in temporal features of word-repetitons.

For this presentation, speech recordings of 80 speakers were selected from two Hungarian speech databases. Speech samples of school children (9-year-olds), and adolescents (13-14-year-olds) were selected from the GABI Hungarian Children Speech Database and Information Repository (Bóna et al. 2014). Speech samples of young adults (20-25-year-olds) and elderly (70+) speakers were selected from BEA Hungarian Speech Database (Gósy 2012). In every age group there were 20 speakers (10 women and 10 men). They were native Hungarian speakers with normal hearing and without any known mental or speech disorders. They all spoke standard Hungarian.

Recordings were made with each subject in two situations which represented different speech tasks: 1) spontaneous narrative (participants spoke about their own lives and families), and 2. narrative recall (the task was to recall two texts they had listened to as accurately as possible).

The annotation and measurements (duration of the components of repetitions) were carried out by Praat. We analysed the ratio of the first instance of the repeated word and the second instance of the repeated word and the pauses before, between and after them. Functions were examined, too.

The preliminary results confirmed our hypothesis on age- and speech-task-dependent properties of the repetitions in relation to both functions and temporal patterns. The presentation discusses these differences in depth. Further, it covers the theoretical and practical implications of the results.

Acknowledgement

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. K-120234.

References

Bóna, J., Imre, A., Markó, A., Váradi, V. & Gósy, M. (2014). GABI – Gyermeknyelvi Beszédadatbázis és Információtár, *Beszédkutatás 2014*: 246–252.

Gósy, M. (2012). BEA – A multifunctional Hungarian spoken language database, *Phonetician* 105–106.: 50-61. http://www.isphs.org/Phonetician/Phonetician_105_106.pdf

Plauché, M., & Shriberg, E. (1999). Data-driven subclassification of disfluent repetitions based on prosodic features. In *Proc. International Congress of Phonetic Sciences* Vol. 2. 1513-1516.

Shriberg, E. (1995). Acoustic properties of disfluent repetitions. In *Proceedings of the international congress of phonetic sciences*. Vol. 4., 384–387.

Exploiting prosodic and word embedding based features for automatic summarization of highly spontaneous Hungarian speech

György Szaszák[1] and András Beke[2]
[1] Dept. of Telecommunication and Media Informatics,
Budapest University of Technology and Economics, Hungary
[2] Research Institute for Linguistics, HAS, Hungary

In this contribution, the authors address speech summarization for Hungarian, using highly spontaneous speech material. As the first step, the audio signal is transcribed using an Automatic Speech Recognizer, and speech summarization is carried out on these transcriptions using various text analysis methods. From the speech stream, we also exploit prosody based information, which is used for tokenization prior to textual analysis. We evaluate this prosody based tokenization approach against human performance. The so obtained intonational phrase like tokens are then converted into virtual sentences, and analyzed by the syntactic parser to help ranking based on thematic terms and sentence position. The thematic term is expressed in two ways: TF-IDF and Latent Semantic Indexing. Word embeddings can also be exploited for more robust thematic term calculation. The sentence scores are calculated as a linear combination of the thematic term score and a positional score. The final summary is generated from the top N candidates. Results show that prosody based tokenization reaches human average performance. Audio summarization shows 0.62 recall and 0.79 precision by an F-measure of 0.68, compared to human reference (N=10). Taking into account the high spontaneity of the speech, this results are very encouraging. A subjective test is also carried out on a Likert-scale to allow for a more complete evaluation.

Speech can be processed automatically in several application domains, including speech recognition, speech-to-speech translation, speech synthesis, spoken term detection, speech summarization etc. These application areas use successfully automatic methods to extract or transform the information carried by the speech signal. However, the most often formal, or at least standard speaking styles are supported and required by these applications. The treatment of spontaneous speech (Neuberger et al., 2014) constitutes a big challenge in spoken language technology, because it violates standards and assumptions valid for formal speaking style or written language and hence constitutes a much more complex challenge in terms of modelling and processing algorithms.

Automatic summarization is used to extract the most relevant information from various sources: text or speech. Speech is often transcribed and summarization is carried out on text, but the automatically transcribed text contains several linguistically incorrect words or structures resulting both from the spontaneity of speech and/or speech recognition errors. To sum up, spontaneous speech is "ill-formed" and very different from written text: it is characterized by disfluencies, filled pauses, repetitions, repairs and fragmented words, but behind this variable acoustic property, syntax can also deviate from standard.

Another challenge originates in the automatic speech recognition step. Speech recognition errors propagate further into the text-based analysis phase. Whereas word error rates in spoken dictation can be as low as some percents, the recognition of spontaneous speech is a hard task due to the extreme variable acoustics (including environmental noise, especially overlapping speech) and poor coverage by the language model and resulting high perplexities (Szarvas et al., 2000). To overcome these difficulties, often lattices or confusion networks are used instead of 1-best ASR hypotheses (Hakkani-Tür et al, 2006).

A possible approach of summarizing written text is to extract important sentences from a document based on keywords or cue phrases. Automatic sentence segmentation (tokenization) is crucial before such a sentence based extractive summarization (Liu–Xie, 2008). The difficulty comes not only from incomplete structure (often identifying a sentence is already problematic) and recognition errors, but also from missing punctuation marks, which would be fundamental for syntactic parsing and POS-tagging. Speech prosody is known to help in speech segmentation and speaker or topic segmentation tasks (Shriberg et al., 2000). In current work we propose and evaluate a prosody based automatic tokenizer which recovers intonational phrases (IP) and use these as sentence like units in further analysis. Summarization will also be compared to a baseline version using tokens available from human annotation. The baseline tokenization relies on acoustic (silence) and syntactic-semantic interpretation by the human annotators.

In addition, a word-embedding based approach is also considered for speech summarization. Word embeddings project individual words into a semantic space, where words with similar meaning are grouped together. Moreover, such semantic spaces are also able to represent inherent logic linked to meaning and can be used for analogical reasoning tasks or representations (Mikolov et al., 2013). We exploit word embeddings for grouping words with similar meaning in the semantic space, and introduce this knowledge into the thematic term calculation process.

Other research showed that using speech-related features beside textual-based features can improve the performance of summarization (Maskey–Hirschberg, 2005). Prosodic features such as speaking rate; minimuma, maximuma, mean, and slope of fundamental frequency and those of energy and utterance duration can also be exploited. Some approaches prepare the summary directly from speech, relying on speech samples taken from the spoken document (Maskey–Hirschberg, 2006).

References

Hakkani-Tür, D., Bechet, F., Riccardi, G., and Tür, G. (2006). Beyond asr 1-best: using word confusion networks in spoken language understanding. *Computer Speech and Language*, 20(4):495–514.

Liu, Y. and Xie, S. (2008). Impact of automatic sentence segmentation on meeting summarization. In *Proc. Acoustics, Speech and Signal Processing*, ICASSP 2008. IEEE International Conference on, 5009–5012.

Maskey, S. and Hirschberg, J. (2005). Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *INTERSPEECH*, 621–624.

Maskey, S. and Hirschberg, J. (2006). Summarizing speech without text using hidden Markov models. In *Proceedings of the Human Language Technology Conference of the NAACL*, Companion Volume: Short Papers, 89–92.

Neuberger, T., Gyarmathy, D., Gráczi, T. E., Horváth, V., Gósy, M., and Beke, A. (2014). Development of a large spontaneous speech database of agglutinative Hungarian language. In Text, *Speech and Dialogue*, 424–431.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1):127–154.

Szarvas, M., Fegyó, T., Mihajlik, P., and Tatai, P. (2000). Automatic recognition of Hungarian: Theory and practice. Int. *Journal of Speech Technology*, 3(3):237–251.

Vowel-formant frequencies of Hungarian–Croatian bilinguals and Hungarian monolinguals in spontaneous speech

Davor Trošelj
Doctoral School of Linguistics, Eötvös Loránd University, Hungary

Hungarian and Croatian vowel systems largely differ from each other. Hungarian has fourteen vowels: a /ɔ/, á /a:/, o /o/, ó /o:/, u /u/, ú /u:/, ö /ø/, ő /ø:/, ü /y/, ű /y:/, e /ɛ/, é /e:/, i /i/ and í /i:/ whereas Croatian has five: a /a/, e /e/, i /i/, o /o/ and u /u/. Hungarian vowels differ in their tongue position, lip rounding and duration. Croatian language does not contrast phonemically short and long vowels like Hungarian, but it does contrast four types of pitch accents on the stressed syllable: short-falling, short-rising, long-falling and long-rising. Figure 1 presents vowel charts of the two languages.

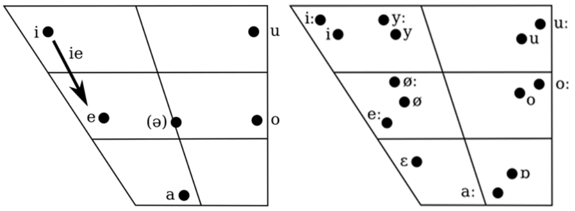


Figure 1: Left: Croatian vocal chart; right: Hungarian vocal chart (Source: https://www.internationalphoneticassociation.org/)

From Figure 1 it is obvious that Hungarian is a language with a crowded phonemic vowel space, while Croatian is a language with a sparser phonemic vowel space. Regarding their phonemically duration, out of the 14 Hungarian vowels, 5 are most similar to the Croatian ones: /i, ɛ, ɔ, o, u/. Previous studies on vowel formants in Hungarian and Croatian have shown differences in formant frequencies between the two languages. F1 of Croatian /i, e, u/ is lower than the F1 of their Hungarian counterparts /i, ɛ, u/ whereas F1 of Croatian /a, o/ is higher than the F1 of Hungarian /ɔ, o/. F2 of Croatian /a, o, u/ is lower than the F2 of their Hungarian counterparts /ɔ, o, u/, while F2 of Croatian /i, e/ is higher than the F2 of Hungarian /i, ɛ/. There have been no researches of vowel-formant frequencies in Hungarian-Croatian bilinguals so far. Since there are substantial differences in vowel quality between Hungarian and Croatian, this study aims to investigate vowel-formant frequencies of Hungarian-Croatian bilingual speakers and compare them to the ones of Hungarian monolinguals. The subjects are 10 Croatian-Hungarian bilingual adult female speakers and 10 Hungarian monolingual adult female speakers. Their spontaneous speech was recorded in a sound-attenuated booth. They were asked to talk about their jobs, free time activities and hobbies. For every speaker 7 – 8 min recording was made. Approximately 150 min material was recorded. The analysis is carried out with Praat 5.4.04 software package. Vowel-formant frequencies of F1 and F2 are analysed. Only words with C_1VC_2 in unstressed position are analysed. Formant frequencies are measured in the middle and the last syllables. In the present study only short vowels (/i, ɛ, ɔ, o, u, y, ø/) are analysed. Each vowel is represented by 20 of its occurrences in various consonant contexts (i.e. Beke & Gráczi, 2010, Auszmann, 2016). The formant values are measured in the middle of the vowel. Before the recording, the bilinguals were asked to fill a questionnaire about their language background. Two groups of speakers were formed: Hungarian dominant and Croatian dominant bilinguals. For a better comparison of the differences in dominance, the results will also be presented for both of these groups separately. Regarding five Hungarian vowels (/i, ɛ, ɔ, o, u/) whose acoustical properties are the most similar to Croatian vowels, it is expected that due to the interference of Croatian language bilinguals will produce formants that differ in their frequencies from the ones of Hungarian monolinguals, i.e. their results will show tendency towards Croatian values. When it comes to other vowels that do not have their counterparts in Croatian (and thus cannot be compared to Croatian values) it is expected that bilinguals will form a new category of vowel-formant frequencies that differs from the Hungarian monolinguals. In a comparison of Hungarian dominants and Croatian dominants it is expected that Hungarian dominants' formant frequencies will show more tendency towards values of Hungarian monolinguals. It is expected that the formant frequencies of Croatian dominants will differ from the values of Hungarian dominants. The results of the research will either provide evidence of monolingual-like production of vowels in Hungarian by Hungarian-Croatian bilinguals, or they will show a cross-language interference between the two languages. These findings will provide an insight on how Hungarian-Croatian bilinguals organize their vowel system.

References

Auszmann, A. (2016) *Magyar gyermekek magánhangzóinak akusztikai-fonetikai jellemzői*. Doktori disszeráció.

Bakran, J. (1996) *Zvučna slika hrvatskoga govora*. Zagreb: Ibis grafika.

Beke, A. & Gráczi, T. E. (2010) A magánhangzók semlegesedése a spontán beszédben. Segédkönyvek a nyelvészeti tanulmányozásához 107. *Nyelv, beszéd, írás. Pszicholingvisztikai tanulmányok* I. Budapest.

Gósy, M. (2004) *Fonetika, a beszédtudománya.* Budapest: Osiris Kiadó.

Gráczi, T. E. & Horváth, V. (2010) A magánhangzók realizációja spontán beszédben. *Beszédkutatás*, 5–16.

Wednesday

May 17th, 2017

Neurolinguistic aspects of speech processing
Preliminary speech by

Vesna Mildner
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Speech is considered to be one of the uniquely human faculties and it is therefore natural that it has been the topic of interest throughout millennia. The study of neurolinguistic aspects of speech (and language) dates back to about 3000 years B. C., when the first mention of brain in relation to speech was found in old Egyptian scrolls. For the longest time the insights into how brain processes speech have been gained from clinical population, by observing speech and language behavior of individuals with diagnosed neurologic impairments and/or by examining *post mortem* data and relating them to available information on speech functioning. It is only in the past several decades, with the advent of new technologies, that it has become possible (and ethical) to study speech in healthy individuals not only indirectly, by behavioral methods, but increasingly more ‘objectively’ and in real time. Techniques such as positron emission tomography (PET), functional magnetic resonance imaging (fMRI), event related potentials (ERP), to name just a few, have enabled researchers to peak into the intact brain as it performs various speech and language tasks. This has lead to re-evaluation of the existing theories and models of speech processing as well as to emergence of new ones.

The presentation will address neurolinguistic aspects of speech processing in general and also try to shed some light on special contexts: developmental, bilingual and disordered. Developmental issues will focus on plasticity and milestones in speech development, including consequences of early deprivation (sensory or otherwise). Studies of bilinguals have given us a wealth of information about speech and language processing in monolinguals as well. The focus here will be on the differences in ‘brain organization’ between monolinguals and bilinguals, as well as on the differences in the representation of bilinguals’ two languages. Finally, the section on disordered speech will, hopefully, complete the picture by showing the relationship between neurologic disorders and/or trauma and speech behavior.

Effects of gemination on the duration and formant frequencies of adjacent vowels in Hungarian voiceless stops

Tilda Neuberger and András Beke
Research Institute for Linguistics, HAS, Hungary

Gemination has been examined in various languages which have contrastive singleton and geminate consonants (e.g., Ham 2001; Ridouane 2007). Numerous production studies have focused on timing properties of geminates. Their findings confirmed that duration is the main acoustic cue for the distinction between singleton and geminate consonants (Ham 2001; Khattab 2007; Ridouane 2007). Closure duration is this primary cue in the case of stop consonants. However, other factors, such as intensity, spectral moments of the burst release etc. may contribute to this opposition (Local–Simpson 1999; Payne 2006; Idemaru–Guion 2008). A few studies have sought evidence for other acoustic correlates of geminates, which concern other units of speech than the target consonants (the given single or geminate consonant). Previous and following segments have also been analysed in terms of gemination. For example, gemination appears to affect the duration and quality of preceding segments in Malayalam (Local–Simpson 1999) and in Tashlhiyt Berber (Ridouane 2007).

The aim of the present study is to analyse whether gemination has an effect on the duration and formant frequencies of the adjacent vowels in Hungarian stops. Two hypotheses were addressed: 1. Vowels would be realized with shorter duration preceding geminate stops than preceding single stops. 2. Vowels would also differ regarding formant frequencies depending on singleton or geminate environment.

Ten Hungarian-speaking male subjects (aged between 20 and 29) were asked to participate in spontaneous conversation about their work and hobbies. The participants have no reported history of speech disorders. Preceding and following vowels in the environment of intervocalic singleton and geminate stops ([p, t, k] and their long counterparts) were annotated manually in Praat software (Boersma–Weenink 2013). Segment boundaries of vowels were marked at the onset and the offset of the second formant of the vowels. Duration measurements were obtained from simultaneous spectrographic and waveform displays. Formant frequencies of surrounding vowel (F1 and F2) were extracted using a script written by Morrison and Nearey (2011) in MATLAB environment. To measure formant frequency 25 ms length Hamming type window and 10 ms time step was used based on LPC analysis. In case of the preceding vowel the formants were measured at the midpoint and offset; and in case of following vowel the formants were measured at the onset and midpoint. R (R Core Team 2012) and MCMCglmm (Hadfield 2010) were applied to perform a generalized linear mixed effects analysis of the relationship between the acoustic features of the vowels and the categories of consonant quantity. As fixed effects, we entered 'consonant quantity' into the model. We present confidence intervals estimated with the Markov Chain Monte Carlo method and p-values that are considered significant at the $\alpha = 0.05$ level.

Out preliminary results showed that gemination shortens vowels preceding voiceless stops. We also found minor formant differences between the adjacent vowels of singletons and geminates. Acoustic correlates of the stop length distinction may play an important role in the perceptual distinction of the contrasting sounds. Durational and formant differences between vowels surrounding singletons and geminates may help listeners in perceptual discrimination of the two phonemic categories. Results of this study may contribute to various speech applications and second language learning.

References

Boersma, P., Weenink, D. 2013. *Praat: doing phonetics by computer* [Computer program]. 5.3. http://www.praat.org/

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, 33(2), 1–22.

Ham, W. 2001. *Phonetic and phonological aspects of geminate timing*. New York: Routledge.

Idemaru, K., Guion, S. G. 2008. Acoustic covariants of length contrast in Japanese stops. *J. Inter. Phon. Assoc.* 38(2), 167–186.

Khattab, G. 2007. A phonetic study of gemination in Lebanese Arabic. *Proc. 16th ICPhS*, Saarbrucken, Germany, 153–158.

Local, J., Simpson, A. 1999, Phonetic implementation of geminates in Malayalam nouns. *Proc. 14th ICPhS*, San Francisco, 595–598.

Morrison, G. S., Nearey, T. M. 2011. *FormantMeasurer: Software for efficient humansupervised measurement of formant trajectories*. (2011). [Software release 2011-05-26].

Payne, E. 2006. Non-durational indices of Italian geminate consonants. *J. Inter. Phon. Assoc.* 36(1), 83–95.

R Core Team (2012). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ridouane, R. 2007. Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *J. Inter. Phon. Assoc.* 37(2), 119–142.

Speaker age estimation by musicians and nonmusicians

Ákos Gocsál
Faculty of Music and Visual Arts, University of Pécs, Hungary

Speaker age is one of the most widely researched human attributes in the context of social perception based on the speaker's voice. Research in this field started in the 1960s and, although some differences were found, the first findings of Hollien and Shipp in 1969 (cited by Huntley, Hollien and Shipp, 1987), have been generally confirmed by many others stating that listeners judge younger adults older than their actual age, while older speakers are usually believed to be younger than their calendar age (e.g. Hughes and Rhodes, 2010; Skoog Waller et al., 2012). Speaker age estimation is not only a function of the speaker's voice parameters. Huntley, Hollien and Shipp (1987) proposed that in addition to individual differences between speakers, some listener attributes seem to influence age judgments, such as the listener's own age. A review by Moyse (2014) also concluded that younger listeners are more accurate than the older ones, and, some gender differences were also found, with women being more accurate in age judgments.

The main focus of this presentation is another aspect that may differentiate listeners in speaker age estimation. In the past decade, there has been an increasing interest in researching differences in sound perception abilities that exist between musicians and non-musicians. Patel's expanded OPERA hypothesis (O = overlap between neural networks processing speech and music, P = higher precision of processing is demanded when music is heard, E = emotion, R = repetition of musical activities, A = focused attention demanded by music), published in 2014, proposes that when music and speech share auditory perceptual mechanisms, and music places higher demands on those auditory mechanisms than speech does, speech processing may be enhanced in musicians (Patel, 2014). Research results have proven that musicians are better at pitch perception (Alexander, Wong and Bradlow, 2005), frequency discrimination (Barrett et al., 2013), voice timbre processing (Chartrand–Belin, 2006) and in other aspects of auditory processing related to speech perception.

Differences in the perception of nonverbal contents of speech by musicians and non-musicians are, however, still largely unknown. To the best knowledge of the author, it is only the perception of emotions that has been researched in this respect, verifying a robust difference between the two groups, musicians being significantly better at identifying emotions from speech (Lima–São Luís, 2011).

The proposed presentation attempts to find similar differences in speaker age estimation. It is hypothesized that musicians' age estimations are more accurate than that of non-musicians, and musicians' age estimations are more coherent, i.e. individual differences in age estimations are smaller than in the non-musician's group. For the experiments, 24 spontaneous speech samples (males, 20-73 years of age) selected from the BEA database (Gósy, 2011) are used as acoustic stimuli. One group of subjects comprises students of music with at least 8 years of classical music training (i.e. playing an instrument), and the other group consists of students of other fields with no previous musical training at any level. After listening to each sound sample, subjects are asked to estimate the age of the speaker in years.

Since listening experiments are in process at the time of abstract submission, only preliminary results are available, mainly because of the limited number of non-musician subjects included so far. Preliminary results with 38 musicians and 11 non-musicians do not support that the musician subjects' age estimations differ significantly from those of non-musicians, however, in most of the cases – but not always –, boxplots representing the musicians' age estimations present narrower ranges than those of non-musicians. If results with the inclusion of more non-musicians remain the same, it will suggest that musical expertise does not necessarily improve age estimations in general, however, may help avoid "extreme" estimations. Another aspect of the phenomenon, i.e. the possible role of f0 and speech rate in age estimation in both groups is also discussed.

References

Alexander, Jennifer A. – Wong, Patrick CM – Bradlow, Ann R. (2005) *Proceedings of the Interspeech 2005 Conference*. Lisbon, Portugal. 397–400.

Barrett, Karen Chan – Ashley, Richard – Strait, Dana L. – Krais, Nina (2013) Art and science: how musical training shapes the brain. *Frontiers in Psychogy* 4. 713. doi:10.3389/fpsyg.2013.00713

Chartrand, Jean-Pierre – Belin, Pascal (2006) Superior timbre processing in musicians. *Neuroscience Letters* 405. 164–167.

Gósy, Mária (2011) BEA – A multifunctional Hungarian spoken language database. *The Phonetician* 105. 50–61.

Hughes, Susan M. – Rhodes, Bradley C. (2010.) Making age assesments based on voice: the impact of the reproductive viability of the speaker. *Journal of Social, Evolutionary, and Cultural Psychology* 4/4. 290–304.

Huntley, Ruth – Hollien, Harry – Shipp, Thomas 1987. Influences of listener characteristics on perceived age estimations. *Journal of Voice* 1/1. 49–52.

Lima, César F. – Castro, São Luís (2011) Speaking to the trained ear: Musical expertise enhances the recognition of emotions in speech prosody. *Emotion* 11(5), 1021–1031.

Patel, Aniruddh D. (2014): Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research* 308. 98–108.

Skoog Waller, Sara – Eriksson, Mårten – Sörquist, Patrik (2015.) Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology* 6. 978.

Comparison of F0 measures for male speakers of Croatian, Serbian and Slovenian

Gordana Varosanec–Skaric, Zdravka Biocina and Gabrijela Kisicek
Department of Phonetics, Faculty of Humanities and Social Sciences,
University of Zagreb, Croatia

Introduction

Fundamental frequency is, among other voice qualities, important in speaker identification (Traunmüller and Eriksson, 1995), especially because the F0 values are preserved in different recording conditions. An average F0 value for 104 male speakers of Croatian language is stated based on continuous neutral speech, with a duration of around 70 s and it is 117.36 Hz (setting between 75 and 176 Hz; S.D. 18.78 Hz) (Varošanec-Škarić, 1998); for male speakers of Serbian language Jovičić (1999) reports average value of 121 Hz and S.D. F0 cca 16 Hz, measured for five standard vowels of Serbian language. Tivadar (2008) reported F0 values in Slovenian which were lower than values measured in Croatian language (Varošanec-Škarić, 1998). However, Slovenian results were based on 6 elite vocal professionals.

Our goal was not to question previously measured values, but to measure and to compare F0 values on symmetrical groups. These findings are important in sociolinguistics and forensic phonetics.

The aim of this paper was to compare fundamental frequency measures expressed in Hz and in semitones which can be used for more robust, future speaker comparison. Second, pragmatic aim of this paper was to collect samples of voices in order to create Croatian database which could be used in future for comparison based on sex, age, education, voice health between speakers of very similar languages, e.g. Croatian, Serbian, Bosnian, Slovenian and for comparison between Croatian speakers with received and dialectal pronunciation.

It was expected that the younger groups of Croatian, Serbian and Slovenian speakers will not differ significantly for average value mean F0.

Method

In the first part of the research native speakers of Croatian (N=15), Serbian (N=15), and Slovenian (N=15) language were recorded under the same conditions during the 2015, 2016 and 2017 year in three country capitals: Zagreb, Belgrade and Ljubljana. Age of recorded speakers was similar with age median of 22.

Considering that the results of the first part of the research between three symmetrical groups of speakers have showed no statistically significant difference of mean F0 measures and that the values of mean F0 between Croatian and Slovenian language were similar, the second part of the research was conducted.

The second part of the research included larger number of speakers in groups which showed differences in F0 measures i.e. Croatian (N=37) and Serbian (N=37) speakers. Spontaneous speech of every speaker was recorded (cca 3 min) and reading passages were used for fundamental frequency measures in Hz and semitones.

F0 mean, median, average baseline value Fb (based on median), average minimal and maximal F0 values, alternative baseline (Alt_Fb), standard deviation (S. D.) of F0 in both Hz and semitones were calculated. For the calculation special programs in Praat (Boersma and Weenink, 2015; Ver. 6.0.05) were used. The calculation was done on a limited setting between 65 and 300 Hz to get all the possible frequencies for the F0 range and to avoid octave jumps. Lindh (2006) used the setting between 75 and 350 Hz for the F0 tracker. Considering that the sample was comprised of only male speakers of received Croatian and Serbian without noticeable dialectal pronunciation, it has been shown that it is enough to limit the pitch ceiling to 300 Hz. The values above 300 Hz have been checked in preliminary testing and it has been shown that they are discontinuous octave jumps, harmonics and stridents. ANOVA: two-factor with replication was used to test the differences between groups and to test the difference between reading passage and spontaneous spoken utterance.

Results and discussion

In the first part of the research based on smaller number of speakers of Croatian, Serbian and Slovenian language there has been no statistically significant difference in mean F0 in Hz (Croatian 118.21 Hz; Serbian 123.70 Hz and Slovenian 119.13 Hz). Other variables of F0 were also not statistically relevant. The second part of the research which was conducted between groups that showed more differences i.e. Croatian and Serbian speakers ANOVA (single factor and two factor) analysis was used.

Results have showed that respectively the greatest difference between groups was in median F0 in semitones (p=0.0005; Figure 1) and in Hz (Croatian 117.25 Hz, Serbian 124.09 Hz; p=0.0007). It has been shown that younger Croatian speakers have significantly lower F0 mean in both measurements than Serbian (Croatian 117.11 Hz, Serbian 126.64 Hz; p=0.001), significantly lower Fb (Croatian 90.86 Hz, Serbian 97.30 Hz; in Hz: p=0.001; in semitones: 0.002) and lower Alt_Fb. The overall results for both groups of speakers show less significant differences between reading passage and spontaneous spoken utterance for measures in semitones and in Hz for Fb (p=0.017; 0.015), respectively, median F0 (p=0.014; 0.018), for F0 minimums (in Hz, p=0.016; 0.022), for F0 mean (p=0.02; 0.024), for F0 maximums (Croatian 198 Hz, Serbian 201 Hz; p=0.04) and in semitones for Alt_Fb (p=0.04). The values of F0 mean from this paper correspond with the value reported in Varošanec-Škarić (1998) for Croatian speakers, and for Serbian speakers it is somewhat higher than the F0 value reported in Jovičić (1999). It can be concluded from the collected data that intonation patterns are more different between groups of Croatian and Serbian speakers which is important for further research in the two similar languages. Overall, it can be concluded that clearer differences were found in the semitone measures, but both type of measurements, Hz and semitones, are useful for comparing speakers with different mother tongues. In wider context these results are similar and can be compared with the results in European languages (Traunmuller and Eriksson, 1995b) and the results of the research conducted aiming to compare Slavic and German languages (Andreevna et al, 2014). Results are also similar with the research in Swedish (Lindh, 2006) and Czech language (Skarnitzl and Vaňková, 2016).

Conclusion

Based on overall results we can conclude that Croatian and Serbian speakers are more different than Croatian and Slovenian based on measures of F0. Which is surprising because RP Croatian and Slovenian languages are more different linguistically. This can be explained with more similarities in intonation between these two languages. However, this should be confirmed with further research planned for the future including greater number of native speakers of Croatian and Slovenian language. These results are valuable also in the sense of the methodology because they indicated how important it is to have larger number of homogeneous groups (higher that 30) which can provide more sensitive intergroup comparisons of acoustical measures in spontaneous speech.

Acknowledgments

We wish to thank colleagues S. Jovičić for the help organizing recording in Belgrade and H. Tivadar for recordings in Ljubljana. We also wish to thank colleagues from the Department of Phonetics J. Bičanić on expert help in statistics and I. Bašić for the help in recordings in Ljubljana.

This research is a part of the project financed by the University of Zagreb: Forensic Phonetics: Measurements of fundamental frequencies (F0) of male speakers (HP044).

References

Andreeva, B., Demenko, G., Möbius, B., Zimmerer, F., Jügler, J.,Oleskowicz-Popiel, M., (2014).: Differences of Pitch Profiles in Germanic and Slavic Languages. In *Proceedings of the Interspeech*, 1307–1311.

Boersma, P. and Weenink, D. (2015). *Praat: doing phonetics by computer* (Version 6.0.05).

Jovičić, S. T. (1999). *Govorna komunikacija*. Beograd: Nauka.

Lindh, J. (2006). Preliminary F0 statistics and forensic phonetics. *Proceedings, IAFPA* ’06, Göteborg University. (http://www.ling.gu.se/konferenser/iafpa2006/Abstracts/Lindh__IAFPA2006.pdf).

Skarnitzl, R. and Vaňková, J. (2016). Population statistics of Common Czech: F0 in multi-style tasks and voice disguise strategies. *IAFPA 25th conference*, York, Uk, pp 128–128.

Tivadar, H. (2008). Kakovost in trajanje samoglasnikov v govorjenem knjižnem jeziku. Unpublished Doctoral Dissertation: Ljubljana.

Traunmöller, H. and Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. (http://www.ling.su.se/staff/hartmut/aktupub.htm).

Traunmöller, H. and Eriksson, A. (1995b). The perceptual evaluation of F0 excursions in speech as evidenced in liveliness estimations. *JSDS 97*, 1905–1915.

Varošanec-Škarić, G. (1998). *Zvučne osobine ugode glasa*. Zagreb: University of Zagreb, Unpublished Dissertation.

On some linguistic properties of spoken Hungarian based on the HuComTech corpus

László Hunyadi
Dept of General and Applied Linguistics, University of Debrecen, Hungary

The HuComTech corpus is the first multimodal corpus of verbal and nonverbal behaviour based on Hungarian. Its building started in 2009 and has by now been virtually completed. It includes about 50 hours of interactions of formal and informal dialogues between two agents and 110 subjects. The formal dialogues (average duration: 8 minutes) have the form of a job interview, the informal dialogues (average duration: 18 minutes) are actually guided interactions on a variety of everyday topics. The original aim of the corpus was to study and learn those verbal and nonverbal aspects of interactions which can be formalised enough to be implemented in human-machine interfaces. It is based on a generative model of multimodal interactions with the essential goal to capture perception/analysis and production/synthesis in a unified system that can both be instrumental in building such interfaces and modelling human cognition.

Annotations are done at multiple levels most of them with a time resolution of 300 ms. As for the medium to be annotated, there are three kinds: video only, audio only, video+audio. Annotation levels include those requiring the description of some physical, measurable properties as well as those requiring the interpretation of observable events. As for video, annotations include physical properties such as gaze, head movement, hand shape, posture, touch motion, interpretations include perceived emotions. As for audio, there are annotations of various aspects of prosody (absolute values of and stylised F0 and intensity), speech rate and silence as well as the interpretations of perceived emotions. Also on the audio level, force alignment of each running word (their beginning and end) is now being implemented using the Webmaus service and manually adjustment. The combined video+audio medium is mainly used for a wide range of pragmatic and turn management annotations.

Special effort has been made to provide sufficient material for traditional linguistic purposes as well. Accordingly, the corpus includes full morphological annotation (using automatic parsing), an automatic shallow syntactic parsing and, manually, a special annotation for syntactic incompleteness, a well known property of spoken language.

At present the size of the corpus is approaching 2 million single annotations. In order to retrieve information from this vast body of data a database has been implemented. The format is .eaf and it can be searched using the freely available ELAN tool. The database is designed to be accessible remotely from two web sites (The Language Archive, Nijmegen, and RIL, Budapest); until all work has been completed, only a subset of the data I can be reached publicly.

The challenge of studying multimodal human behaviour is that virtually none of the stereotypical primitives of an event of any kind is obligatory and – in sharp contrast to syntax or morphology – the primitives constituting an event may or may not follow adjacency so that a chain of them may also include “noise”, i.e. irrelevant to the event data. In order to cope with this seemingly discouraging situation we are applying a special research environment with a specific methodology designed to understand this complex nature of social interaction. The talk will present some research data and results based on pattern recognition using the Theme software. We will highlight verbal and nonverbal behavioural patterns of spoken dialogues drawing examples from syntax organisation as well as prosodic patterning as a function of certain thematic and pragmatic properties, all reflecting the variability of spoken interactions.

Stem and suffix durations in words of increasing length in children's spontaneous utterances

Valéria Krepsz and Mária Gósy
Research Institute for Linguistics, HAS, Hungary

Introduction

Duration is a physical parameter of spoken language in that words exist in time. A great many studies discussed various factors that influence the duration of a word and the variability therein (e.g., Losiewicz 1995; Bell et al. 2009) including the effect of the increasing number of syllables within a word on the duration of the syllables. Individual syllables in a word become shorter as word length increases (e.g., Lehiste 1972; Bell et al. 2002; Tily et al. 2009). Age is also acknowledged to be of considerable importance when spoken words are considered. Young children were reported to show longer durations for words than groups of adults (e.g., Smith 1992) that is explained by their underdeveloped speech motor control. For words, produced in spontaneous utterances during language acquisition, the temporal interrelations of stems and suffixes might carry cue information in an agglutinating language about the children' lexical access and speech planning process of the articulation of lexemes.

In this study we seek to explore the internal temporal patterns of the words of various lengths across several ages (in terms of a cross-linguistic analysis). The core question of the study is whether there is a morphologically conditioned shortening of stems and suffixes across the increasing number of stem syllables, on the one hand, and whether this shortening phenomenon exists across various ages, on the other. Our current hypotheses are that (i) reduction of stems and suffixed words will occur as word length increases (equalization tendency), (ii) this reduction will take place after the age of 7, (iii) suffixes will not show durational changes irrespective of word length and age.

Methodology

Thirty Hungarian-speaking children participated in the study forming three age groups (mean ages: 5, 7, and 14 years). Each group consisted of 10 speakers (with an equal number of males and females). More than 8 hours of spontaneous speech material was carefully hand-labeled using Praat (Boersma–Weenink 2014). Stems, suffixes and suffixed words were marked by one of the authors while the other author checked each word (with an agreement ratio of 98%). The word boundaries (between acoustically distinct regions in the signal) were identified in the waveform signal and spectrogram display via continuous listening to the words.

Suffixed verbs and nouns with similar distribution (about 7,000 items) were selected according to the following criteria: (i) stems consisted of various numbers of syllables from 1 to 4, together with suffix syllables to 2 to 5, (ii) five frequent monosyllabic suffixes (*-ban*/*-ben* 'in', *-nak*/*-nek* 'for', *-val*/*-vel* 'with', *-tam*/*-tem* '1sg past', and *-nak*/*-nek* '3pl' were selected that indicated grammatical relationships, (iii) all suffixes were the last syllables of the words, (iv) all words occurred in the middle of phrases (in order to avoid phrase-final lengthening), (v) the suffixes occurred in similar ratios across stems and speakers. Durations of both the stems and suffixes were taken by measuring the phase between the onsets and the offsets of the stems, the suffixes and the whole words according to common acoustic-phonetic procedures. A specific script was written to obtain the values automatically. All data were normalized across speakers using the z-normalization method in order to avoid speech rate differences. To test statistical significance, linear regression analysis, repeated measures ANOVA and the Mann–Whitney test were used, as appropriate (using SPSS 19.0 version). Measured durations of stems, suffixes, and suffixed words were dependent variables while number of syllables of the stems and age were the independent factors. The confidence level was set at the conventional 95%.

Results

As expected, the longest articulation of the suffixed words irrespective of word length were found with the five-year-olds followed by the older children while the fastest articulation was produced by the 14-year-olds. Duration of the suffixed words showed significant differences depending on both word length and age. The increasing number of syllables in the words had an effect on the duration of the syllables. Statistical analysis confirmed significant differences in the reduction of words between 14-year-olds and both groups of younger children. Five-year-olds produced words containing more than two syllables substantially longer than all the other participants. Five- and seven-year-olds articulated suffixed words with practically no reduction.

Durations of the stems also showed significant differences depending on both the number of syllables and age. As the length of stems increases the durational differences in the production of the stems become significantly different across ages. The largest differences in the durations of both stems and suffixed words were found in those containing 5 syllables.

Suffixes had an average duration of about 380 ms across ages. Significant differences were found depending on age; however, groups of younger children showed similar values. The length of stems did not have any significant effect on suffix durations.

Conclusions

This study revealed significant differences in the durations of stems and suffixed words depending both on word length and age. The equalization tendency of the duration of the syllables, however, was characteristic only of 14-year-olds. Immature speech motor control might explain the lack of syllable reduction of words in five- and seven-year-olds. However, speech motor control does not seem to explain the relatively stable durations of suffixes across the analyzed age groups. Slightly variable duration in suffixes may be supposed to be a consequence of the agglutinative character of Hungarian, in which coordination of stems and suffixes takes place relatively early in the process of language acquisition. Finally, we conclude that word duration is influenced by stem length, and reduction tendency in word production does not apply to all analyzed age groups.

References

Bell, A. Gregory, M. L., Brenier, J. M., Jurafsky, D., Ikeno, A. & Griand, C. 2002. Which predictability measures affect content word durations? *Proc. of the ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language* Technology, 1–5.

Bell, A., Brenier, J. Gregory, M., Girand, C. & Jurafsky, D. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60, 92–111.

Boersma, P. & Weenink, D. 2014. *Praat: doing phonetics by computer.* http://www.fon.hum.uva.nl/praat/download_win.html (accessed 18 November 2014)

Lehiste, I. 1972. The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America* 51(6B), 2018–2024.

Losiewicz, B. L. 1995. Word frequency effects on the acoustic duration of morphemes. *Journal of the Acoustical Society of America* 97, 3243.

Smith, B. L. 1992. Relationships between duration and temporal variability in children's speech. *Journal of the Acoustical Society of America* 92, 2165–2174.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A. & Bresnan, J. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* 1(2), 147–165.

Challenges in automatic annotation and the perception of prosody in spontaneous speech

István Szekrényes
Universtiy of Debrecen, Hungary

Introduction

The proposal aims at demonstrating a rule-based annotation method and an experimental approach for the automatic analysis and the perceptual aspect of speech prosody. The development is inspired by the work of Piet Mertens [5] using its objectives and the psychoacoustic model of tonal perception [2]. The main purpose of our implementation is to extend the *HuComTech* Hungarian multimodal corpus[1] with prosodic labels for the analysis of non-verbal behaviour in human-human interactions [3]. The most recent repairs and adaptations - resulting in the current exibility of the program - are made in collaboration with the *SegOrg* project[2] during the analysis of various kinds of recordings (with 2-14 speakers) from the FOLK [6] German corpus. Details of the algorithm under the name *ProsoTool*[3] are described in [8] [7]. The same algorithm as part of the e-magyar project [4] referring to that project is mentioned as *emPros*[4].

Methodology

The algorithm is implemented as a Praat[1] script including a speaker isolation and an intonation processing module which stylizes and categorizes F0

[1] https://hdl.handle.net/1839/00-0000-0000-001A-E17C-1@view

[2] http://www1.ids-mannheim.de/prag/muendlichekorpora/segcor.html

[3] https://github.com/szekrenyesi/prosotool

[4] http://e-magyar.hu/hu/speechmodules/empros

curves as perceptually relevant melodic sequences labelling the shape (rise, fall, ascending etc.) and the relative (compared to the individual vocal range) and absolute (in Hertz) position of every movement. The input is a speech sound file (in WAV format) and the acoustic representation of turn-taking (in Praat TextGrid) to isolate the voice segments of the speakers excluding overlapping speeches. Based on the F0 distribution of the isolated segments, the algorithm divides the individual vocal range into five levels (see in Figure 1). F0 smoothing and stylisation are performed in every single speech segments resulting the melodic sequences of intonation as it can be seen in Figure 2. The categorizations (the resulting labels) are based one the global F0 distribution and some parameters (the amplitude and the duration of movements) which are also used in the Tilt intonation model [9].

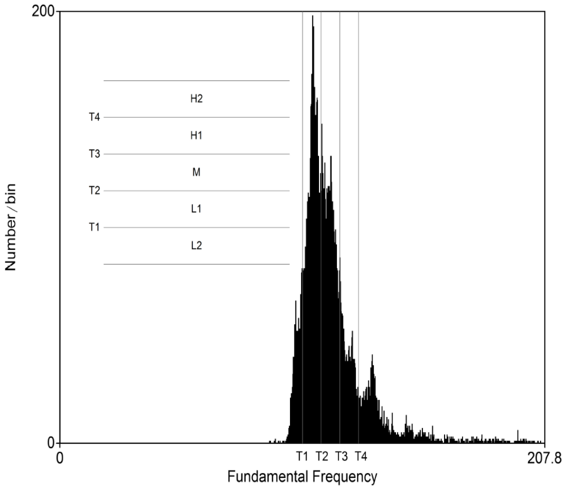


Figure 1: Individual vocal ranges based on F0 distribution

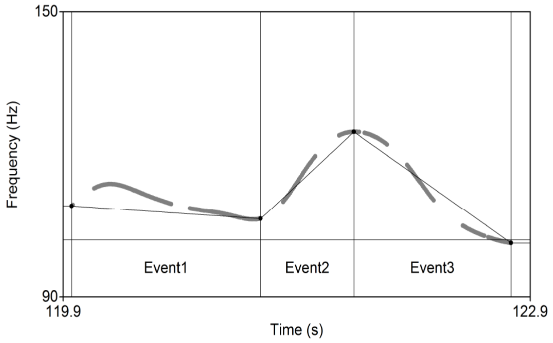


Figure 2: The result of smoothing and stylization

Results

The validation is the most complicated and a still on-going part of the project. In Figure 3, one can see the output (the annotation at the bottom and the measured F0 with transcription above) for an utterance with a prototypical intonation of a Hungarian yes-no question.

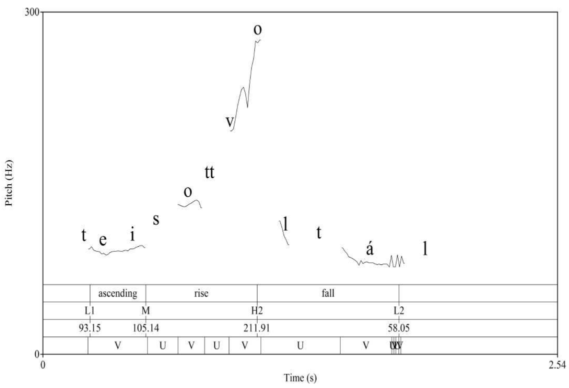


Figure 3: Output for a Hungarian yes-no question: "Te is ott voltál? [Were you there too?]"

In spontaneous speech (see Figure 4), the perceptual alignment of the resulting labels is less evident or verifiable.

For validation, some experiments are also designed to explore the perception of prosody in spontaneous conversations using various conditions (see Figure 5). The results are still in the process of evaluation.

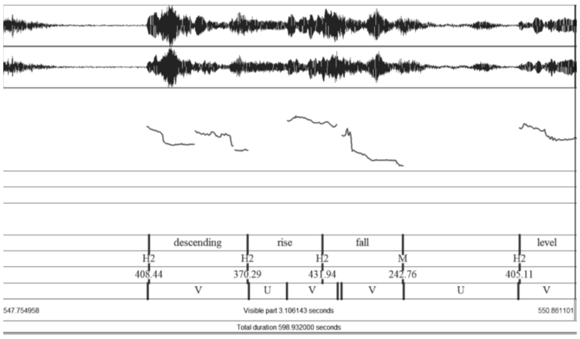


Figure 4: Results for a spontaneous conversation

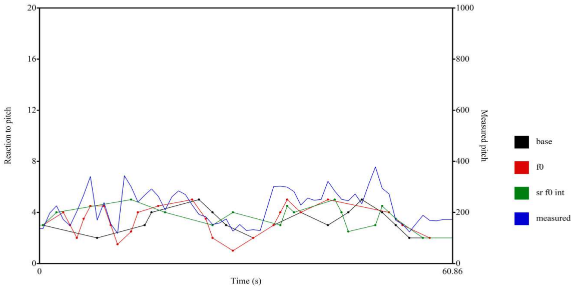


Figure 5: Perception of speech prosody in different conditions

References

[1] David Boersma, Paul & Weenink. *Praat: doing phonetics by computer* [computer program]. version 6.0.22. http://www.praat.org/, 2016. retrieved 15 November 2016.

[2] J. 't Hart. 1976. Psychoacoustic backgrounds of pitch contour stylisation. *IPO-APR*, 11:11–19.

[3] László Hunyadi, András Földesi, István Szekrényes, Alexandra Staudt, Hermina Kiss,

Ágnes Abuczki, and Alexa Bódog. 2012. Az ember-gép kommunikáció elméleti-technológiai modellje és nyelvtechnológiai vonatkozásai. In *Általános Nyelvészeti Tanulmányok XXIV: Nyelvtechnológiai kutatások*, Akadémiai Kiadó, Budapest, 265–309.

[4] András Kornai and István Szekrényes. 2011. e-magyar beszédarchívum. In V. Vincze, editor, *XIII. Magyar Számítógépes Nyelvészeti Konferencia* (MSZNY2017), Szegedi Tudományegyetem Informatikai Tanszékcsoport, 103–109.

[5] Piet Mertens. 2004. *The prosogram: Semi-automatic transcription of prosody based on a tonal perception model.* In Proceedings of Speech Prosody.

[6] Thomas Schmidt. 2016. Good practices in the compilation of folk, the research and teaching corpus of spoken German. In John M. Kirk and Gisle Andersen, editors, *Compilation, transcription, markup and annotation of spoken corpora*, Special Issue of the International Journal of Corpus Linguistics [IJCL 21:3], 396–418.

[7] István Szekrényes. 2015. Prosotool, a method for automatic annotation of fundamental frequency. In 6th *IEEE In- ternational Conference on Cognitive Infocommunications (CogInfoCom)*, New York, IEEE 291–296.

[8] István Szekrényes. 2014. Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. *Journal on Multimodal User Inter- faces*, 8:(2):143–150.

[9] P Taylor. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107(3):1697–1714.

A perceptual comparison: Spontaneous speech of speakers' today and 40 years ago

Anita Auszmann
Research Institute for Linguistics, HAS, Hungary

Linguistic investigations periodically focus on changes in the language that reflect changes in both the society and the speakers (e.g. Labov 2006). As the society, language itself is changing continuously (Wardhaugh 1995, Labov 2006). The pronunciation of speech sounds is changing both in terms of an individual's life and across speaker generations. The perception of these small differences in articulation goes unnoticed because the listener constantly adapts to the modifications (Ohala 2012). It is often heard subjective observations that people long ago spoke differently than today: with slower tempo, less disfluencies, more accurate articulation. Our previous experiments – carried out on the same database – did not substantiate a significant change in articulation and speech rate through the decades (Auszmann 2016a). However we found a change in the way and the strategies of expressing thoughts (the duration and length of utterances, the duration and occurrences of pauses, proportion of disfluencies) (Auszmann 2015; Auszmann 2016b).

In the present research we would like to make the previous results more detailed with a perception test. The aim of the investigation is to found out whether university students are able to distinguish the spontaneous speech of speakers' today and 40 years ago, and if so, along what kind of criteria.

For the compilation of the perception test we used the same samples (12 male speakers) as in previous research and we set them in pairs. Half of the recordings were selected from BEA (Hungarian Spoken Language Database) database (Gósy 2012; Neuberger et al. 2014) and the other half from the 'Szalag' corpus. The test contains some coherent sentences from all of the speakers. No fillers was added to the speech material. In the perception test university students will participate who are studying phonetics (minimum 50 people). The protocol of the test will be the following: after a short training period in the first part of the test participats have to compare two tokens (one from BEA database – one from 'Szalag' corpus), in the second part of the test participats have to make decisions for each tested speech token (played randomly) whether the speaker lives today or lived 40 years ago. We would like to investigate the reaction time of the participants and ask them to clarify their choice with some words (professional commentary).

We assume in the light of the results of research carried out earlier that the separation of the two speaker groups will not be clear perceptually, but owing to the phonetics experience students will be successful in the task in a certain degree.

References

Auszmann Anita 2015. A spontán beszéd időviszonyai 40 évvel ezelőtti és mai beszélőknél. [The temporal pattern of spontaneous speech today and 40 years ago] In: Gósy Mária (szerk.) *Diszharmóniás jelenségek a beszédben*. Budapest: MTA Nyelvtudományi Intézet. 219¬234.

Auszmann Anita 2016a. *Formant structure and duration of Hungarian vowels as distinctive features in spontaneous speech of speakers' today and 40 years ago*. (Poster presentation at ExAPP conference)

Auszmann Anita 2016b. *Megakadásjelenségek 40 évvel ezelőtti és mai beszélők spontán beszédében*. [Disfluencies in spontaneous speech of speakers' today and 40 years ago]. (Oral presentation at Beszédkutatás 2016 conference)

Gósy, Mária 2012. BEA—A multifunctional Hungarian spoken language database. *The Phonetician* 105(106), 50–61.

Labov, W 2006. A sociolingvistic perspective on sociophonetic research. *Journal of Phonetics* 34. 500-515.

Neuberger, T., Gyarmathy, D., Gráczi, T. E., Horváth, V., Gósy, M., & Beke, A. (2014). Development of a Large Spontaneous Speech Database of Agglutinative Hungarian Language. In *Text, Speech and Dialogue*. Springer International Publishing. 424–431.

Ohala, John J. 2012. The listener as a source of sound change (perception, production, and social factors). In Maria-Josep Solé and Daniel Recasens (eds.): *The initiation of sound chang*e. Amsterdam: John Benjamins. 21–36.

Wardhaugh, Ronald 1995. *Szociolingvisztika*. Osiris–Századvég, Budapest.

The temporal characteristics of teenagers in the various spontaneous speech genres

Laczkó Mária
Faculty of Paedagogy, University of Kaposvár, Hungary

The speech tempo is one of the most explored areas of spontaneous speech research. The researchers refer to speech rate as a measure of the entire cognitive and articulatory activity involved in the production of an utterance, and the articulation rate for the amount of speech produced in the time actually taken to articulate it. In other words the tempo of speech is the rate at which utterances and their smaller units are pronounced. Consequently tempo of speaking is usually defined as speaking rate or as articulation rate. Speaking rate as a gross rate refers to the entire speaking phase including pauses versus articulation rate as a net rate refers to phases of articulation excluding pauses.

Both of them are influenced by different factors, regarding the speaker's characteristics like his/her age, gender, individuality, physical/emotional condition or the characteristics of his/her pauses occurring in the spontaneous speech. The topic of the speech, the speaking context and type of the text, consequently the style of the speech can also have an effect on speaking and articulation rates.

The effect of the age on speech tempo has already proved in Hungarian language and the results showed significant differences among the people with various age including the teenagers whose both speech and articulation rate was the highest. There are also some experimental results regarding the different speech tempo values in various types and styles of the texts including the spontaneous speech, interviews, conversations, or reading some kind of text and reading news.

It is still a question what is the speech tempo of teenagers in the various speech genres like and how it is characterized by various types of the pauses and hesitation phenomena in the various communication situations which require different cognitive activities and skills. In other words the question is whether it is possible to claim that their fast speech tempo in Hungarian language can also occur in the various communication situations and to what extent. So how can their speech tempo depend on the type of the text and speaking style. Our previous hypothesis was that the speed of teenagers' speech samples in the given communication situations will be fast too, however it can also be determined by the type and style of the texts: the fastest tempo categories can occur in the narrative speech samples, the shortest tempo categories can describe the processes reading aloud. We have also thought that the tempo categories have close interrelation with speech planning process, and the examined communication situations require different cognitive activities in terms of planning, consequently the differences can also be seen in tempo rates.

In order to answer the questions and to discuss the hypothesis the series of experiments were carried out with the participation of the same aged secondary school children (teenagers). Their mean age was 16,7 years, the students' age was between 16 and 17.

Their speech samples (aproximately 3 minutes per person) were recorded in three communication situations: in narrative, rhetorical speech and in reading aloud. In the first case the students had to speak about the family life after they had 1,5 minutes to think of the given topic. In the second communication situation the teenagers had to memorize their texts written in advance, and they had to tell them by heart to the audience. In this case they previously had to choose one between two kind of topics (the value of family nowadays or the use of Internet versus books) in order to collect their arguments and discuss the topic, so they have enough time for the preparation of their speech. In the case of reading aloud they had to read the short text which was given to them before reading, however they have time for looking at all the text. The number of the students taking part in the experiments was 6-6, and in all three communication situations the same teenagers' speech was digitally recorded.

The analysis was based on the determination of speech rates, articulation rates, types and duration (length) of pauses (silent pauses and hesitation phenomena) and their functions (start/continue the speach, mistake and correction, uncertainty) in each situations in terms of each children. For the acoustic analysis the Praat program was used, while the statistical analysis was done by the SPSS 13.00 version. The tempo categories were measured by the number of sounds per seconds, the duration of pauses was given in milliseconds. Our preliminary hypothesis was proved, as the results show that the style of various examined genres of speech have different effect on temporal parameters of teenagers's speech both regarding the speaking rates, articulation rates and also the frequency, the types, the duration and the function of pauses. The speed of teenagers in terms of different texts was not only fast as it has been proved before but it varied depending on the types and styles of the texts. Both the number and length of the silent pauses and hesitation phenomena were different in the various communication situations and we have also found some differences in terms of the function of hesitation phenomena occurring in the different speaking styles. Our paper is focused on the presentation of our data observed and results, and we also emphasize the consequences both in linguistic and pedagogical aspects.

References

Duez, Danielle 2001. Acoustico-phonetic characteristics of filled pauses in spontaneous French speech: preliminary results. in: DISS'01.41-44. htp//www.isca-speech.org/archive_open/archive_papers/diss_01/dis1_041.pdf.

Gósy Mária – Bóna Judit – Beke András – Horváth Viktória 2013. A kitöltött szünet fonetikai sajátosságai az életkor függvényében. Beszédkutatás.121-1

Laczkó, Mária 2013. e Why do teenagers create hesitation phenomena in their mother tongue and in their foreign language? The International Journal of Assesment and Evaluation 2014. Volume 20. Issue 2. 63-74. ISSN: 2327-7920. Champaign, Illinois, USA. Common Ground Publishing LLC

Vasilescu, Ioana – Adda Docker, Martine – Nemeto, R. 2007. Acoustic and prosodic Characteristic of vocalic hesitations across languages. Scientific Report. 2007.

Watanabe, Michiko – Hirose Keikichi – Den, Yasuharu – Minematsu, Nobuaki 2008. Filled pauses as cues to the complexity of upcoming phrases for native and nonnative listeners. Speech Communication 50. 81-94. Synchronized speech, tongue ultrasound and lip movement video recordings with the “Micro” system

Synchronized speech, tongue ultrasound and lip movement video recordings with the “Micro” system

Tamás Gábor Csapó[1,4], Andrea Deme[2,4], Tekla Etelka Gráczi[3,4], Alexandra Markó[2,4] and Gergely Varjasi[2,4]

[1]Budapest University of Technology and Economics,
[2]Dept. of Phonetics, Eötvös Lorand University,
[3]Research Institute for Linguistics, HAS, Hungary
[4]MTA-ELTE “Momentum” Lingual Articulation Research Group

This demonstration will show the details of the articulatory investigations (tongue ultrasound and lip movement) of the MTA-ELTE “Momentum” Lingual Articulation Research Group, including the technical aspects, applied hardware and software elements, sample recordings, and current and planned research.

Analysis of the shapes and dynamics of the human tongue during speech is important for modeling the vocal tract. Ultrasound imaging of the tongue is an attractive solution because it images tongue motion at a rapid frame rate (up to 100 Hz), which can capture subtle and swift movements during speech production (Stone, 2005). Csapó et al. (2017) have demonstrated the type and quality of the first speech and ultrasound recordings with the "Micro" (previously SonoSpeech) system (Articulate Instruments Ltd.). In the current demonstration, we will extend this by showing how to record video of the lips in synchrony with the ultrasound and speech signals.

Five Hungarian subjects (three females and two males) with normal speaking abilities were recorded while reading aloud sentences and nonsense words. The tongue movement was recorded in midsagittal orientation using a "Micro" ultrasound system (Articulate Instruments Ltd.) with a 2-4 MHz / 64 element 20mm radius convex ultrasound transducer at 80-100 fps. During the recordings, the transducer was fixed using an ultrasound stabilization headset (Articulate Instruments Ltd.). The video of the lips was recorded at 59.94 fps (interlaced) either from front or from side view with an NTSC microcamera that was attached to the helmet (see Fig. 1). The video was digitized using a DFG2USB device. The speech was recorded with an Audio-Technica - ATR 3350 omnidirectional condenser microphone that was clipped approximately 20cm from the lips. The ultrasound and the audio signals were synchronized applying the frame synchronization output of the equipment with the Articulate Assistant Advanced software (Articulate Instruments Ltd.). The lip video and the audio signals were synchronized using a SyncBrightUp Unit (Articulate Instruments Ltd.) which adds a white mark to several video frames at the same time as putting a trigger signal to the audio. Both the microphone signal and the ultrasound synchronization signals were digitized using an M-Audio – MTRACK PLUS external sound card at 22050 Hz sampling frequency.

After the recordings, the ultrasound frames were extracted as raw scan line data and converted to JPG images. Next, videos were constructed from the raw ultrasound data, lip movement and synchronized speech recordings (for sample images, see Fig. 2).

The demonstration will include a general introduction of the hardware and software components of the "Micro" system, and sample tongue ultrasound and lip movement videos from the five speakers.

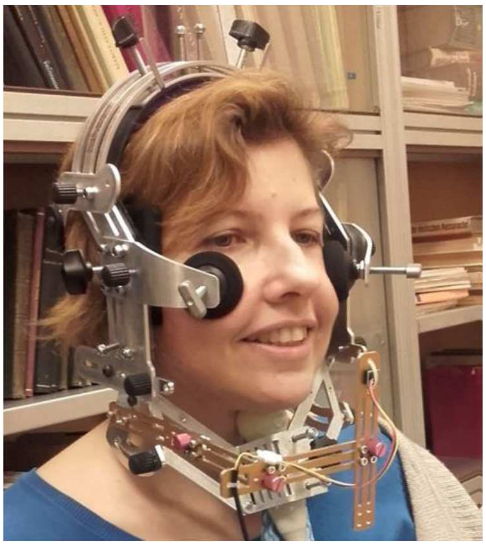


Figure 1: Ultrasound stabilization helmet with the fixed ultrasound transducer and lip camera (in the front of the lips)

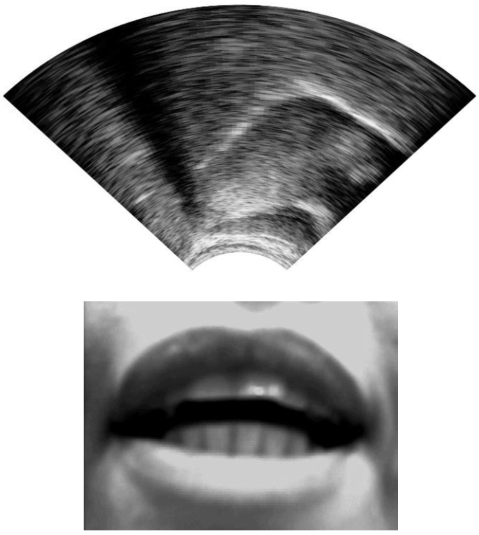


Figure 2: Sample ultrasound and lip images from a female speaker

References

Csapó T. G., Deme A., Gráczi T. E., Markó A., Varjasi G., 2017. *Szinkronizált beszéd- és nyelvultrahang-felvételek a SonoSpeech rendszerrel* [Synchronized speech and ultrasound recordings with the SonoSpeech system], In: XIII. Magyar Számítógépes Nyelvészeti Konferencia [13th Conference on Hungarian Computational Linguistics] (MSZNY2017), Szeged, Hungary, 339–346.

Stone, M. 2005. A guide to analysing tongue motion from ultrasound images. *Clin. Linguist. Phon.* 19, 455–501.

Notes