Speech Research Conference

ELTE Research Centre for Linguistics

Budapest, 9–10 October 2025

Beszédkutatás konferencia

ELTE Nyelvtudományi Kutatóközpont

Budapest, 2025. október 9–10.

Editors/Szerkesztők:

Tekla Etelka Gráczi, Kornélia Juhász, Anna Kohári,

Katalin Mády

ELTE Research Centre for Linguistics

ELTE Nyelvtudományi Kutatóközpont

Budapest

2025

Scientific committee/Tudományos bizottság

Balogné Bérces, Katalin
Bučar Shigemori, Lia Saki
Bárkányi, Zsuzsanna
Cser, András
Deme, Andrea
Fejes, László
Gocsál, Ákos
Hunyadi, László
Huszthy, Bálint
Jordanidisz, Ágnes
G. Kiss, Zoltán
Kiss, Angelika
Liker, Marko
Markó, Alexandra
Sparvoli, Carlotta
Svindt, Veronika
Szekrényes, István
Sztahó, Dávid
Tar, Éva
Zainkó, Csaba

Organisers/Szervezők

Gráczi, Tekla Etelka
Juhász, Kornélia
Kohári, Anna
Mády, Katalin
Mihajlik, Péter
Pirsel, Katalin

DOI: 10.18135/BeszKutKonf.2025
ISBN: 978-615-6678-12-6

Contents

Tartalomjegyzék

Invited talks/Meghívott előadások

Conversational entrainment and its challenges

Štefan Beňuš[1,2]
[1] Constantine the Philosopher University in Nitra, Slovakia
[2] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia

Speech entrainment refers to the tendency of interlocutors to adapt to one another in various aspects of their verbal and non-verbal behavior. In various disciplines, this phenomenon has been also described using terms such as alignment, accommodation, synchrony, or convergence. In this talk, the primary focus will be on the acoustic–prosodic dimensions of this behavior as they manifest in spoken communication. Previous research indicates that speech entrainment is strongly associated with productive, engaging, and stimulating interactions. It also correlates with several socially desirable outcomes, including success in collaborative tasks, smoother and more positively perceived conversations, the expression of romantic interest or physical attractiveness, heightened trust, stronger rapport, and other markers of social connection.

Furthermore, understanding and successfully modeling entrainment opens a wide range of promising research directions into cognitive mechanisms and theoretically relevant aspects of spoken interaction. For instance, it creates opportunities to evaluate the relative importance of different acoustic–prosodic features and units of analysis in relation to the cognitive systems that support speech communication. It also highlights the value of investigating the temporal dynamics of entrainment—how it develops and unfolds during interaction—as well as identifying universal tendencies versus those that are language- or culture-specific. Such inquiries can deepen our knowledge of how communicative behavior can be formally modeled. Beyond advances in understanding human cognition, this work has several lines of applied research and development. For example, equipping robots and automatic spoken dialogue systems with entrainment capabilities may ultimately lead to more natural, effective, and seamless human–machine communication. In this way, entrainment research not only enhances our understanding of human interaction but also contributes to the design of future communication technologies.

However, in an effort to characterize the current state of the art, I will review several studies suggesting that entrainment research is, at present, difficult to interpret. Different languages and cultures appear to employ entrainment in distinct ways, even when comparable datasets and methodologies are used. Features extracted from a single dataset can simultaneously reveal both entrainment and disentrainment, complicating interpretation. Variables such as biological sex or conversational role do have measurable effects, but they may not fully account for individual variation. Moreover, results derived from different metrics applied to the same corpus often fail to corroborate one another, underscoring the fact that the choice of operationalization plays a critical and complex role. Similarly, analyses across linguistic levels—such as prosody, word choice, or semantics—and their interrelationships with entrainment do not tend to yield a coherent systematic patterns. Hence, the variability of observed results points toward a very complex behavior, or a set of behaviors with elusive latent structure. This, in turn, poses significant challenges for developing comprehensive formal models of entrainment capable of a)

capturing its dynamic and context-dependent nature and b) being implemented, deployed and tested in human-machine interactions.

Despite this complexity, the motivations outlined in the first two paragraphs above remain relevant, and existing measures and metrics do capture some aspects of entrainment in diverse and meaningful ways. I do not have a magic solution, but looking ahead, one important step seems to be the development of annotated benchmark corpora that integrate perceptual data, such as human ratings of alignment, to provide a reliable ground truth. At the same time, perception studies are needed to determine how listeners identify and respond to entrainment during interaction, thereby linking acoustic–prosodic features with social and cognitive evaluations. In addition, analysis-by-synthesis approaches offer a promising methodological avenue, enabling researchers to generate and test controlled speech patterns in order to refine theories of entrainment.

After several decades of research, speech entrainment is far from a well-understood behavior, but this is precisely why it still offers a very interesting and challenging agenda for future inquiry, with the potential to connect cognitive, social, and technological perspectives on speech.

Cross-fertilization between speech science and technology for the study of conversational speech

Barbara Schuppler[1]

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

For several decades, the study of conversational speech has come to interest to many areas of speech science and technology, spanning phonetics [6, 13, 34, 2], psycholinguistics [8, 33], automatic speech recognition (ASR) [16, 19], speech synthesis [5], and dialogue systems [30]. Compared to research on prepared or read speech, research on conversational speech can be complex, starting from the collection of conversational speech databases [32, 31, 26], their manual annotation, the development of automatic annotation tools, and the development of multi-layered turn-taking annotation [27, 12]. One of the key challenges for all of the disciplines working with conversational speech is to find approaches to statistically analyze and interpret variation [7], and/or how to make models more robust against variation. Conversational speech exhibits variation at all linguistic levels: intra-speaker and inter-speaker variation, differences across conversational settings and speaker relationships [3], as well as variation stemming from the type of task and the topic. And even in corpora, where many of *the most obvious* factors for variation were controlled for, as soon as people engage in spontaneous conversation, there is high variation. In this talk, I will present insights on pronunciation variation and prosodic variation in conversational speech, based on studies that combined methods from speech science and technology into one methodological cycle[1] (cf. Figure 1), allowing us to draw conclusions about the general strategies employed by most speakers, while still also paying attention to speaker-dependent variation.

This interdisciplinary undertaking requires an interdisciplinary team. At the SPSC Laboratory, my team and and collaborators bring together expertise from signal processing, (room) acoustics, speech recognition, natural language processing (NLP), phonetics and conversation analysis. Together, we develop models for pronunciation variation and prosodic variation in conversational speech, subsequently tested in ASR systems and human perception experiments. In doing so, we investigate which contextual, lexical and acoustic cues affect ASR performance, and also interpret them from a human speech processing point of view. When building classifiers for linguistic labels (e.g., prosodic prominence levels), we use explainable methods that not only allow us to know which features contribute to performance improvements, but are also cues to human perception of the mentioned labels (e.g., perceived prosodic prominence). The (quantitative) phonetic studies are facilitated by technologies from ASR and NLP. For instance, we develop tools for data annotation [21, 14], for acoustic feature extraction, and we apply advanced statistical analysis methods and classification approaches together with explainable AI tools [15]. These quantitative analyses

[1]For a more comprehensive literature review on the cross-fertilization between speech science and technology, please consult the book *Rethinking Reduction*, which presents contributions that show how interdisciplinary approaches yield new insights into acoustic reduction phenomena [2]. Our recent survey demonstrates that a cross-fertilization between speech science and technology is particularly fruitful when addressing with atypical speech and speech from low-resourced languages and language varieties [24].

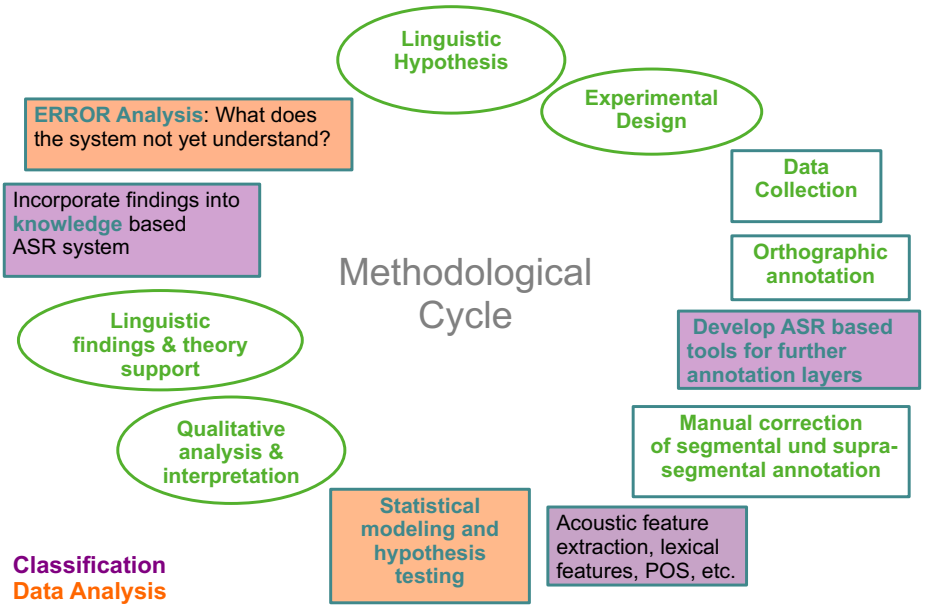


Figure 1: Methodological cycle to study variation in conversational speech allowing for a cross-fertilization between speech science (green) and technology (blue). In our research on conversational speech, the cycle may have started and ended at a different point depending on the aims and research questions of a specific study. Together as a team, however, we completed the full cycle for our models on pronunciation variation and prosodic variation.

are combined with qualitative approaches [35]. Also here, we follow a combined approach, using tools (e.g., *PyCaret* [1]) to facilitate the process of combining quantitative analysis (e.g., for speaker-independent strategies, for estimating the relative impact of features) with a qualitative analysis, for instance, for specific speakers, for specific types of tokens, or for outlier/error analysis. When combining quantitative and qualitative analyses, I believe it is important to document transparently which principles and selection criteria where followed for the choice of speech examples for the qualitative analysis (e.g., following methods from Conversation Analysis [23, 20]).

Qualitative vs. quantitative methods – this is only one of the ongoing debates in conversational speech analysis. Do we use controlled experiments or corpus studies? Do we use data from reading stories, task-oriented dialogues or free, casual conversations? My personal opinion very much aligns with Wagner, P. et al. [34, p. 10], who stated that "phonetics does not need to continue the debate about what is the most appropriate data, but we need a better understanding about how our methods, including the speaking style recorded, influence our results." In order to increase our understanding of the processes underlying conversational speech, we need corpus studies on materials that are as natural as possible, and we need to combine these corpus studies with controlled experiments (with humans or machines), using stimuli from read, synthetic or spontaneous speech. In psycholinguistics, new hybrid models of speech processing have developed based on insights from experiments with stimuli from

spontaneous speech material (e.g., [8]). Certain psycholinguists[2] even state that "... we need to investigate casual speech to truly understand language production, processing and the mental lexicon." [33]. From a speech technology perspective, we have found that improvements in signal processing methods achieved on read speech do not automatically transfer to conversational speech [25, 9]. In this talk, I will present insights from corpus studies, controlled production experiments and experiments with spontaneous speech stimuli.

Each experimental design comes with different challenges to the statistical analysis. The combination of prosodic, large contextual and local variables poses problems for traditional statistical methods (e.g., ANOVAs), even more so for conversational speech, where the variation is higher than in controlled experiments. A widely used and suitable modeling technique for the combination of categorical, continuous and random factors is mixed-effects logistic/linear regression [11, 4]. Regression-based approaches, however, run into problems with high-order interactions between predictors, correlating predictors, and the combination of a small sample size and a large number of predictors. Conditional inference trees (CIT) and random forests (RF) perform well under such conditions. Random forests were introduced to linguistics some time ago [29], and thereafter also applied to conversation and turn-taking analysis [22]. In the works presented in my talk, we used both CITs and RFs for feature importance analysis, and in some cases also for classification, outperforming other approaches that are not so well suited for small data sets (e.g., outperforming CNNs for homophone disambiguation [28]). For feature importance analyses, we used different measures: Gini importance (or mean decrease in impurity), permutation importance (or mean decrease in accuracy), as well as Shapley values [18]. What random forests cannot do, but regression approaches can, is to find relevant interaction terms. Thus, in earlier work, we used random forests to find the overall importances of all features of interest, and separate mixed-effects logistic/linear regression to investigate whether there were significant interactions between (a subset of not correlating) variables [17]. This gap has recently been filled by the so-called *Interaction Forests* [10]. Interaction forests are a recently developed variant of RFs, which provide effect importance measures (EIMs) for univariable effects as well as quantitative and qualitative interaction effects. Thus, in contrast to classical variable importance measures (VIMs), EIM values rank not only univariable effects, but also effects due to quantitative and qualitative interactions, which can be illustrated in a comprehensible manner. Recently, we introduced interaction forests to the field of speech science and applied them to the analysis of WERs of different ASR systems on conversational speech [15].

Summing up, in this talk I will present works that combine quantitative and qualitative methods from speech science and technology with the aim to increase our understanding of conversational speech processing. While the specific findings on pronunciation variation and prosodic variation may be language dependent, the methodological approach will hopefully stimulate further analyses in the field, not only by providing additional or more precise insights into previously posed research questions, but also by allowing to ask entirely new research questions.

[2]It is not a coincidence that I mention the work of my PhD supervisor Mirjam Ernestus here, who strongly contributed to my passion for investigating casual, conversational speech.

References

[1] Ali, M. (2020). 'PyCaret: An open source, low-code machine learning library in python (version 1.0.0)'. In: *https://www.pycaret.org, last viewed 18-03-2024*.

[2] Cangemi, F. et al., eds. (2018). *Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*. Berlin: Mouton de Gruyter. ISBN: 978-3-11-052417-8.

[3] Clopper, C. & R. Turnbull (2018). 'Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors'. In: *Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*. Ed. by Cangemi, F. et al. De Gruyter Mouton, pp. 25–72. DOI: 10.1515/9783110524178-002.

[4] Clopper, C. G. (2013). 'Mixed effects modeling: Applications and practice in speech research'. In: *Proc. Meetings on Acoustics*. Vol. 19, pp. 1–6.

[5] Dali, R. et al. (2016). 'Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis'. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5155–5159. DOI: 10.1109/ICASSP.2016.7472660.

[6] Dilley, L. C. & M. A. Pitt (2007). 'A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition'. In: *Journal of the Acoustical Society of America* 122.4, pp. 2340–2353.

[7] El Zarka, D. et al. (2024). 'The prosody of theme, rheme and focus in Egyptian Arabic: A quantitative investigation of tunes, configurations and speaker variability'. In: *Speech Communication* 160, p. 103082. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2024.103082. URL: https://www.sciencedirect.com/science/article/pii/S0167639324000542.

[8] Ernestus, M. (2014). 'Acoustic reduction and the roles of abstractions and exemplars in speech processing'. In: *Lingua* 142, pp. 27–41.

[9] Gabler, P. et al. (2023). 'Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition'. In: *Information* 14.2. ISSN: 2078-2489. DOI: 10.3390/info14020137. URL: https://www.mdpi.com/2078-2489/14/2/137.

[10] Hornung, R. & A.-L. Boulesteix (2022). 'Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects'. In: *Computational Statistics & Data Analysis* 171, p. 107460. DOI: https://doi.org/10.1016/j.csda.2022.107460.

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

[12] Kelterer, A. & B. Schuppler (2025). 'Turn-taking annotation for quantitative and qualitative analyses of conversation'. In: *arXiv preprint arXiv:2504.09980*. DOI: https://doi.org/10.48550/arXiv.2504.09980.

[13] Kohler, K. J. (2001). 'Articulatory dynamics of vowels and consonants in speech communication'. In: *Journal of the International Phonetic Association* 31, pp. 1–16.

[14] Linke, J., G. Kubin & B. Schuppler (2023). 'Using word-level features for prosodic prominence detection in conversational speech'. In: *Proc. of ICPhS 2023*, pp. 3101–3105.

[15] Linke, J. et al. (2025). 'What's so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures'. In: *Computer Speech & Language* 90, p. 101738.

[16] Lopez, A., A. Liesenfeld & M. Dingemanse (2022). 'Evaluation of Automatic Speech Recognition for Conversational Speech in Dutch, English and German: What Goes Missing?' In: *Proc. of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pp. 135–143.

[17] Ludusan, B. & B. Schuppler (2022). 'An analysis of prosodic boundaries across speaking styles in two varieties of German'. In: *Speech Communication* 141, pp. 93 –106. DOI: 10.1016/j.specom.2022.05.002.

[18] Lundberg, S. M. & S.-I. Lee (2017). 'A Unified Approach to Interpreting Model Predictions'. In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[19] Nakamura, M., K. Iwano & S. Furui (2008). 'Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance'. In: *Computer Speech and Language* 22.2, pp. 171–184.

[20] Ogden, R (2012). 'The Phonetics of Talk in Interaction – Introduction to the Special Issue'. In: *Language and Speech* 55.1, pp. 3 –11. DOI: https://doi.org/10.1177/0023830911433559.

[21] Paierl, M. et al. (2023). 'Creapy: A Python-based tool for the detection of creak in conversational speech'. In: *20th International Congress on Phonetic Sciences : ICPhS 2023, ICPhS 2023 ; Conference date: 07-08-2023 Through 11-08-2023*, p. 1716. URL: https://www.icphs2023.org/call-for-papers/.

[22] Roberts, S. G., F. Torreira & S. C. Levinson (2015). 'The effects of processing and sequence organization on the timing of turn taking: a corpus study'. In: *Frontiers in Psychology* 6, p. 509. DOI: 10.3389/fpsyg.2015.00509.

[23] Sacks, H., E. Schegloff & G. Jefferson (1974). 'A simplest systematics for the organisation of turn-taking for conversation'. In: *Language* 50, pp. 696–735. URL: https://www.jstor.org/stable/412243.

[24] Schuppler, B. et al. (2023). 'An introduction to pluricentric languages in speech science and technology'. In: *Speech Communication* 156.103007. DOI: https://doi.org/10.1016/j.specom.2023.103007.

[25] Schuppler, B. (2017). 'Rethinking classification results based on read speech, or: Why improvements do not always transfer to other speaking styles.' In: *International Journal of Speech Technology* 20.3, pp. 699 –713. DOI: 10.1007/s10772-017-9436-y.

[26] Schuppler, B., M. Hagmüller & A. Zahrer (2017). 'A corpus of read and conversational Austrian German'. In: *Speech Communication* 94C, pp. 62–74.

[27] Schuppler, B. & A. Kelterer (2021). 'Developing an Annotation System for Communicative Functions for a Cross-Layer ASR System'. In: *Proceedings of Workshop the First Workshop on Integrating Perspectives on Discourse Annotation (DiscAnn)*. URL: https://aclanthology.org/2021.discann-1.3/.

[28] Schuppler, B. et al. (2022). 'Homophone Disambiguation Profits from Durational Information'. In: *Proc. Interspeech 2022*, pp. 3198–3202. DOI: 10.21437/Interspeech.2022-10109.

[29] Tagliamonte, S. & R. Baayen (2012). 'Models, forests and trees of York English: Was/were variation as a case study for statistical practice'. In: *Language Variation and Change* 24.2, pp. 132–178.

[30] Tian, L., J. Moore & C. Lai (2016). 'Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features'. In: *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 565–572. DOI: 10.1109/SLT.2016.7846319.

[31] Torreira, F., M. Adda-Decker & M. Ernestus (2010). 'The Nijmegen Corpus of Casual French'. In: *Speech Communication* 52.3, pp. 201–212.

[32] Torreira, F. & M. Ernestus (2012). 'Weakening of Intervocalic /s/ in the Nijmegen Corpus of Casual Spanish'. In: *Phonetica* 69.3, pp. 124–148.

[33] Tucker, B. V. & M. Ernestus (2016). 'Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon.' In: *The Mental Lexicon* 11.3, pp. 375–400.

[34] Wagner, P., J. Trouvain & F. Zimmerer (2015). 'In defense of stylistic diversity in speech research'. In: *Journal of Phonetics* 48, pp. 1–12. URL: https://www.sciencedirect.com/science/article/abs/pii/S0095447014000941.

[35] Wepner, S. & B. Schuppler (2025). '(When) Does it Harm to Be Incomplete? Encoding ASR Mistranscriptions of Syntactically Disfluent Structures'. In: *Proc. of the 3rd Graz-Vienna Speech Workshop Vienna. Connecting with Health Sciences*, pp. 23 –24. ISBN: 978-3-903477-11-7. URL: https://repositorium.meduniwien.ac.at/obvumwoa/download/pdf/12365441.

Conference presentations/Konferenciaelőadások

From Brainwaves to Speech: A Diffusion-Based Approach

Fatima Ezzahra Akaaboune, Mohammed Salah Al-Radhi

Budapest University of Technology and Economics, Department of Telecommunications and AI, Hungary

1 Proposed Approach and Findings

Brain-to-speech synthesis offers a pathway for restoring communication in individuals with severe speech impairments. We present a two-stage diffusion-based model that reconstructs speech from intracranial EEG (iEEG) signals. In the first stage, a 5-layer temporal convolutional network (TCN) maps neural activity to coarse Mel-spectrograms, capturing temporal context via dilated convolutions. In the second stage, a U-Net style denoising diffusion probabilistic model (DDPM) refines these predictions over 50 iterative steps, enhancing spectral detail and reducing noise. Training was performed on the Verwoert *et al.* iEEG dataset [1], which contains synchronized recordings of brain activity and overt speech, with subject-specific models trained independently. Preprocessing included band-pass filtering (0.1–200 Hz), downsampling to 1 kHz, and segmentation into 50 ms windows; log-Mel spectrograms with 80 Mel bands (hop size 10 ms) served as targets.

Quantitative evaluation used Pearson correlations between predicted and ground-truth Mel-spectrograms, computed across frequency bins and cross-validation folds. Our diffusion-enhanced system consistently achieved high correlations (≥ 0.94 across subjects, peaking at 0.977), clearly outperforming linear regression (0.72) and a TCN-only baseline (0.88). These results indicate both robustness and generalization, with diffusion refinement adding measurable accuracy beyond deterministic mappings. For perceptual assessment, spectrograms were inverted into waveforms using Griffin–Lim [2]. Informal listening tests (N=5) confirmed smoother phonetic transitions, reduced background noise, and higher clarity compared to baseline reconstructions. Figures 1 and 2 illustrate reconstruction stability across participants and the fidelity of predicted spectrograms.

These findings suggest three key insights: (1) robust cross-subject generalization, indicating that fundamental mappings between neural activity and acoustic features are captured; (2) the diffusion refinement plays a critical role in suppressing spectral artifacts and producing smoother, more natural-sounding speech; and (3) limitations remain, as the current system is offline and relies on pre-aligned data, which constrains real-time deployment. Nonetheless, this two-stage diffusion pipeline advances the field of neural speech prosthetics and opens avenues for future work on streaming inference, end-to-end optimization, and integration with high-quality neural vocoders [3], such as WaveGlow or HiFi-GAN.

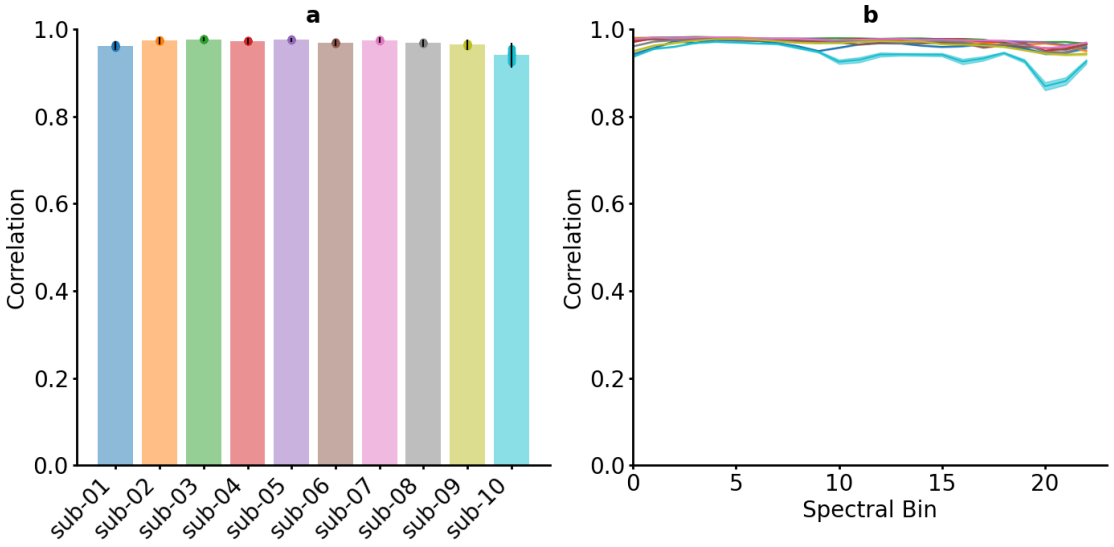


Figure 1: Mean correlation coefficients across participants and frequency bins. Reconstruction is reliable for all 10 subjects.

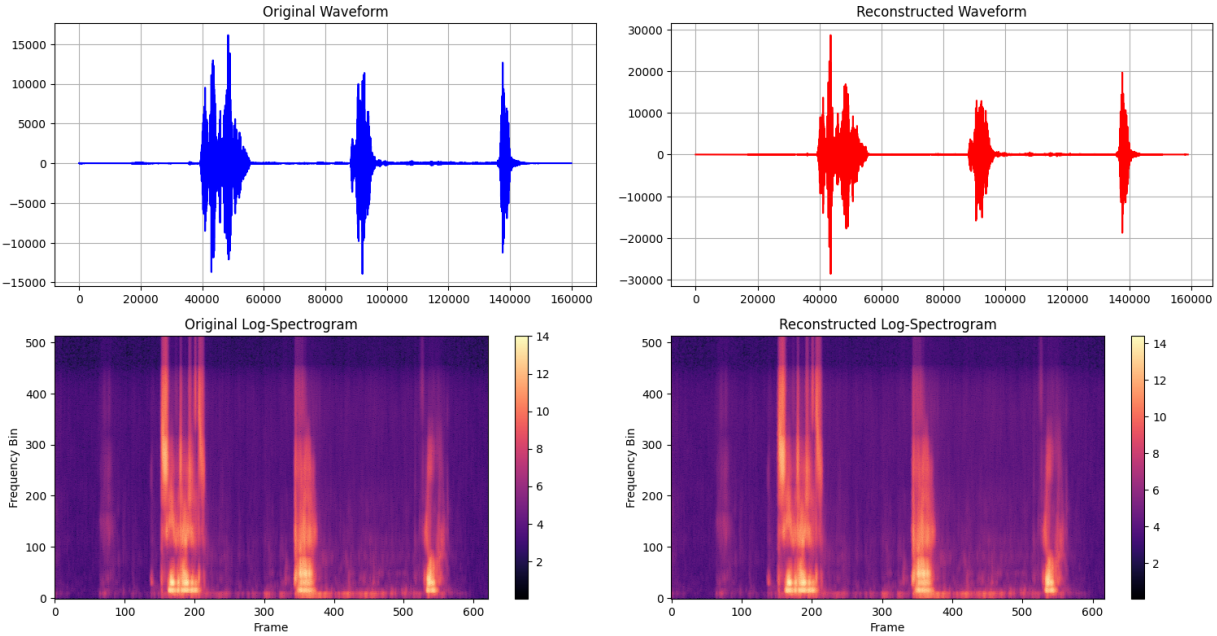


Figure 2: True vs. Predicted Mel-Spectrogram for Subject 03. The model captures key spectral features for accurate reconstruction.

2 Acknowledgement

This work is supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) . M. S. Al-Radhi's research was supported by the EKÖP-24-4-II-BME-197, through the National Research, Development and Innovation (NKFI) Fund.

References

[1] M. Verwoert, M. C. Ottenhoff, S. Goulis, A. J. Colon, L. Wagner, and S. Tousseyn, *et al.*, "Dataset of speech production in intracranial electroencephalography," *Scientific Data*, vol. 9, no. 434, pp. 1–9, 2022, doi: 10.1038/s41597-022-01542-9.

[2] H. Liu, T. Baoueb, M. Fontaine, et al., "GLA-grad: A griffin-lim extended waveform generation diffusion model", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 11611-11615, 2024.

[3] M. S. Al-Radhi, G. Németh, and B. Gerazov, "MiSTR: Multi-modal iEEG-to-speech synthesis with transformer-based prosody prediction and neural phase reconstruction", in *Proc. Interspeech*, Rotterdam, Netherlands, pp. 2930-2934, 2025.

Quantifying voiced stop lenition: data from L3 Spanish

Zsuzsanna Bárkányi[1,2] (zsuzsanna.barkanyi@open.ac.uk),
Zoltán G. Kiss[3] (gkiss.zoltan@btk.elte.hu)
[1]The Open University; [2]HUN-REN Research Centre for Linguistics;
[3]Eötvös Loránd University, Budapest

Although interest in *multilingual* (e.g., L3) phonetic and phonological acquisition has grown over the past decade (e.g. Wrembel 2023 for a review), studies on the acquisition of allophonic processes in this domain remain scarce, especially those focusing on advanced learners in their non-native languages. The aims of the study are: (1) to investigate Spanish voiced stop lenition (spirantisation) in the speech of L1 Hungarian, L2 English and L3 Spanish advanced sequential learners; (2) to investigate the methodological ramifications of the acoustic quantification of this process.

We conducted a production experiment with 14 participants using a randomized reading task which looked at Spanish voiced stop lenition in intervocalic position (i) word-internally (e.g., *sa**b**otaje, ca**d**a, a**g**uja,* etc.) as well as (ii) across the word boundary, where the stop is word-initial (e.g., *última **b**ala, era **d**ama, cada **g**allo,* etc.). For L1 Hungarians, this process creates sounds absent from their phonology, whereas in English (their L2), the voiced dental fricative /ð/ is part of the inventory, though /b d ɡ/ lenition is also absent. We investigated how advanced Hungarian learners produced intervocalic voiced stops (approximant vs. stop-like) and whether word position affected acquisition. To do so, a suitable quantification method was necessary.

Voiced stops /b d ɡ/ after non-nasal sonorants are realized as approximants [β̞ ð̞ ɣ̞]. Lenition is seen as gradient and can be quantified by the *intensity difference* (Idiff) between the consonant (its intensity *minimum*) and the following vowel (its intensity *maximum*: the lower the value, the more spirantized the consonant (Eddington 2011; Rogers & Alvord 2014; Amengual 2023). The procedure is superior to intensity averages, since lenited consonant boundaries are hard to define. The main issue concerns the cut-off between lenited vs. stop-like articulation. In our L1 Hungarian data, mean Idiff values were above 10 dB (Fig. 2, Table 1). Still, word-internal stops averaged 2.8 dB lower than across-word ones, with 40% lenited word-internally vs. 24% across boundaries. We argue that although L1 Hungarians perceive lenition, they do not treat it as linguistically relevant in L3 Spanish. This likely reflects L1 sociophonetic cross-linguistic interference rather than strict L1 phonology. Our results support earlier findings: lenition is not acquired as a postlexical process but resembles transitional sound change (Hualde 2013).

These results are solely based on the intensity difference measurements. In this paper we will look at other acoustic measures (like HNR, burst quality, center of gravity). Finally, we will also discuss whether the acoustic quantification of stop lenition should be based on binary classification (into "stop" vs. "non-stop/continuant" groups), using specific cut-off values (like the 10 dB Idiff boundary), or continuous measurements (like Idiff), or that the two should be used in tandem to best determine (i) the presence of lenition, and once present, (ii) the degree of lenition.

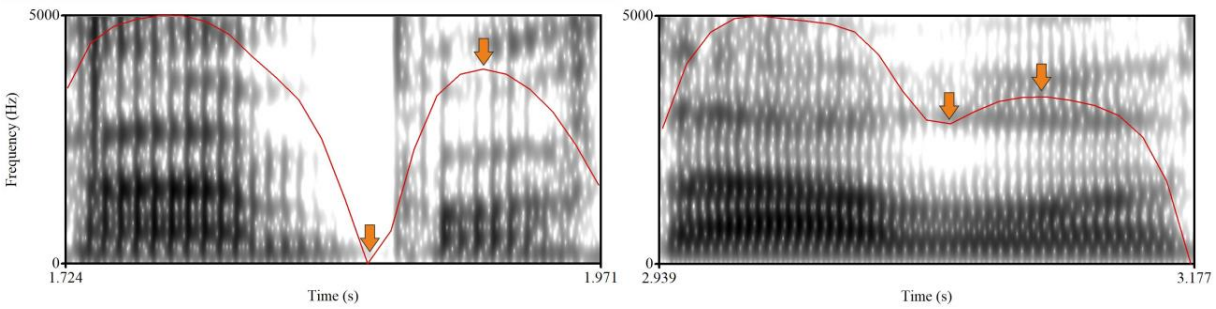


Fig. 1: Idiff of intervocalic /b/ in the Spanish word *(s)a**b**o(taje)* 'sabotage' by two participants. The difference is relatively large in the first case (left), hence /b/ is more stop-like, while it is relatively small in the second (right), hence a more approximant-like /b/-realization.

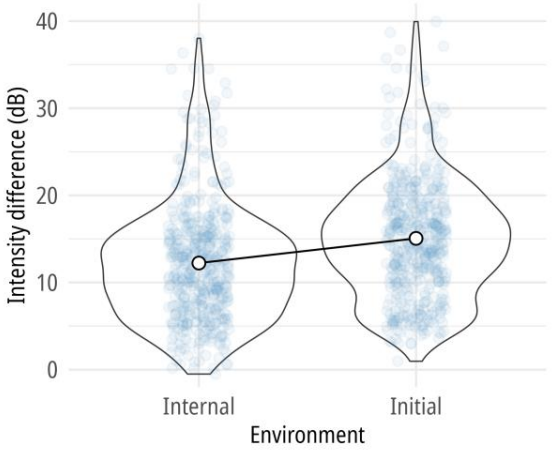


Fig. 2: Idiff of intervocalic /b d g/ in the Spanish words in two environments by L1 Hungarians. The white circles inside the violin plots represent the means.

Table 1: Descriptive statistics for Idiff of intervocalic /b d g/ in the Spanish words. B10 = number of observations below a 10dB intensity difference; A10 = number of observations above or equal to a 10 dB intensity difference.

Environment	N	Mean	SD	Median	Min	Max	B10	A10
Internal	448	12.23	6.88	11.54	−0.50	38.03	180 (40.18%)	268 (59.82%)
Initial	446	15.06	6.98	14.65	0.97	39.94	107 (23.99%)	339 (76.01%)

References

[1] Amengual, Mark. 2023. The Acoustic Realization of L2 Spanish Phonetic Categories and Allophonic Alternations by English-Speaking Immigrants. In R. Skarnitzl & J. Volín (eds) *Proceedings of the 20th International Congress of Phonetic Sciences* (ICPhS).

[2] Eddington, David. 2011. What are the contextual phonetic variants of /b d ɡ/ colloquial Spanish? *Probus* 23: 1–19.

[3] Hualde, José. 2013. Intervocalic Lenition and Word-Boundary Effects: Evidence from Judeo-Spanish. *Diachronica* 30(2): 232–266.

[4] Rogers, Brandon & Scott M. Alvord. 2014. The gradience of spirantization. *Spanish in Context* 11(3): 402–424.

[5] Wrembel, Magdalena. 2023. Exploring the Acquisition of L3 Phonology: Challenges, New Insights, and Future Directions. In: J. Cabrelli et al. (eds) *The Cambridge Handbook of Third Language Acquisition*. Cambridge: Cambridge University Press. 115–141.

Dizartriás beszéd gépi leiratozásának javítása implicit nyelvmodellekkel

Barta Piroska Zsófia[1, 2], Mihajlik Péter[1, 3]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem
[2]SpeechTex Kft.
[3]ELTE Nyelvtudományi Kutatóközpont

A modern beszédfelismerő rendszerek egyre hatékonyabban képesek leiratozni az egynyelvű beszédet, azonban a nem tipikus beszédet, mint amilyen például a beszédmotoros zavarokkal élőké, gyakorlatilag nem tudják kezelni.

A dizartria a beszédprodukció különböző aspektusait befolyásoló neurológiai eredetű beszédzavarok összefoglaló neve [2]. Bár a modern beszédfelismerési modellek elviekben többékevésbé képesek megtanulni a különböző beszédzavarokból adódó egyedi mintákat [5], a leiratozás minősége még mindig elmarad a tipikus beszédétől – már csak az atipikus beszédadatok gyűjtésének és feldolgozásának nehézségei miatt is.

Egy korábbi tanulmány [3] többek között egy Connectionist Temporal Classification (CTC)-alapú FastConformer_hu [1] mélyneurális modellt alkalmazott dizartriás beszédfelismerésre, ez a modelltípus önmagában azonban a CTC következtében nem képes hasznosítani a szöveg kontextusát a különböző hibák javítása érdekében, hiszen a CTC feltételes függetlenséget tételez fel a tokenek között. Ennek következtében az említett típusú modellek kimenetei gyakran tartalmaznak a mondat kontextusába nem illő, vagy értelmetlen szavakat, melyek számát ezen modellek esetében egy külső nyelvmodell alkalmazásával lehet csökkenteni. Jelen tanulmány célja egy alternatív modelltípus, a FastConformer-Transducer [4, 1] alkalmazása, amely önmagában képes kezelni a szöveges kontextust, tehát nem szükséges külső nyelvi modell integrálása.

Az eddigi kutatások [3] alapján az a tendencia figyelhető meg, hogy a dizartriás beszédfelismerő rendszerek vonatkozásában elengedhetetlen a beszédfelismerő modellek által használt adathalmazok személyre szabása. A perszonalizált adatgyűjtési folyamat azonban időigényes, illetve korlátozza az amúgy is kevés rendelkezésre álló adat újfelhasználhatóságát. Az egyedi modellek tanításhoz ezért két magyar anyanyelvű dizartriás személy hanganyagait hasznosítottuk, amint azt az 1. táblázat mutatja. Az első személy azonos volt egy előző tanulmányban szereplővel [3], a tanító, validációs és teszt adathalmazokat képző felhasznált hanganyagokon nem változtattunk. A második személy beszélőspecifikus tanító halmaza nem tartalmazott Text-To-Speech (TTS) rendszerrel előállított hanganyagokat, a rendelkezésre álló természetes minták mennyisége is kevesebb volt. A tanító halmaz időbeli terjedelme körülbelül az előbbi személy modelljéhez alkalmazott emberi beszédminták időtartamának fele volt, öt eltérő beszédtempóban. Míg a validációs halmazok leirata a két személy esetében minimális módosításokat leszámítva megegyezett, a második személy modelljéhez a teszt adathalmaz az eredetileg tanító adathalmaznak használt egyénre szabott Train_dys egy részének elkülönítésével jött létre.

Az első személyre vonatkozóan a FastConfromer-Transducer modellarchitektúra alkalmazá-

sával két metrika tekintetében is jelentős javulást sikerült elérni a FastConformer_hu modellhez képest. A WER a teszt halmazon 13,68%-ra csökkent, ami az eredeti értékhez képest 25,2%-os relatív javulást jelent, a CER pedig 21,4%-os relatív javulással 5,77%-ra csökkent. A Zero-Shot (ZS) Word Error Rate (WER) a FastConfromer-Transducer modellel a teszt adathalmazon 47,54% a Character Error Rate (CER) 24,94% volt.

A második, súlyosabb dizartriával rendelkező személy hanganyagainak segítségével személyre szabott FastConfromer-Transducer modellel a ZS WER 89,43% volt, a CER 52,66%. A CTC modellhez képest a WER és a CER relatív javulása rendre 29,03% és 23,89% volt. Bár az eredmények ebben az esetben is javuló tendenciát mutatnak, az abszolút hibaértékek még túl magasak a gyakorlati használathoz. A vonatkozó eredmények a 2. táblázatban találhatóak.

A jövőben szeretnénk bővíteni a második személy hanganyagait, többek között TTS-el szintetizált mintákkal. A többnyelű kommunikáció széles körben való elterjedéséből adódóan továbbá szeretnénk olyan többnyelvű ASR modellt létrehozni, ami alkalmas dizartriás beszéd felismerésre, illetve szeretnénk megvizsgálni a nagy nyelvi modellek alkalmazásának potenciális előnyeit a beszélőpartner által biztosított kontextus kihasználásának céljából.

	1. Személy			**2. Személy**		
	Tanító halmaz	**Validációs halmaz**	**Teszt halmaz**	**Tanító halmaz**	**Validációs halmaz**	**Teszt halmaz**
Szegmensek száma	195	40	107	200	125	123
Szavak száma	1406	363	1382	607	365	350
Karakterek száma	9523	2331	8326	3992	2258	2389

1. táblázat. A felhasznált dizartriás adathalmazok metaadatai

		FastConformer_hu		**FastConformer-Transducer**	
	Halmaz	**WER**	**CER**	**WER**	**CER**
1. Személy	Validációs	26,72	8,62	19,83	6,95
	Teszt	18,3	7,34	13,68	5,77
2. Személy	Validációs	78,08	43,84	65,48	36,85
	Teszt	72,86	34,37	51,71	26,16

2. táblázat. A különböző architektúrájú beszédfelismerő modellek WER és CER metrikák szerinti teljesítménye [%] finomhangolás után

Hivatkozások

[1] Dobsinszki, G. és tsai. (2025). „Best Data is More Supervised Data – Even for Hungarian ASR". *International Conference on Speech and Computer (SPECOM)*.

[2] Duffy, J. R. (2019). *Motor Speech Disorders E-Book: Motor Speech Disorders E-Book*. Elsevier Health Sciences.

[3] Mihajlik, P. és tsai. (2025). *Improved Dysarthric Speech to Text Conversion via TTS Personalization*. arXiv: 2508.06391 [cs.SD]. URL: https://arxiv.org/abs/2508.06391.

[4] Rekesh, D. és tsai. (2023). „Fast Conformer With Linearly Scalable Attention For Efficient Speech Recognition". *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–8. old. DOI: 10.1109/ASRU57964.2023.10389701.

[5] Shor, J. és tsai. (2019). „Personalizing ASR for Dysarthric and Accented Speech with Limited Data". *Interspeech 2019*, 784–788. old. DOI: 10.21437/Interspeech.2019-1427.

A spontán beszéd jellemzői óvodás és kisiskolás korban a szocioökonómiai sajátosságok tükrében

Bóna Judit és Murányi Sarolta
Eötvös Loránd Tudományegyetem

Számos szociolingvisztikai és pedagógiai kutatás rávilágított már évtizedekkel ezelőtt is arra, hogy a szociálisan hátrányos helyzetű gyermekek nyelvi és tanulási készségeinek fejlődése elmaradhat a nem hátrányos helyzetű gyermekekéhez képest megfelelő segítség és fejlesztés nélkül. A vizsgálatok szerint leginkább az anya iskolai végzettségének hatása kimutatható napjainkban is ([1, 6, 4]). A nyelvi készségek a szókincs, a narratív készségek és a beszédfeldolgozás terén is elmaradást mutatnak ([2, 3, 5]). Ugyanakkor keveset tudunk arról, hogy milyen grammatikai és pszicholingvisztikai jellemzői vannak a hátrányos helyzetű gyermekek beszédének különböző életkorokban és beszédtípusokban. A jelen előadás célja ennek vizsgálata.

A kutatási kérdés megvizsgálásához spontán beszédet rögzítettünk 15 hátrányos helyzetű nagycsoportos óvodással (5-6 évesek) és 15 korban és nemben illesztett kontroll gyermekkel, illetve 15 hátrányos helyzetű kisiskolással (8-9 évesek) és 15 korban és nemben illesztett kontroll kisiskolás gyermekkel. Minden gyermekkel két beszédtípusban készítettünk felvételt: a gyermekeket szabadidejükről és a mindennapi óvodai/iskolai tevékenységeikről kérdeztük, illetve egy képsor alapján egy történetet meséltettünk el velük.

A hangfelvételeket lejegyeztük, majd megvizsgáltuk a szövegek nyelvi komplexitását (MLU szám), illetve a beszédprodukciók folyamatosságát. Ez utóbbit a beszédszakaszok hossza és a megakadás-jelenségek előfordulási gyakorisága mutatja. A következő megakadásokat vettük figyelembe: kitöltött szünet, töltelékszó, nyújtás, ismétlés, újraindítás, szünet a szóban, önjavítás. A gyakorisági értékeket 100 szótagra vetítve számítottuk ki.

Az adatokon statisztikai elemzést végeztünk az SPSS szoftverrel. Összevetettük ugyanazon életkorú gyermekek két csoportjában az adatokat, illetve a hátrányos helyzetű gyermekeket kétféle életkorban, és a nem hátrányos helyzetű gyermekek adatait is a kétféle életkorban. Az adatokat a beszédtípusok szerint is megvizsgáltuk. Az MLU-szám és a beszédprodukciók folyamatosságára vonatkozó mutatók között korrelációs elemzést végeztünk.

Eredményeink igazolták a két csoport közötti különbséget mindegyik vizsgált paraméter szempontjából. A hátrányos helyzetű gyermekek beszédének grammatikai komplexitása és beszédük folyamatossága óvodás korban nagyobb elmaradást mutatott a nem hátrányos helyzetű gyermekekéhez képest, mint kisiskolás korban. Emellett mindkét beszélői típusban különbség volt a két életkor között. A beszédtípus is hatással volt az eredményekre: a mindennapi életről szóló spontán beszédben kisebb különbség volt a gyermekek két csoportja között az egyes életkorokban, mint az irányított spontán beszédben (képsorról való történetmesélés).

Az adatok felhívják a figyelmet arra, hogy bár az iskolának és a spontán fejlődésnek van kompenzáló hatása, a hátrányos helyzetű gyermekek nyelvhasználatát az iskolában is szükséges célzottan fejleszteni.

Irodalom

[1] Dollaghan, C. A, T. F. Campbell, J. L. Paradise, H. M. Feldman, J. E. Janosky, D. N. Pitcairn & Kurs-Lasky M. (1999). Maternal education and measures of early speech and language. *Journal of Speech, Language and Hearing Reserach* 42.6, 1432-1443. old. DOI: 10.1044/jslhr.4206.1432. PMID: 10599625.

[2] Hódi Á. & Tóth E. (2016). A különböző szocioökonómiai státuszú tanulók iskolakezdéskor mért elemi alapkészségeinek és a későbbi szövegértés teljesítményének alakulása az óvodában eltöltött évek tükrében. *Iskolakultúra* 26.9, 51–72. old. DOI: 10.17543/ISKKULT.2016.9.51.

[3] Mozzanica F., F. Ambrogi, R. Salvadorini, E. Sai, R. Pozzoli, M. R. Barillari, L. Scarponi & A. Schindler (2016). The Relationship between Socioeconomic Status and Narrative Abilities in a Group of Italian Normally Developing Children. *Folia Phoniatria Logopedia* 68.3, 134-140. old. DOI: 10.1159/000452443.

[4] Sipos Zs. (2022). Szocioökonómiai sajátosságok hatása az olvasástechnika elsajátítására második évfolyamos tanulók vizsgálata alapján. *Anyanyelv-pedagógia* 15. 1-20. old. DOI: 10.21030/anyp.2022.3.1.

[5] Szántó A. (2016). 'Beszédfeldolgozási folyamatok változásai a fejlesztés függvényében. doktori értekezés'. ELTE Bölcsészettudományi Kar. Budapest.

[6] Vakula T. & Bóna J. (2018). Characteristics of storytelling in preschool- and schoolchildren depending on family environment. Előadás a 2nd International Conference on Sociolinguistics konferencián. ELTE, Budapest.

Voice quality as a marker of phrase-final boundaries in different sentence types

Fruzsina Csébfalvi[1] and Anna Kohári[2]
[1] Radboud University, Nijmegen, The Netherlands
[2] ELTE Research Centre for Linguistics, Hungary

Introduction. Modal voice stems from quasi-periodic vocal fold vibration, yet various non-modal voice qualities may also be produced under different laryngeal configurations. Usually, three categories of voice quality are distinguished, which can be perceptually differentiated from each other: modal, breathy and creaky voice [5]. The difference between the types of voice quality correlates to the relative distance between the vocal folds (Figure 1). During breathy phonation the vocal folds are spreading, and the produced friction noise can be detected by the listener, and it shows in the acoustic signal as well [5]. When producing a creaky voice, the vocal folds are typically constricted. Because of this compression a much bigger subglottal pressure is needed for the flowing air to spread the vocal folds. As a result, they will separate less often and in a less periodic rhythm, therefore during glottalization the frequency or amplitude of the periods of the voice in waveform can fluctuate dynamically [5]. Non-modal voicing can have a wide range of varying purposes in language, one of which is signaling intonational phrase boundaries [2]. Regarding Spanish language, González [6] found that creaky voice is an important prosodic marker in declarative sentences, and non-modal voicing in general is more likely to appear at the end of intonational phrases. Based on experiments made with Hungarian speakers, Markó [8] concluded that glottalization is a typical tool for marking boundaries among Hungarian speakers as well. The results suggest that the closer a syllable is to the end of the utterance, the more likely it is to be glottalized. The boundary-marking function of creaky phonation in Hungarian language is primarily examined irrespective of sentence types. Markó [8] conducted an experiment with Hungarian speakers where they paid attention to the possible connection between sentence types and voice quality. Among typical speakers they found that, similarly to declarative sentences, yes/no interrogatives tend to have a glottalized last syllable. In other languages, some studies suggest that different kinds of voice qualities (e.g., breathy voice) may serve as acoustic cues signaling different illocutionary acts [3, 4].

The aim of this study is to determine whether phrase-final glottalization systematically occurs in Hungarian, regardless of sentence types. Our research questions were the following: do the final syllables in Hungarian exclamatives, imperatives, interrogatives and optatives realize generally with glottalized vowels, similarly to declarative sentences? Is there any connection between a sentence's last syllable's voice quality and the sentence type?

Methods. Voice recordings were made with 10 native Hungarian speakers, 5 men and 5 women and they were all between the ages of 19 and 26. The participants uttered different types of sentences in an appropriate linguistic context. We used a semantic framework [1] in which the material was constructed, as it defines the semantic profiles of the three major (declarative, interrogative, imperative) and two minor (exclamative, optative) sentence types with direct illocutionary acts in Hungarian. Overall, 60 sentences were uttered, which consisted of 10 sentence-groups. Each group had 6 sentences with the exact same number of syllables, exactly matching each other on the last word, but were all of different sentence-types. All examined sentences ended with words structured as CVCVCVC. The material consisted of 118 target sentences and matching contexts to implicitly evoke the appropriate intonation structure for the

sentence. This method of the experiment was inspired by the study of Dehé & Braun [3]. The recordings were manually labelled in Praat. The last three vowels of every target sentence were examined. Every vowel was perceptually and visually classified as modal, breathy or creaky. A vowel was modal if the voicing was regular, no signs of breathy or creaky voicing could be detected (Figure 2). A vowel was breathy if clearly detectable noise appeared in the acoustic sign (Figure 2). A vowel was creaky (Figure 2) if there was irregularity in frequency or amplitude, or if it audibly showed the characteristics of prototypical creaky voice [5, 8]. An independent expert reviewed 10% of the labels, and the annotations for different phonation types showed a 96% interrater agreement. For the statistical evaluation of the occurrence rates of different voice quality categories, Fisher's exact test and post hoc analyses were performed in R.

Results. Our preliminary results focus on exclamatives and imperatives, while also comparing them to declarative sentences. Declaratives and imperatives exhibited similar patterns regarding the voice quality of the final syllable (Figure 3). In declaratives, the vowel in the final syllable was realized with glottalization in 80.00% of cases, breathy voice in 10.00%, and modal voice in another 10.00%. In the case of imperative sentences, the phrase-final vowel was produced with glottalization in 83.33% of cases, breathy voice in 13.33%, and modal voice in only 3.33%. A significant difference in voice quality across the three sentence types was found using Fisher's exact test ($p < .001$). The paired post hoc test, however, indicated no difference between declaratives and imperatives ($p = .17$). The voice quality pattern at phrase-final position in exclamatives differed markedly from those observed in declaratives and imperatives. In exclamatives, only 38.2% of phrase-final vowels were realized with glottalization, while the majority, 55.06%, were produced with modal voice, and 6.74% with breathy voice. The final vowel in exclamatives showed a statistically significant difference in voice quality relative to both declaratives and imperatives based on the post hoc tests ($p < .001$).

Discussion. This study is the first to investigate utterance-final voice quality across three sentence types in Hungarian. Earlier studies have shown that declarative sentences and yes/no interrogatives both tend to have a glottalized vowel in the final syllable [8]. However, papers examining other languages than Hungarian suggest that sentences with different illocution types involve the use of different voice qualities [3, 4]. According to our results, using glottalization on the last syllable as a marker of utterance boundaries varies depending on sentence type and is not universally present across all sentence types. While declarative and imperative sentences both had the tendency to have a creaky vowel in the final syllable, this pattern was not observed in exclamatives. Exclamatives realized with a modal vowel in the final syllable more than half of the cases. It has been observed that exclamatives typically end with a mid boundary tone, whereas other sentence types, such as interrogatives, are generally characterized by a phrase-final low boundary tone [7]. Since glottalization typically occurs at low fundamental frequencies, it is plausible that the different phrase-final tones across sentence types influence the type of voice quality. Another possible explanation could be that exclamative sentences contain irregular voicing (e.g., breathy voice) in earlier segments, causing the phrase-final portion to differ from the rest of the utterance in voice quality, even if realized with modal phonation. These potential explanations should be explored in future studies.

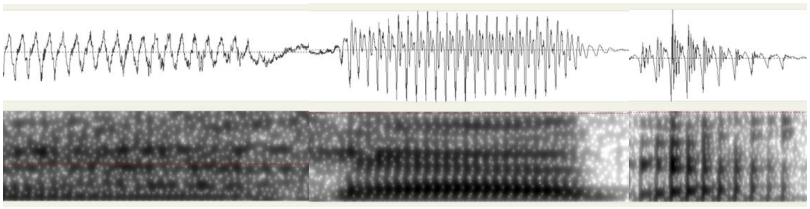


Figure 1: Oscillogram and spectrogram of breathy (left), modal (center), and creaky (right) voice.

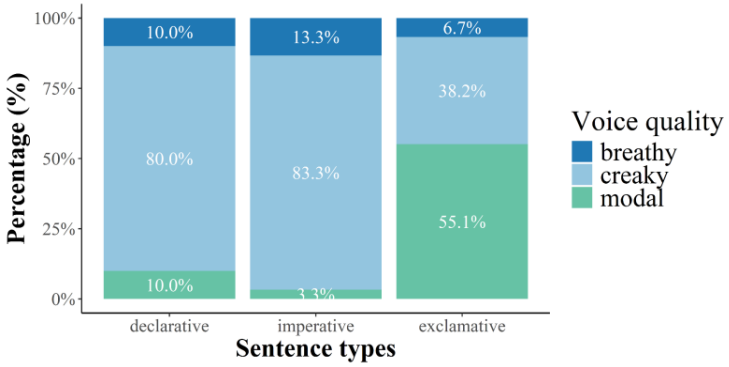


Figure 2: Proportional occurrence of voice quality types across different sentence types.

This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00160/25/1).

References

[1] Alberti, G., Dóla, M., Kárpáti, E., Kleiber, J., Viszket, A., & Szeteli, A. (2021). Lehetséges lehetséges világaink. *Jelentés és Nyelvhasználat*, 8(1), 105–145.

[2] Davidson, L. (2021). The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages. *Cognitive Science*, 12(3), e1547.

[3] Dehé, N., & Braun, B. (2020). The prosody of rhetorical questions in English. English *Language & Linguistics*, 24(4), 607–635.

[4] Dehé, N., & Wochner, D. (2024). Voice quality and speaking rate in Icelandic rhetorical questions. *Nordic Journal of Linguistics*, 47(1), 111–120.

[5] Garellek, M. (2019). The phonetics of voice. In W. Katz, & P. Assmann (Eds.), *Handbook of phonetics* (pp. 75–106). Routledge.

[6] González, C., Weissglass, C., & Bates, D. (2022). Creaky voice and prosodic boundaries in Spanish: An acoustic study. *Studies in Hispanic and Lusophone Linguistics*, 15(1), 33–65.

[7] Gyuris, B., & Mády, K. (2014). Approaching the prosody of Hungarian wh-exclamatives. In *VLlxx: Papers in Linguistics*. (pp. 333–349).

[8] Markó, A. (2013). *Az irreguláris zönge funkciói a magyar beszédben*. ELTE Eötvös Kiadó.

A magánhangzók hosszúsági szembenállása a nyelvi környezet függvényében

Deme Andrea[1,2,3] – Szánthó Zsuzsa[1,2,3] – Gráczi Tekla Etelka[2,3]
[1]ELTE Eötvös Loránd Tudományegyetem, [2]MTA–HUN-REN NYTK Lendület Neurofonetikai Kutatócsoport, [3]ELTE Nyelvtudományi Kutatóközpont

A nyelv(i hangzás) leírásában meghatározó fogalom a kontraszt. Ez a hangszintű jelenségek esetében azt jelenti, hogy a beszédhangok azon tulajdonságait azonosítjuk, amelyek elkülönítik azokat más hangoktól, így hozzájárulhatnak eltérő jelentésű hangsorok, ún. minimális párok megformálásához. A H&H elmélet [2] szerint a kontraszt megvalósítására számos tényező hat, amik két ellentétes erő alá rendeződnek: míg a beszédprodukciós rendszer a gazdaságosságra (a megformálásra fordított energia minimalizálása) törekszik, a hallgató felől a kontraszt detektálhatóságának szükségessége a nagyobb erőfeszítést motiválja. Így a beszéd minden helyzetben az alul- vagy túlartikuláltság extremitásai közötti kontinuumon belül valósul meg, ami rendre a kontrasztok gyengülésével vagy erősödésével jár.

A kutatások szerint a beszédhang-kontrasztok erősítése jelentkezik akkor, amikor a beszélő összetéveszthető szóalakok, minimális párok elkülönítésére törekszik úgy, hogy mindkét szóalak jelen van a nyelvi kontextusban (pl.: *Nem **X** hanem **Y***) [1, 3]. Olyan eredmények is ismertek, amelyek szerint a szavak fonológiai szomszédjai (a tőlük csak egy fonémában eltérő szavak) a szó relatív gyakoriságával együtt [8], de attól függetlenül is [5] a beszédhangok szélsőségesebb ejtése, ill. a minimális párok közti kontraszt erősítése felé hatnak. Ez azt jelenti, hogy azok a minimális párok is befolyásolják a szó/beszédhangok ejtését, amelyek nincsenek jelen a szó megvalósuló nyelvi környezetében. Valószínűleg a fenitek együttesen képezik az alapját annak a közkeletű feltételezésnek, amely szerint a minimális párok megjelenése egy beszédhelyzeten, például egy kísérlet nyelvi anyagán belül fokozza a párok közti kontraszt kifejezésének szándékát – még akkor is, ha a minimális párok további elemekkel vegyesen szerepelnek egy olyan felsorolásban, ahol egyenként mutatjuk be a szavakat. Erre nézve azonban közvetlen bizonyítékokról nincs tudomásunk, miközben ez a feltevés a kísérletes beszédtudományi vizsgálatok módszertani megfontolásait és kritikai fogadtatását is nagymértékben meghatározza. A szó előfordulásának gyakoriságáról is úgy vélik, hogy önmagában is befolyásolja az ejtést. A nagyobb gyakoriság vezethet törléshez, ill. redukcióhoz [6], tehát az alulartikuláció felé mozdíthatja a megvalósítást. A ritkább szavakban megjelenhet a kontrasztok, így például a magánhangzós hosszúsági szembenállás erősítése [7]. A nulla gyakoriságú szavakban (az álszavakban) pedig várható, hogy szintén túlartikulált az ejtés, hiszen az alacsony bejósolhatóság miatt a beszélő vélelmezheti az azonosítás nehezítettségét, de ezt az összefüggést csak nagyon korlátozottan írták még le eddig [4].

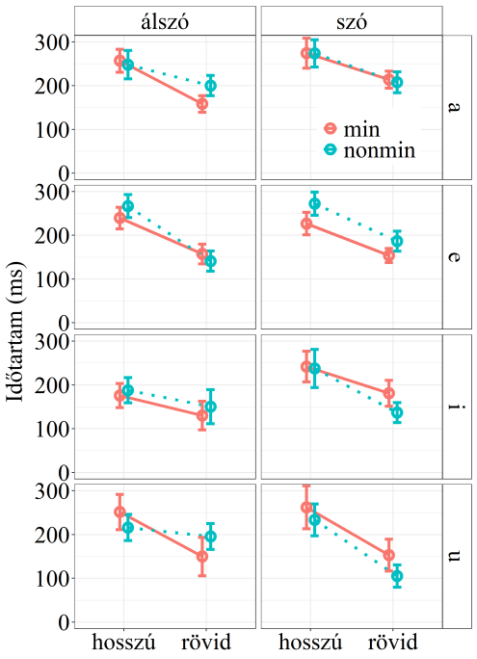
Vizsgáltunkban azt elemezzük, hogy miként hat a gyakoriság, szűkebben az álszóság ténye, ill. az anyagban megjelenő szószomszéd (minimális pár) a hosszúsági magánhangzó-kontraszt megvalósulására a magyarban. Ennek első lépéseként itt a magánhangzók időtartamát vizsgáljuk. Feltételezésünk szerint a beszélők eltérően ejtik ugyanazon magánhangzókat a különböző feltételekben (l. lentebb) úgy, hogy álszavakban, ill. minimális párokban (min) jobban elkülönítik a rövid-hosszú párok tagjait egymástól, mint valódi szavakban és nem minimális párokban (nonmin).

Kísérletünkben az [ɒ]-[aː], [ɛ]-[eː], [i]-[iː], [u]-[uː] magyar magánhangzópárok elkülönítését elemeztük 10 felnőtt magyar anyanyelvű nő ejtésében. A magánhangzókat négyféle feltételben produkálták a beszélők: 1. olyan valódi szavakban, amelyek nem alkotnak minimális párokat a kísérletben ejtendő további szavakkal a célhangok mentén (pl. *z**á**r-d**a**l*); 2. olyan valódi szavakban, amelyek minimális párokat alkotnak a kísérletben szereplő további szavakkal a célhangok mentén (pl. *t**á**r-t**a**r*); 3. álszavakban, amelyek mellett nem jelennek meg az anyagban

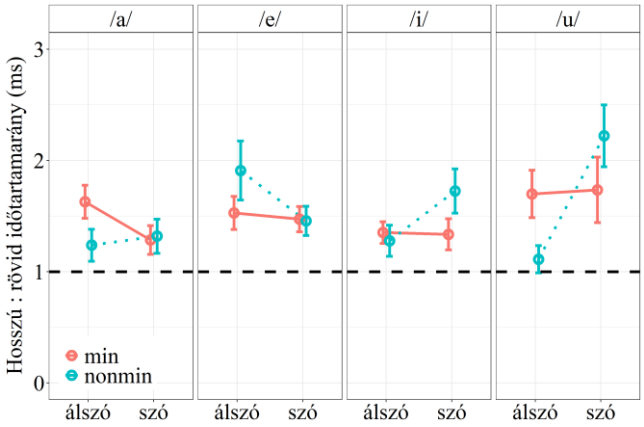
további, tőlük csak a célhangban eltérő további álszók (nem „minimális párban” jelennek meg, pl. ***lász-zal***); és 4. álszavakban, amelyek mellett megjelennek a kísérletben ezektől csak a célhangban eltérő további álszók („minimális párban” jelennek meg, pl. ***nát-nat***). Minden szót minden beszélő háromszori ismétlésben ejtett (mivel az anyag egy nagyobb kísérleti anyag része, több ismétlésre nem volt mód), egyénenként randomizált sorrendben úgy, hogy a célszavak önálló megnyilatkozásokként szerepeltek. A szó valódi szóként vagy álszóként való felismerését képekkel segítettük. A célszavakat és -hangokat kézzel szegmentáltuk, mértük szavak és célhangok tartamát, a kontrasztok vizsgálatához elemeztük a szegmentumok abszolút időtartamait és a hangpárok tartamarányát. Az adatokat a magánhangzó típusa szerint kialakított csoportokban vetettük össze lineáris kevert modellekkel (ill. egymintás *t*-próbával), az ábrákon a típust a hosszú magánhangzó minőségével jelezzük.

Az elemzések szerint a magánhangzók abszolút időtartamára sem a gyakoriság (az álszó vs. valódi szó), sem a minimális pár jelenléte az anyagban nem hatott robusztusan (1. ábra). A gyakoriság („álszóság”) minden pár esetében mutatott hatást legalább a kondíciók 1-1 kombinációjában, de nem mindenhol, és nem mindig azonos (ill. a várt) irányban tértek el a megfelelő feltételekben ejtett beszédhangok (pl. az [ɒ] rövidebb volt álszókban, ha minimális párban szereplő szóban állt, mint ha nem minimális párban állóban, míg az [aː] nem minimális párokban álszókban volt rövidebb, mint valódi szavakban). A hosszú és rövid beszédhangok aránya minden esetben eltért az 1-től, és szintén nem mutatta a gyakoriság vagy a minimális pár jelentének robusztus hatását, a két tényező minden magánhangzó-pár esetében interakcióban hatott az adatokra (2. ábra). Egy kísérlet anyagának megtervezése szempontjából a legfontosabb megállapítások a következők lehetnek. Az [ɒ] és [aː] a legnagyobb mértékben a várt módon akkor tért el egymástól, amikor „minimális párokban” megjelenő álszavakban ejtették őket a beszélők, de valódi szavakban nem volt különbség a minimális párok és nem minimális párok ejtése között. Az [ɛ] és [eː] legnagyobb mértékben szintén álszavakban különült el, de akkor, ha azok nem „minimális párokban” jelentek meg, ami csak részben vág egybe a várakozásokkal. Valódi szavakban – nem várt módon – itt sem okozott különbséget a célszó minimális párjának jelenléte az anyagban. Meglepő módon az [i] és [iː] legnagyobb mértékben valódi szavakban, nem minimális párokban tért el egymástól, míg minimális párokban ennél kisebb eltérés mutatkozott álszók és valódi szavak esetében. Ehhez hasonlóan alakultak az [u] és [uː] adatai: a legnagyobb elkülönítést szintén valódi szavakban, nem minimális párokban láttuk, a legkisebbet álszókban, nem „minimális párokban”, míg minimális párokban ugyanolyan mértékben különült el a hosszú és rövid hang álszókban és valódi szavakban is.

A kísérlet eredményei nem támasztották alá a várakozásokat (legalábbis nem egyértelműen), hiszen az egyes magánhangzó-párokban nem mutatkozott szisztematikusan a két vizsgált hatás: eltért az, hogy mely hatások fokozták a hosszúsági kontraszt tartambeli megvalósítását, és bizonyos esetben a hatások a várttal ellentétes irányúak voltak. Ez alapján kétségbe vonható, hogy az álszavakkal, ill. a minimális párokkal végzett vizsgálatok eredményei eredendően „félrevezetőek” lennének, ahogyan az a tudományos gyakorlatban nemritkán kritikaként megfogalmazódik. Ugyanakkor hasznos módszertani javaslatok tehetők az alapján a két megfigyelés alapján, amelyek szerint a) jellemzően létező szavakban találni nagyobb eltérést a párok tagjai között (szemben az álszavakkal), és b) ezekben (bizonyos beszédhangok esetében legalábbis) nem minimális párokban szintén erősebb lehet a kontrasztban résztvevő elemek elkülönítése, mint minimális párokban.



1. ábra: A magánhangzók abszolút időtartama



2. ábra: A hosszú és rövid magánhangzók időtartamaránya (vízszintes szaggatott vonal = a hosszú és rövid magánhangzók időtartama azonos)

Irodalom

[1] Granlund S., Hazan, V. & Baker, R. (2012). An acoustic-phonetic comparison of the clear speaking styles of Finnish–English late bilinguals. Journal of Phonetics, 40(3), 509–520.

[2] Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In Speech Production and Speech Modelling, Kluwer, Dordrecht, Boston, 403–439. DOI: 10.1007/978-94-009-2037-8_16

[3] Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. Journal of Phonetics 41, 249–263.

[4] Maxwell, O. Baker, B. Bundgaard-Nielsen, R. L. & Fletcher, J. (2015). A comparison of the acoustics of nonsense and real word stimuli: coronal stops in Bengali. In Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015), 10-14 August 2015, Glasgow, Scotland, UK.

[5] Munson, B. & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. Journal of Speech, Language, and Hearing Research 47: 1048–1058. DOI: 10.1044/1092-4388(2004/078)

[6] Pluymaekers, M., Ernestus, M., Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America* 118(4):2561–2569. DOI: 10.1121/1.2011150

[7] Tomaschek, F. Wieling, M., Arnold, D & Baayen, H. (2013). Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience. In Proceedings of Interspeech 2013, 25-29 August 2013, Lyon, France, 1302–1306.

[8] Wright, R. (2004). Factors of lexical competition in vowel articulation. In Local, J. Ogden, R. & Temple, R. (eds): Phonetic Interpretation Papers in Laboratory Phonology VI, Cambridge University Press, 75–87. DOI: https://doi.org/10.1017/CBO9780511486425.005

Köszönetnyilvánítás

A kutatás a Bolyai János Kutatási Ösztöndíj, az EKÖP-24 és az EKÖP-25 szakmai támogatásával készült.

Nőies és férfias beszélt nyelvi jellemzők a BEA adatbázis beszélőinél

Főző Eszter
Nemzetbiztonsági Szakszolgálat Szakértői Intézet
fozo.eszter@nbsz.gov.hu

Az NBSZ Szakértői Intézetben (és jogelődjeiben) az 1970-es évek óta működik kriminalisztikai nyelvészet. A szakterület egyik feladata, hogy a névtelen szerzőjű fenyegető, rágalmazó, zsaroló stb. szövegek nyelvhasználati jegyei alapján megállapítsa az ismeretlen fogalmazó nyelvi profilját, vagyis behatárolja a szerző nemét, életkorát, iskolázottsági szintjét, anyanyelvét stb., hogy a nyelvi profil alapján a nyomozóhatóság szűkíthesse a lehetséges elkövetők körét.

A kriminalisztikai nyelvészek azt tapasztalják, hogy a női(es) és férfi(as) nyelvhasználati jegyek közötti határ elmosódni látszik. Kevés a témában készült, modern, reprezentatív kutatás; a kriminalisztikai szövegnyelvészet szempontjából az 1980-as években készült egy átfogó vizsgálat ([7]), melyben a korpuszt 200 magyar nyelvű, névtelen bejelentés, feljelentés, panaszlevél adta, a referenciaszövegeket pedig 100–100 női és férfi fogalmazótól származó hivatalos levél szolgáltatta. Az összehasonlító vizsgálat szerint a nők bizalmaskodó, trágár szavakat használnak, halmozzák a minőségjelzőket, a mondanivalójukat lényegtelen, de emocionálisan erős részletekkel dúsítják, gyakran motívumismétléssel is élnek. Ezzel szemben a férfiak kedvelik a bonyolult birtokos szerkezeteket, a 4-nél több tagmondatból álló mondatokat, zsargonban írnak, a stílusuk tárgyszerű, hivatalos nyelvezetű, és gyakran hivatkoznak valaki másra/más által elmondottakra. A nyelvhasználat változása miatt azok a jellemzők, amelyek a baby boomer generációban kifejezetten nőies és férfias stílusjegynek számítottak, nem feltétlenül azok az X-generációban, és különösképpen nem azok az Y- és a Z-generációk beszélőinél/fogalmazóinál.

Amikor nőies és férfias nyelvhasználati jegyek vizsgálatáról van szó, elsősorban a magyar nyelvű kutatásokat részesítjük előnyben, ugyanis bizonyos (nemi) nyelvi markerek nyelvfüggőek – pl. a névmás használata (pl. [1]); a szavak hossza (pl. [5]), az elöljárószók gyakorisága (pl. [3]) –, így a nemzetközi tanulmányok tanulságai nem minden esetben alkalmazhatók a magyar szövegelemzési gyakorlatban. Bár a téma még forenzikus szakkönyvi fejezetben is helyet kap (pl. [4]), a női és férfi nyelvhasználati jegyekre vonatkozó hazai kutatások száma kevés, ezek elsősorban az írott nyelvi sajátosságok feltárására helyezik a hangsúlyt (pl. [10]), és a nemi különbségeket érintő, beszélt nyelvi spontánbeszéd-kutatások száma még kisebb (pl. [6]).

A jelen kísérletben a beszélt nyelvi jegyekre koncentrálunk, különösen a két nem szintaktikai, lexikai, pragmatikai hasonlóságainak és különbségeinek feltárására. A vizsgálatokhoz a beszélőket a BEA adatbázisból ([2]) választottuk ki a csoportjegyeik alapján, és mind a nem, mind az életkor, mind az iskolázottság szempontjából kiegyensúlyozott korpusz összeállítására törekedtünk. A 40-40 női és férfi beszélő mindegyike a kriminalisztikai szempontból releváns életkori sávba tartozik ([8], [11]). A beszélőket két korcsoportba soroltuk aszerint, hogy a felvétel idején a mintaadó hány éves volt; a korcsoportok az X-, illetve a babyboomer-generációnak feleltethetők meg ([9]). A korpusz összeállításánál a vizsgálati anyag terjedelmét is optimalizáltuk, így az egyes korcsoportok beszédleiratainak tokenszámai közel azonosak (I.: 16411, II.: 15737). A beszélőcsoportok adatait az 1. táblázat tartalmazza.

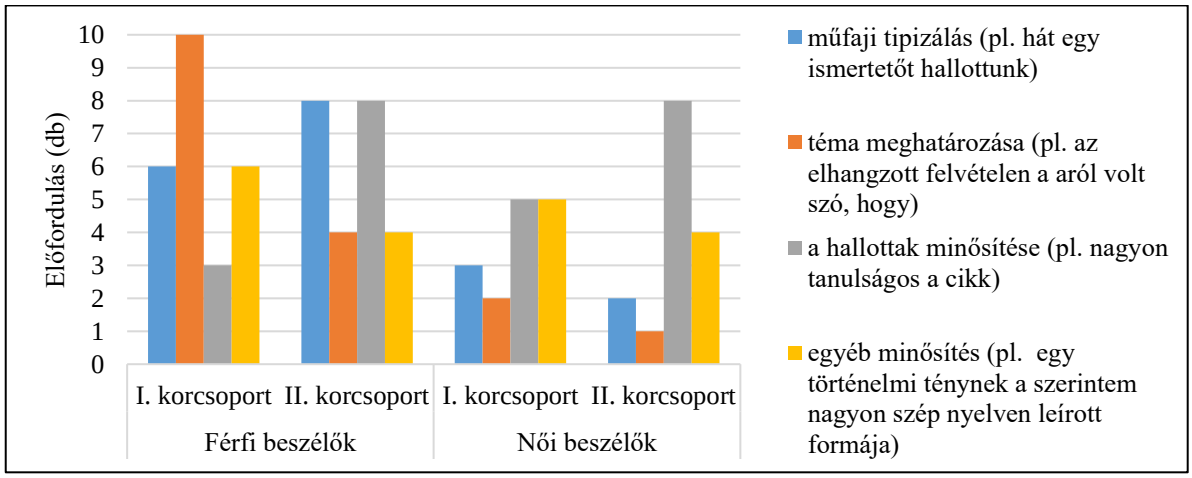
	Nők		Férfiak	
	I. 29–40 év	II. 41–59 év	I. 29–40 év	II. 41–59 év
érettségi	3	4	3	4
főiskola/egyetem	17	16	17	16
beszélők összesen	20	20	20	20

1. táblázat: Beszélői adatok

A vizsgálatokhoz a *Növény* és a *Varkocs* összefoglalókat (tartalomösszegzéseket) választottuk, mely fél-spontán beszédfeladatnak vagy irányított spontán beszédnek felel meg. A beszédleiratok gépi leiratozóval készültek (https://www.speechmatics.com), majd humán ellenőrzésen estek át. A beszédleiratokat az NBSZ gépi szövegfeldolgozójával (SZIRA, [12]) és manuális úton dolgoztuk fel.

A vizsgálódásainkban a női és férfi nyelvhasználati jegyekre vonatkozó, korábbi szakirodalmi megállapításokra és a nemekkel kapcsolatos sztereotípiákra összpontosítunk, a cél azon beszélt nyelvi markerek azonosítása, melyek a nyelvi profilalkotás gyakorlatában segíthetik a nyelvészeket. Vizsgáltuk a megnyilatkozások kezdetét a beszédtervezés elemzéséhez; az emocionális töltetet a 8 érzelemhez – harag, bizonytalanság, undor, bánat, félelem, öröm, szeretet, meglepettség – rendelt emóciószótár alapján; a nyelvi bizonytalanságot; az ún. puhító, óvatoskodó kifejezéseket (hedge), a szófaji lefedettséget (pl. melléknév), konkrét szóelőfordulásokat (pl. *érdekes, sajnos*), a szinonimákat stb.

Az eredmények alapján például a női beszélők kétszer olyan gyakran használták a ***nem tud-*** kifejezést, mint a férfiak (pl. *nem tudom, mi történt, teljesen elgondolkodtam; onnantól kikapcsoltam, nem tudtam követni*); a beszélt nyelvi jegyek közül a **szóismétlés** mint megakadásjelenség (pl. *tehát hogy hogy; egy egy közösségnek*) a férfiak nyelvhasználatára jobban jellemző témától függetlenül (előfordulás átlagosan: 3,3 és 1,72 db/beszélő). A férfi és női beszélőknél pregnáns eltérés mutatható ki a **bevezető gondolat** meglétében (pl. *Tehát az elhangzott bejátszás arról szólt, hogy*; *Na, hát az elhangzott szövegben arról kaptunk egy híradást hogy*), ezt a férfi beszélők 80–85%-a alkalmazta (az arány a női beszélőknél 45–65%). A bevezetés tartalmi vizsgálata azt mutatja, hogy a 40 év alatti férfi beszélők feladatorientáltak, a női beszélők pedig életkortól – és feladatszabástól – függetlenül véleményezik az elhangzottakat (1. ábra).



1. ábra: A bevezetés tartalmai elemeinek megoszlása

Az eredményeink azt mutatják, hogy a morfológiai és szemantikai jellemzők kevésbé hordoznak meggyőző és konzekvens, a nemekre jellemző sajátosságot a korpusz női és férfi beszélőinél, a pragmatikai sajátosságokban azonban mutatkoznak különbségek a nemek nyelvi viselkedésében.

Köszönetnyilvánítás
A kutatást az NKFIH az FK128814-es és az Európai Unió az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében támogatta.

Irodalom

[1] Ford, J. (2021). *Gender Disguise and Linguistic Identity Performance in Online Writings: Production, Perception, and Forensic Applications*. PHD thesis, Aston University.
[2] Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T. E., Beke, A., Neuberger, T. & Nikléczy, P. (2012) BEA: Beszélt nyelvi adatbázis. In.: Gósy M. (szerk.) *Beszéd, adatbázis, kutatások*. Akadémiai Kiadó. 9–24. old.
[3] Koppel, M., Argamon, S. & Shimoni, A.R. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, Volume 17, Issue 4, 401–412. old. https://doi.org/10.1093/llc/17.4.401
[4] Leemann, A., Perkins, R., Buker, G.S. & Foulkes, P. (2025). *An Introduction to Forensic Phonetics and Forensic Linguistics.* Routledge, London and New York.
[5] Litvinova, T., Seredin, P.V., Litvinova, O. & Zagorovskaya, O. (2018). Identification of gender of the author of a written text using topic-independent features. *Pertanika Journals Social Science & Humanities.* Vol. 26, Issue 1: 103–112. old.
[6] Markó, A., Főző, E. & Gráczi, T. E. (2022). Alkalmazható-e a diskurzusjelölők vizsgálata a beszélői profilalkotásban. In.: Markó A.–Deme, A. (szerk.). *Alkalmazott Nyelvtudomány különszám: Alkalmazott Nyelvészet és kriminalisztika*. 73–91. old.
[7] Nagy, F., Szakácsné, F.J. & Vágóné, M.I. (1983). Nőies és férfias szövegsajátosságok. In.: Rácz E.– Szathmári I. (szerk.). *Tanulmányok a mai magyar nyelv szövegtana köréből*. Budapest. 239–248. old.
[8] Skarnitzl, R., Vaňková, J. (2017). Fundamental frequency statistics for male speakers of common Czech. *Acta Universitatis Carolinae: Philologica*, 2017/3, 7–17. old.
[9] Steigervald, K. (2022). *Generációk harca. Hogyan értsük meg egymást?* Partvonal Kiadó, Budapest.
[10] Szabó, M.K., Lázár, B., Nyíri, Zs. & Morvay, G. (2017). A negatív emotív elemek vizsgálata a nemek közötti nyelvhasználati különbségek szempontjából. *XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia*. 161-176. old.
[11] Tatár, Z. (2013). Beszélőprofil-alkotás lehetőségei a kriminalisztikai fonetikában. *Alkalmazott Nyelvtudomány*, XIII/1–2, 121–130. old.
[12] Vincze, V., Főző, E., Kicsi, A. & Vidács, L. (2022). SZIRA: Szövegfeldolgozó Információs Rendszer és Adattár a szerzőazonosítás szolgálatában. In.: Markó A.–Deme, A. (szerk.). *Alkalmazott Nyelvtudomány különszám: Alkalmazott Nyelvészet és kriminalisztika*. 52–74. old.

How positive is VOT in Hungarian?

Zoltán G. Kiss (gkiss.zoltan@btk.elte.hu)
ELTE Eötvös Loránd University, Budapest

Voice Onset Time (VOT) has been a staple of phonetic and phonological research ever since the publication of Lisker & Abramson [4]. Already in this pioneering work, Hungarian was included as an illustration of how VOT can be used as a quantifiable measure of laryngeal contrast in stops, in this respect, this work provides the very first VOT data on Hungarian. VOT is such a phonetic/phonological commonplace [1] that anyone doing research on the laryngeal aspects of stops can certainly feel that they can reach back to the well-established sources, and quote their results when issues such as classifying stop systems arise, without the need of doing new experimental work: the work seems to have been already done, if a VOT number is required, then surely, that number is already available in the literature. In this paper, we will argue that this may not be completely the case for Hungarian.

The aim of the paper is to provide a systematic survey of papers analysing the VOT of the voiceless stops /p t c k/ in Hungarian, therefore, focussing is on *positive* VOT. 17 papers have been reviewed, yielding altogether 31 different analyses of the voiceless stops spanning 60 years (1964–2024). Variables examined include: VOT of the Hungarian voiceless stops; place of the stop; year; measurement methodology concerning the beginning and end of the VOT interval; consideration of multiple bursts in the measurement; number of participants; use of repetitions; environment of the stops (initial vs. intervocalic, etc.); type of the sound following the stop (vowel vs. sonorant); type of test words (natural vs. nonsense); number of syllables of test words; segmentation method (manual vs. automatic); statistical test type; statistical significance of result. It needs to be stressed that a proper meta-analysis of the literature was not possible as in many cases – which we will argue to be one of the problematic issues of past literature on Hungarian VOT – the standard deviation and/or standard error is not quoted, or only few, sometimes only a single, participant was used in the experiment. Nevertheless, the data provide valuable insights into VOT values in Hungarian over the past 60 years.

The literature reviewed shows that VOT of the voiceless stops increases as we move from anterior to posterior places, as in many other languages (Fig. 1, Table 1). Most studies report a statistically significant difference in VOT between the anterior stops /p t/ and the posterior ones /c k/. However, the mean VOT of /c/ is the highest, higher than that of /k/, although this result is based on few studies and is therefore less reliable. Most studies did not even consider /c/, reasons for its exclusion is not even discussed in the papers. If we set the upper limit of VOT at around 35 ms for stops in a "non-aspirating" language (as done in e.g., [2]), then the anteriors /p t/ are within this limit, while the posteriors /c k/ are beyond it, even though it can be demonstrated that this is still something non-aspirating languages display for posterior voiceless stops. However, we do see an elongated distribution upwards, into the 50–70 ms range for /k/, resembling aspirating languages (like English). Looking at the values over the years may shed some light on this (Fig. 2). Relatively high values, especially for /k/ were published in 2024, an unexpected jump from the previous 60 years. A possible interpretation is that Hungarian might have taken the first step towards being an "aspirating" language. Instead of this reading, we suggest that methodological factors might be behind the higher values, especially regarding the inconsistent

measuring of VOT in the case of multiple release bursts, which are often present in posteriors [3]. Most analyses are inexplicit about the methodology of the measurement of VOT, regarding both the start and the end of the measurement interval. Papers that *are* explicit about this mostly say they measured it from the "release". Overall, over 60% do not mention multiple release bursts at all, and so it is unclear if there were any, and if there were, from which burst they measured the VOT (the last?, the first?, the strongest?). The duration of the multiple-burst interval can be relatively long, and this may influence the overall VOT values. As far as the end of the VOT interval is concerned, the majority of the analyses measured it until the appearance of vocal fold vibration, only few papers considered the role of F2 in this respect, and around 20% of the analyses are unclear about what end point they opted for at all.

The literature review also indicates how difficult it is to provide an overall, general picture of VOT in Hungarian voiceless stops as they use rather disparate test designs, measure VOT in different environments, and ignore certain environments completely. For example, there is no analysis of VOT in onset clusters, all papers only considered prevocalic single onsets. 20% of the analyses relied on simple word lists read out by the participants, some only used single syllables for their measurements, and a third of the papers relied on nonce words. Spontaneous natural speech data were involved in only 32% of the analyses reviewed. As far as the number of participants is concerned, most papers relied on only five speakers, only four used more than ten participants. The low number of participants may be counterbalanced by repetitions, but again, only a quarter of the analyses reviewed employed this, and actually, most papers do not state whether repetitions were included, and so we can only assume they were not. Potential repeated measurements also have consequences for the statistical method chosen for the analysis, and so stating this aspect of the analysis should be important. Furthermore, interestingly, no paper used automatic segmentation methods, either alone or in tandem with manual methods, all employed manual inspection of waveforms and spectrograms in specifying VOT intervals.

VOT plays a fundamental role in laryngeal phonetics and phonology, making reliable and replicable data essential. This review highlights the need for a standardized measurement protocol – one that accounts for multiple release bursts, includes a wider range of environments, incorporates natural-speech data, and relies on larger participant samples. This would allow us to answer the question, "How positive is VOT in Hungarian?" more reliably.

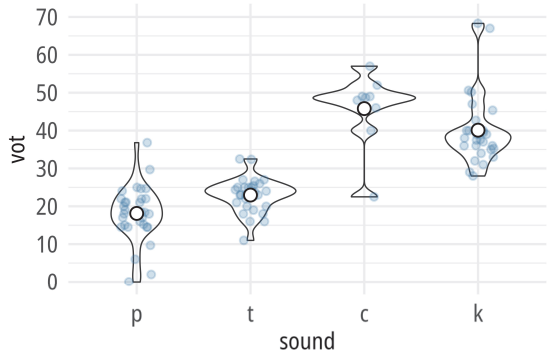


Figure 1: Distribution of VOT (ms) of the four voiceless stops in Hungarian.
White circles in the plots show the means.

Table 1. Descriptive statistics of VOT (ms) for the four voiceless stops in Hungarian

Sound	*N*	Mean	SD	Min.	Max.
/p/	31	18.12	7.32	0	36.8
/t/	31	22.97	4.43	11.0	32.5
/c/	9	45.80	9.84	22.5	57.0
/k/	29	40.10	9.35	28.0	68.3

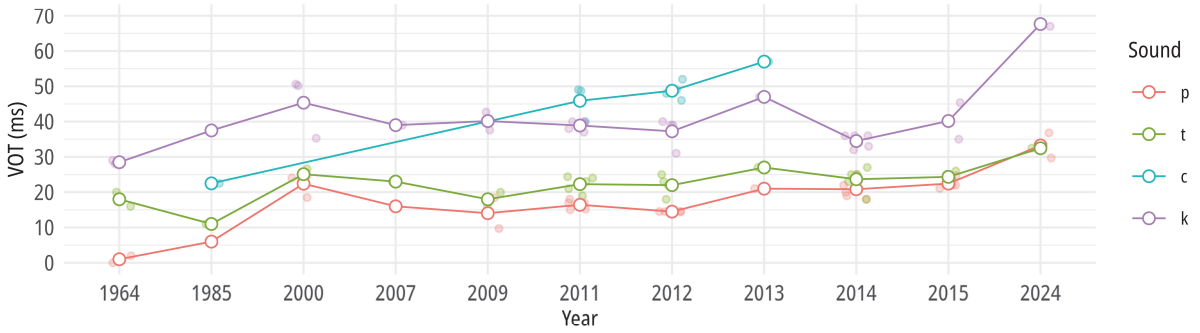


Figure 2: VOT (ms) of the four voiceless stops in Hungarian between 1964–2024,
based on the literature reviewed. White circles are the means.

References

[1] Abramson, Arthur S. & Douglas H. Whalen (2017). 'Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions.' In: *Journal of Phonetics* 63, pp. 75–86. DOI: https://doi.org/10.1016/j.wocn.2017.05.002.

[2] Keating, Patricia (1984). 'Phonetic and phonological representation of stop consonant voicing.' In: *Language* 60, pp. 286–319. DOI: https://doi.org/10.2307/413642.

[3] Keating, Patricia, John Westbury & Kenneth N. Stevens (1980). 'Mechanisms of stop consonant release for different places of articulation'. In: *Journal of the Acoustical Society of America* 67, pp. 1–18. DOI: https://doi.org/10.1121/1.2018489.

[4] Lisker, Leigh & Arthur Abramson (1964). 'A cross-language study of voicing in initial stops: Acoustical measurements'. In: *Word* 20, pp. 384–422. DOI: https://doi.org/10.1080/00437956.1964.11659830

Beszédtípusok és beszélők az 1933-ban készült magyar filmhíradókban

Gocsál Ákos
Pécsi Tudományegyetem Művészeti Kar
ELTE Nyelvtudományi Kutatóközpont

Bevezetés

Magyarországon 1931 szeptemberében jelent meg a hangosfilm, amit általában médiatörténeti mérföldkőnek tekintenek, ám azzal, hogy a filmszalag ekkortól hangot, beszédet is rögzített, a filmek a beszédtudomány számára is új kutatási lehetőségeket nyitnak. A hangosfilm a beszédhelyzetek képi ábrázolásával lehetővé teszi azok tágabb kontextusban történő vizsgálatát, s így a filmeken látható nyelvi jelenségek is pontosabban értelmezhetők. Ennek ellenére a hangos filmek, és ezen belül a híradók beszédanyagával kevesen foglalkoznak. A kutatók egy része nyelvészeti elemzést végzett, ismeretesek eredmények például filmhíradók beszédének prozódiájáról [1], illetve személyes névmások használatáról [7]. Mások történeti vagy médiatudományi kontextusban említik a beszédet, például a propagandában betöltött szerepét [8] vagy a hírolvasó hangjának változásait a politikai változások függvényében [3]. További kutatók módszertani és technikai kérdésekkel, például a filmhíradók automatizált annotálásával [4] vagy metaadatok kinyerésével, adatbázisok építésével [6] foglalkoznak. A magyar filmhíradókban rögzített beszédről is állnak már rendelkezésre adatok, azonban ezek a felvételek nyelvi-kommunikációs szempontból jelentős részben még feltáratlanok. Úgy véljük, a híradók beszédjelenségeinek részletesebb kutatásához elengedhetetlen a beszédanyag egészének általános jellemzése a beszédesemények kvantitatív és kvalitatív leírásával. A kvantitatív indikátorok előfordulási gyakoriságot jelentenek, a kvalitatív indikátorok pedig nyelvészetileg releváns kategóriák, amelyeket vagy korábban leírt elméletek alapján, vagy tartalomelemzéssel létrehozott, saját szempontok alapján állítunk fel. Jelen tanulmányunk célja az eddig rendelkezésre álló [2], 1931-ben és 1932-ben készült magyar filmhíradók jellemzésének kiegészítése az 1933-as filmhíradók beszédeseményeinek kvantitatív és kvalitatív leírásával. A kvalitatív jellemzés során kinyert kategóriákat metaadatként külön adatbázisban rögzíthetjük. Ezek segítségével további kutatásokhoz azonosíthatókká válnak a kutató számára egyébként rendezetlenül rendelkezésre álló filmhírek.

Anyag és módszer

A kutatást 1933-ban készült filmhíradókon végeztük, ezek elérhetők a www.filmhiradokonline.hu oldalon és kutatási célra felhasználhatók. Először meghatároztuk a beszédet tartalmazó híradók és az azokon belüli filmhírek számát, majd azonosítottuk a filmhírekben megfigyelhető közlési egységeket. Közlési egységnek tekintettük egy filmhíren belül ugyanazon beszélőtől származó, ugyanazon beszédtípushoz tartozó megnyilatkozásokat [2]. A közlési egységeket a Wacha, illetve Krepsz alapján, a spontaneitás mértéke szerint meghatározott kategóriák szerint végeztük, felolvasásokat, reproduktív, fél-reproduktív és spontán megnyilatkozásokat kerestünk [5, 9]. Mivel a fél-reproduktivitás (megtanult szöveg kötetlen elmondása) utólag nem állapítható meg, ennek a kategóriának az alkalmazásától eltekintettünk. Az egyes kategóriákon belül további, egyedi beszédtípusokat kerestünk a tartalomelemzés módszerével. Egyes beszédtípusok több tágabb kategóriához is tartozhattak. Például a vezényszavakat reproduktívnak tekintettük, ha a helyzet kiszámítható, ismétlődő, begyakorolt volt, de spontán beszédnek, ha a helyzet folyamatosan változott, pl. sportesemény

során. Ahol lehetett, azonosítottuk a beszélő személyeket is a rendelkezésre álló adatok alapján. A feltárt kategóriákat és beszélői adatokat a filmhírek más adataival együtt Excel-táblában rögzítettük, így a későbbiekben ezek alapján visszakereshetők, azonosíthatók a filmhírek.

Eredmények

A vizsgált időszakban 51 filmhíradó és 392 filmhír készült, ezek közül 44 híradóban 370 egyedi filmhír maradt fenn. 127 filmhír tartalmaz beszédanyagot, amelyek közül 99 magyar nyelvű, a többi idegen, a leggyakrabban német, angol, francia nyelvű. A beszédanyagban a fenti eljárással összesen 134 közlési egységet adatoltunk, majd kategorizáltuk beszédtípusok szerint. Az alábbiakban a korábbi kutatásokból rendelkezésünkre álló, 1932-ben készült filmek adataival együtt közöljük az új adatokat, így az előző teljes évhez viszonyítva tendenciák is megfogalmazhatók. Az 1. táblázat a közlési egységek számát mutatja. Ebből megállapítható, hogy az előző évvel összehasonlítva lényegesen kevesebb közlést találtunk, a visszaesés különösen a spontán beszédesemények számának csökkenésével magyarázható. A 2. táblázat részletezi a leggyakrabban előfordult, spontán beszédként kategorizált beszédtípusokat. Minden esetben jelentős csökkenést figyelhetünk meg. A felolvasások száma azonban 1933-ban növekedett, ez a narrációs szövegek gyakoribb előfordulásából adódik (3. táblázat). Az azonosított személyek száma is csökkenő tendenciára enged következtetni. 1932-ben 43 közlési egységben tudtuk meghatározni a beszélőt, ez 33 különböző beszélőt jelentett, míg 1933-ban 30 közlési egységben 22 különböző beszélőt találtunk. A leggyakrabban országos és önkormányzati politikusok jelentek meg (Gömbös Gyula, Eckhardt Tibor, Fabinyi Tihamér), ritkábban művészek, sportolók vagy más közéleti szereplők (Herczeg Ferenc, Piller György).

Következtetések

A feltárt adatok azt mutatják, hogy 1933-ban az előző évvel összehasonlítva kevesebb beszédeseményt mutatott be a híradó. A visszaesés főleg a spontán helyzetek, így a társalgások és a szónoklatok számában figyelhető meg. Emelkedett a narrációs szövegek száma, ami arra utalhat, hogy a korabeli alkotók egyre tudatosabban szerkesztették a híradókat, azaz a helyszínen rögzített hangok helyett nagyobb hangsúllyal szerepeltek az utólag megírt kommentárok. Ennek igazolására a következő évek beszédanyagát is vizsgálni kell. Ugyanakkor még 1933-ban is számos spontánnak tekinthető beszédhelyzetet rögzítettek az alkotók, ami lehetővé teszi a korabeli nyelvhasználat tanulmányozását autentikus környezetben. A beszélő személyek azonosításával pedig közel száz év távlatában beszélők közötti vagy beszélőn belüli különbségek, és akár longitudinális változások is kutathatók.

1. táblázat. A közlési egységek számának eloszlása tágabb kategóriák szerint

	spontán	reproduktív	felolvasás
1932	130	16	61
1933	58	10	66

2. táblázat. A spontán beszédként megvalósult, gyakoribb beszédtípusok előfordulása

	társalgás	szónoklat	vezényszavak	bejelentések	interjú	bekiabálás
1932	41	29	17	16	9	6
1933	8	17	3	3	4	3

3. táblázat: A felolvasásként kategorizált beszédtípusok előfordulása

	narráció	szónoklat	előadás	bejelentés	rádióbeszéd, mikrofonba beszéd
1932	48	6	6	1	0
1933	60	3	0	0	3

Irodalom

[1] de Mareüil, B. P., Rilliard, A., & Allauzen, A. (2012). A Diachronic Study of Initial Stress and other Prosodic Features in the French News Announcer Style: Corpus-based Measurements and Perceptual Experiments. *Language and Speech*, *55*(2), 263–293. https://doi.org/10.1177/0023830911417799

[2] Gocsál, Á. (2022). Beszédtípusok és artikulációs tempóik a korai hangos filmhíradókban. In L. Csárdás & J. Bóna (Szerk.), *Sokszínű beszédtudomány*. Akadémiai Kiadó. https://mersz.hu/csardas-bona-sokszinu-beszedtudomay

[3] Imesch, K., Schade, S., & Sieber, S. (2016). Introduction: Constructions of Cultural Identities in Newsreel Cinema and Television after 1945. In K. Imesch, S. Schade, & S. Sieber (Szerk.), *MedienAnalysen* (1. kiad., Köt. 17, o. 7–20). transcript Verlag. https://doi.org/10.14361/9783839429754-001

[4] Koržinek, D., Wołk, K., Brocki, Ł., & Marasek, K. (2019). Automatic transcription of the Polish newsreel. *Poznan Studies in Contemporary Linguistics*, *55*(2), 183–209. https://doi.org/10.1515/psicl-2019-0008

[5] Krepsz, V. (2016). Fonetikai hasonlóságok és különbözőségek a beszédtípusokban. In J. Bóna (Szerk.), *Fonetikai olvasókönyv* (o. 175–188). ELTE Fonetikai Tanszék. http://fonetikaitanszek.elte.hu/wp-content/uploads/2017/11/OlvasokonyvTeljes.pdf

[6] Oiva, M., Mukhina, K., Zemaityte, V., Karjus, A., Tamm, M., Ohm, T., Mets, M., Chávez Heras, D., Canet Sola, M., Juht, H. H., & Schich, M. (2024). A framework for the analysis of historical newsreels. *Humanities and Social Sciences Communications*, *11*(1), 530. https://doi.org/10.1057/s41599-024-02886-w

[7] Pozharliev, L. (2017). Newsreels from 1968 Communist Bulgaria: The Encompassing Us vs. The Different Them. *On_Culture: The Open Journal for the Study of Culture*, *4*, 1–25. http://geb.uni-giessen.de/geb/volltexte/2017/13390/

[8] Spohlberg, M. (2014). The din of gunfire: Rethinking the role of sound in World War II newsreels. *European Journal of Media Studies*, *3*(2), 113–129.

[9] Wacha, I. (1974). Az elhangzó beszéd főbb akusztikus stíluskategóriáiról. *Általános Nyelvészeti Tanulmányok*, *X*, 203–216.

Alveoláris réshangok zöngéssége Down-szindrómában

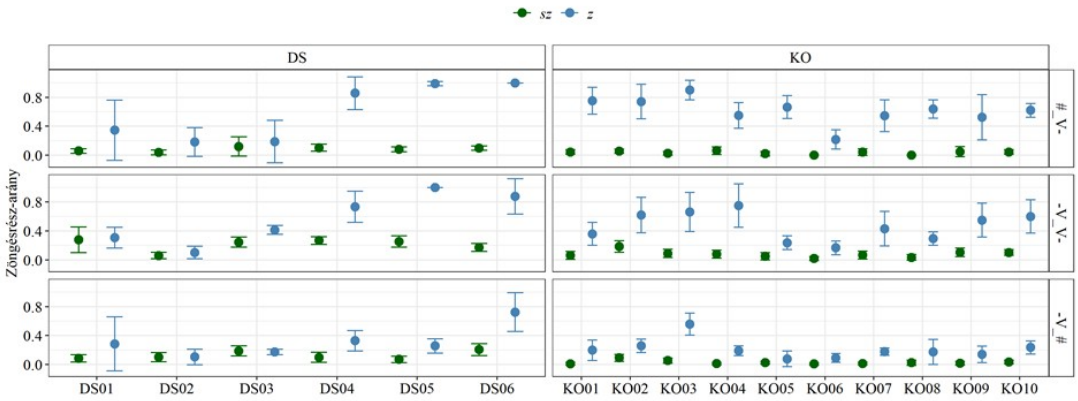
Gráczi Tekla Etelka[1, 2], Szánthó Zsuzsa[3, 1, 2] és Deme Andrea[3, 1, 2]
[1] ELTE Nyelvtudományi Kutatóközpont
[2] MTA–HUN-REN NYTK Lendület Neurofonetikai Kutatócsoport
[3] Eötvös Loránd Tudományegyetem

1. Bevezetés. A Down-szindróma (DS) olyan genetikai kondíció, amely atipikus intellektuális, neurológiai és fizikai fejlődéssel jár [1]. A DS-val diagnosztizált beszélők beszéde gyakran alig érthető, pragmatikai, grammatikai és fonológiai tekintetben atipikus mintázatok jellemzik [7]. DS-ban nem csak az intellektuális és neurológiai, de a fiziológiás (pl. gége és orofaciális terület) atipikus fejlődése is hozzájárul a szegmentumok ejtésének és akusztikai realizációinak sajátosságaihoz. Jellemző például a mássalhangzók zöngétlenedése [3], ami bizonyos mértékben tipikus fejlődésű felnőttekben is jellemző (a magyarban pl. [2, 5]). Fontos hangsúlyozni, hogy a DS-val diagnosztizált beszélők között fonetikai és fonológiai tekintetben is jelentős variabilitást mutattak ki [pl. 6], és a zöngésség fonológiai kontrasztját eddig csak szubjektív, észleleti minősítés alapján végezték el. Felmerül a kérdés, hogy akusztikai értelemben milyen jellegű zöngétlenedés tapasztalható ebben a beszélői csoportban: 1) A mássalhangzó képzése alatt milyen mértékben van jelen a zönge, és ez alapján elkülönülnek-e a zöngésségi párok az egyes DS-val diagnosztizált beszélők ejtésében? 2) A zöngés hangokat zöngétlenül ejtő beszélők esetében a zöngésség másodlagos kulcsai, így a mássalhangzó tartama elkülönítik-e a kontrasztban álló hangzókat, vagy a zöngétlenítésnek a tipikus fejlődésű beszélőktől eltérhető mintázatai a kontraszt teljes neutralizációjához vezetnek? A jelen esettanulmányban ezeket a kérdéseket a /z/ és /s/ alveoláris réshangok elemzésével vizsgáljuk DS-val diagnosztizált és tipikus fejlődésű felnőtt beszélők ejtésében.

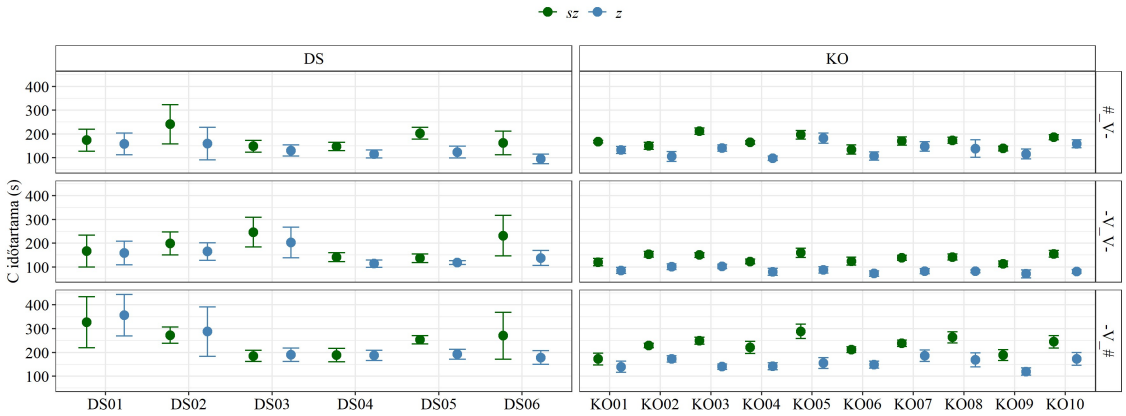
2. Módszertan. 6 DS-val diagnosztizált női beszélő (DS01–DS06, 23–40 év) és 10 tipikus fejlődésű női beszélő (CO01–CO10, 19–23 év) vett részt a kísérletben. A /z/ és /s/ hangzókat #_V-, -V_V- és -V_# pozícióban olvasták fel minimális párt alkotó szavakban, minimális párt nem alkotó szavakban, valamint „minimális párt” alkotó álszókban. Az itemeket önálló mondatokként jelenítettük meg. A létező szavak esetében egy, a szót ábrázoló képet, a logatomok esetében egy (minden esetben azonos) absztrakt képet mutattunk a résztvevőknek. Az 54 célszó mellett 46 disztraktort is tartalmazott a lista, melyet háromszori ismétléssel, random sorrendben olvastak fel. A felvételeket csendesített szobában, fejmikrofonnal készítettük. A felvétel végén egy kérdőívben a beszédet érintő kérdéseket is feltettünk a beszélőknek. A hanganyagokat Praatban [4] dolgoztuk fel: automatikus címkézés után kézi korrekciót alkalmaztunk. Mértük a hordozó szó, a mássalhangzó és az esetleges megelőző magánhangzó időtartamát, valamint a mássalhangzó ejtésében a zöngétlen rész arányát a *voice report* funkcióval (500 Hz-ig aluláteresztő szűrő, f_0 érték: 100–300 Hz-es tartomány). A további beállításokban a sztenderd értéket tartottuk meg. A kapott értékeket lineáris kevert modellekkel elemeztük A zöngétlen rész arányát a zöngés rész arányára váltottuk, a 0 és 1 értékeket 0,001-gyel növeltük/csökkentettük, hogy bétaeloszlásúként kezelhessük az adatokat a modellben. A DS csoportban 0–22 db szót olvastak félre a beszélők (2 fő 0-t, 1-1 fő 1-et, 4-et, 5-öt, 1 fő 22-t). A kontrollcsoport beszélőivel minden esetleges félreolvasást tudtunk javíttatni. Emiatt a kontrollcsoportban 972, a DS-csoportban 940 elemezhető itemet kaptunk.

3. Eredmények. A zöngésrész-arányra a csoport főhatás szignifikánsnak bizonyult, ahogyan a fonológiai zöngésség és a pozíció faktorok is, valamint ezek interakciója. Szókezdő és szóbelseji pozícióban szignifikáns eltérést találtunk a mássalhangzópár tagjai között mindkét csoportban, de szóvégen egyikben sem. A /z/ zöngésrész-aránya szóvégi pozícióban szignifikánsan alacsonyabb volt mindkét csoportban, mint a további, vizsgált helyzetekben. Az 1. ábrán azonban megfigyelhető, hogy a két csoport esetében eltért a beszélők közötti variancia mértéke. A DS csoportban 3 beszélő (DS01, DS02, DS03) esetében a két fonéma realizációi nem különböztek a zöngés rész arányában semmilyen pozícióban, tehát ezek a beszélők egyáltalán nem különítették el a kontrasztban álló mássalhangzókat. Ezzel szemben a másik három beszélő ejtésében csak szóvégi helyzetben találtuk ezt (a zöngés /z/ teljes zöngétlenedésével), szókezdő és intervokális helyzetben pedig változatos mintázatot láttunk. A hangzók időtartama a kontrollcsoport beszélőinél minden esetben, a DS-csoportban két főnél mutatta a várt mintázatokat, tehát a zöngés /z/ rövidebb tartamát a zöngétlen /s/-hez képest (2. ábra). A DS csoportban a nehézkesebb olvasás miatt a teljes szóra vetített időtartamot is elemeztük, de ugyanazon eredményt kaptuk. Lényegében két olyan beszélő esetében látható a kontroll csoportra jellemző időviszony-mintázat, akiknél a zöngésség is hasonlóan alakult (DS05, DS06).

4. Következtetések. Eredményeink szerint a Down-szindrómával diagnosztizált csoportban a zöngésségi oppozíció megtartása egyénfüggően alakult. A beszélők egy részére jelentős zöngétlenítés volt jellemző olyan pozíciókban is, ahol a kontroll csoportnál ez nem volt jellemző, valamint az időtartam-viszonyok csoport szinten is eltérő mintázatot mutattak a kontroll beszélőkben látottaktól. Feltételezhető tehát, hogy a DS-ban jellemzőnek tekintett zöngésségikontraszt-semlegesedés nagymértékben beszélőfüggő jelenség. Az időtartammintázatok esetében a DS-val diagnosztizált beszélők ejtésében további kérdésként merül fel, hogy a beszédritmus alakulása általában tér-e el a tipikus fejlődésű beszélőkétől, avagy csak a kontraszt fenntartásában nem játszik szerepet. A jövőbeli elemzésekben (további beszélők bevonása mellett) több hangzó bevonása és percepciós teszt végzése adhat támpontot a jelenség pontos leírására. Az előadásban az ingertípus hatására is kitérünk.



1. ábra. A zöngésrész-arány átlaga és szórása a pozíció és a zöngésség függvényében az egyes beszélők ejtésében a DS-val diagnosztizált (DS), és a kontrollcsoportban (KO)



2. ábra. A mássalhangzó időtartamának átlaga és szórása a pozíció és a zöngésség függvényében az egyes beszélők ejtésében a DS-val diagnosztizált (DS), és a kontrollcsoportban (KO)

Köszönetnyilvánítás. A kutatás részben a Bolyai János Kutatási Ösztöndíj (D. A.) és az EKÖP-24, EKÖP-25 (Sz. Zs.) támogatásával készült.

Irodalom

[1] Antonarakis, S. E. et al. (2021). Down syndrome. Nat Rev Dis Primers. 6(1), 9. doi:10.1038/s41572-019-0143-7.

[2] Bárkányi, Zs., and G. Kiss, Z. (2019). A fonetikai korrelátumok szerepe a zöngekontraszt fenntartásában. [Role of phonetic correlates in preserving the voicing contrast]. Beszédprodukciós és észleléses eredmények. Általános Nyelvészeti Tanulmányok, 31, 57–102.

[3] van Borsel, J. (1996). Articulation in Down's syndrome adolescents and adults. European Journal of Disorders of Communication, 31, 415-444.

[4] Boersma, P. & Weenink, D. (2025). Praat: doing phonetics by computer [Computer program]. Version 6.4.34, http://www.praat.org/. Retrieved 20 May 2025.

[5] Gráczi, T. E. (2012). Zörejhangok akusztikai fonetikai vizsgálata a zöngésségi oppozíció függvényében. [Acoustic phonetic analysis of obstruents regarding voicing opposition]. PhD-thesis. Budapest : ELTE Eötvös Loránd University.

[6] Karmiloff-Smith, A. et al. (2016). The importance of understanding individual differences in Down syndrome. F1000Research, 5, 389. https://doi.org/10.12688/f1000research.7506.1

[7] Kent, Ray D., and Vorperian, Houri K. (2021). Speech intelligibility and communication in Down syndrome: Intervention approaches and the role of speech supplementation across the lifespan. In K. Wilkinson, and L. H. Finestack (Eds.), Multi-modal AAC for individuals with Down syndrome across the lifespan (37–60). Brookes.

A fordulókezdés fonetikai eszköztára és beszélőváltás-időzítési szerepe 5, 9 és 13 éves gyermekeknél

Hámori Ágnes[1] – Bóna Judit[2]
[1]ELTE Nyelvtudományi Kutatóközpont, Budapest/ME BTK, Miskolc
[2]Eötvös Loránd Tudományegyetem, Budapest

1. Bevezetés

A társalgásbeli beszélőváltások temporális jellemzői, különösen a fordulók gyors és gördülékeny váltakozása kiemelt figyelmet kap a kortárs pszicholingvisztikai és társalgásfonetikai kutatásokban [8, 11]. Különböző nyelvek társalgási korpuszain végzett fonetikai kutatások azt mutatják, hogy a beszélőváltások legnagyobb része hasonló időtartam (FTO vagy floor transfer offset, az előző forduló vége és a következő forduló kezdete közötti néma szünet) alatt megy végbe, átlagosan +250 és -250 milliszekundum között ([3, 12, 7]). Ez meglepően rövid időtartam azon pszicholingvisztikai adatoknak a fényében, melyek szerint a beszédprodukció megkezdése egy szó előhívása esetén legalább 600, többszavas megnyilatkozás esetén mintegy 1500 ms-nyi időt igényel ms [10]; mindez egy összetett pszicholingvisztikai mechanizmus működésére utal, amely magában foglalja a beszédpartner fordulója végének előre jóslását, a saját válasz megtervezését, a partner beszédében a beszélőváltási pont azonosítását, majd a válasz meghangosításának megindítását. Ez a rendívül összetett mechanizmus és a gyors, gördülékeny váltások nagy száma arra utal, hogy a precíz fordulóidőzítésnek és a beszélőváltás szűk időablakának fontos szerepe van az emberi nyelvben.

Lényeges kérdés, hogyan ez a komplex beszélőváltási mechanizmus hogyan alakul ki a gyermeknél a nyelvi fejlődés során [6]. 1 éves kor körül, a beszéd indulásakor a gyermekek átlagos válaszideje jóval hosszabb a felnőttekénél (1200 ms körül), majd az életkor előrehaladtával fokozatosan csökken, és 5-6 éves korra megközelíti a felnőtteknél mért átlagos értéket [2], azonban még 9 éves korra sem éri el teljesen [5]. A szóátváteli időtartam csökkenése szorosan összefügg az életkorral járó nyelvi és kognitív fejlődéssel, de ez nem egyszerű lineáris csökkenés: a váltás gyorsaságát a beszélő kora mellett számos tényező befolyásolja (pl. a beszélgetőpartnerek száma, kora, a társalgástípus, az előző forduló vagy az új forduló hossza), valamint ezek kombinációja. A kutatásokból továbbá az is látszik, hogy a gyermekek beszélőváltásainak nagy része a felnőtteknél szokásos normatív időkeretek közt marad, és viszonylag ritka a hosszú szünet, valamint a közbevágás vagy az egyszerre beszélés, ami a szociokulturális normák követésére való törekvésre utal.

Néhány kutatás ([2, 5]) azt mutatja, hogy a gyermekeknél megjelennek jellegzetes hezitációs technikák fordulókezdéskor: ezek a diszfluencia-jelenségek osztályozásaiban a bizonytalanságra visszavezethető megakadásokhoz [4] vagy a hezitációs jelenségekhez állnak közel, azonban helyzetük és funkciójuk alapján egy sajátos csoportot alkotnak, mivel nem beszédprodukció közben, hanem annak megkezdése előtt jönnek létre, és funkciójuk kifejezetten a szóátvétel jelzése és a beszédszándék kifejezése [7]. Ezek kapcsán [2] a diskurzusjelölőként használt kötőszókat vizsgálta 2-3 éves gyermekek társalgásaiban, Hámori és Bóna [5] a kitöltött szünetek és a diskurzusjelölők szerepét mutatta be. Eredményeik azt is feltárják, hogy ezeknek az eszközöknek a használata és típusa is összefügg az életkorral. A gyermeknyelvi diszfluencia- jelenségek kutatásaiból is tudjuk, hogy a nyelvi-kognitív fejlődés során a megakadások és hezitációs stratégiák is változnak (13, 1, 9]), „hezitálni is meg kell tanulni", illetve az életkorral a gyakoriság mellett a stratégiák típusainak aránya is megváltozik, ezért érdemes ezeket a fordulókezdő helyzetben is vizsgálni.
A jelen kutatás célja, hogy átfogó jelleggel feltárja a fordulóváltáskor használt hezitációs

eszközöket 5, 9 és 13 éves korú magyar gyermekek társalgásában, különös tekintettel az eddig nem vizsgált jelenségekre és az életkorral összefüggő jellegzetességekre.

2. Anyag és módszer

A kutatás során 30 gyermek-felnőtt beszélgetést vizsgáltunk a GABI adatbázisból, ebből 12 ötéves, 12 kilencéves és 6 tizenhárom éves gyermek, fiúk és lányok egyaránt; mindegyikük ép hallású és tipikus beszédfejlődésű. Beszélgetésenként kb. 5 percnyi felvételt elemeztünk, az elemzett anyag összesen 140 perc. A hangfelvételeket a Praat szoftver segítségével annotáltuk, kézzel határoztuk meg a fordulók határait és a hezitációs jelenségek típusát. A fordulók időzítését az FTO-érték (az előző forduló vége és az új forduló kezdete közötti néma szünet időtartama, [3]) alapján mértük. Fordulókezdő hezitációnak tekintettünk formai alapon minden hezitációs jelenséget, amely a forduló első szavának első fonémája előtt hangzott el, így különösen a nyújtást, kitöltött szünetet, diskurzusjelölőket, a nevetést és az ismétlést. Nem soroltuk ide a hangos lélegzetvételt, valamint itt a diskurzusjelölők közé nem számítottuk be a mondat eleji kötőszókat.

3. Eredmények

Mind a három korosztályban találtunk fordulókezdő hezitációkat, ezek hossz, forma, és jelleg szerint különböztek. Az 5 éveseknél átlagban a fordulók 27%-a kezdődött így, a 9 éveseknél a fordulók 40%-a. A fő típusok közül kiemelhetők az alábbiak: nyújtás (a megnyilatkozás első hangjának megnyújtása); nyújtás ismétléssel (pl.: BF005F05 *ak akkor én nem játszottam semmit);* kitöltött szünet (többnyire *mm, aa* vagy *öö*); kitöltött szünetek társulása (*aaa mmm);* diskurzusjelölő (*hát*); diskurzusjelölő kitöltött szünettel (*hát öö*).

Életkor szerinti megoszlást is meg lehetett figyelni: az 5 éveseknél a nyújtás és a kitöltött szünet volt fordulókezdésen a leggyakoribb, ezek sokszor nem is válnak el világosan egymástól. A kitöltött szünet az 5 éveseknél többnyire *m* vagy *a*, a 9 éveseknél legtöbbször *ö* hanggal realizálódott. Fordulókezdő diskurzusjelölő az 5 éveseknél viszonylag ritka volt, és szinte mindig a *hát* állt ebben a szerepben. A 9 éveseknél már a diskurzusjelölők (*hát*) vannak nagyobb arányban, és a kitöltött szünet gyakorisága csökkent. A 13 éveseknél a fentiekhez képest jelentős eltéréseket találtunk: a fordulók túlnyomó része hezitációval kezdődött (átlag 80%), és ez elsősorban a *hát* és/vagy nyújtás jelenségével valósult meg. Funkciójuk szerint az 5 és 9 éveseknél ezek a hezitációs technikák elsősorban a beszélőváltás normatív időtartamához való illeszkedést szolgálták; ezt mutatja, hogy ebben a két korcsoportban ezek jellemzően a nagyobb kognitív terheléssel járó, hosszabb tervezést igénylő megnyilatkozások elején jelennek meg, míg a kognitív szempontból könnyebb beszédprodukciós fordulókban, pl. az egyszavas választ tartalmazó fordulókban, vagy az eldöntendő kérdésre adott rövid válaszokban nem szerepelnek. A 13 éveseknél ez az időzítés-illesztési funkció nem volt jellemző, hanem a fordulókezdő hezitációs elemek inkább diskurzusjelölőként – általános válaszjelölőként vagy attittűdjelölőként –, társas jelentéssel összekapcsolva (szerénység, bizonytalanság jelzése stb.) szerepelnek.

Szakirodalom

[1] Bóna J. (2019). Megakadáskapcsolatok és komplex megakadások óvodások és kisiskolások beszédében. In: Bóna Judit–Horváth Viktória (szerk.): Az anyanyelv-elsajátítás folyamata hároméves kor után. ELTE Eötvös Kiadó, Budapest. 259–272.

[2] Casillas, M. (2014). Taking the floor on time: Delay and deferral in children's turn taking. In Arnon, I., M. Casillas, C. Kurumada, & B. Estigarribia (eds.): *Language in Interaction*. Amsterdam: John Benjamins. 101–114.

[3] De Ruiter, J. P., H. Mitterer, N. J. Enfield (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* 82/3: 515–535.

[4] Gósy M. (2003). A spontán beszédben előforduló megakadásjelenségek gyakorisága és összefüggései. *Magyar Nyelvőr 127/3:* 257–277.

[5] Hámori Á. & J. Bóna (2023). A beszélőváltások időzítése 5 és 9 éves gyermekek társalgásában: a milliszekundumok változásai a nyelvi fejlődés során. In: Empirikus társalgáskutatás Magyarországon. HUN-REN Nyelvtudományi Kutatóközpont, Budapest, pp. 245–276.

[6] Hilbrink, E. E., M. Gattis, & S. C. Levinson (2015). Early developmental changes in the timing of turn-taking: a longitudinal study of mother–infant interaction. *Frontiers in Psychology* 6. doi.org/10.3389/fpsyg.2015.01492

[7] Horváth V. (2010). "Filled pauses in Hungarian: Their phonetic form and function." Acta Linguistica Hungarica (Since 2017 Acta Linguistica Academica) 57.2-3: 288-306.

[8] Horváth V., J. Bóna, CS. I. Dér, D. Gyarmathy, Á. Hámori, A. Huszár, V. Krepsz, & Zs. Weidl (2021). Fonetika és társalgáskutatás: a beszélőváltások kérdésköre. In Markó A. (szerk.): Tanulmányok a beszédtudományok alkalmazásainak köréből. Budapest: ELTE Eötvös Kiadó. 251–278.

[9] Horváth V. & V. Krepsz (2021). Filled pauses in children's spontaneous speech – aspects of timing and complexity. In Bóna J. (szerk.): *(Dis)fluencies in children's speech*. Akadémiai Kiadó, Budapest. Paper: m873dfics_19#m873dfics_19

[10] Indefrey, P. & W. J. M. Levelt (2004). The spatial and temporal signatures of word production components. Cognition 92: 101– 144.

[11] Levinson, S. C. (2016). Turn-taking in Human Communication – Origins and Implications for Language Processing. Trends in Cognitive Sciences 20/1: 6–14.

[12] Markó A. & M. Gósy (2015). A megszólalás stratégiái társalgásban. In: Bárdosi V. (szerk.): *A nyelvi pragmatika kérdései szinkrón és diakrón megközelítésben*. Tinta Könyvkiadó, Budapest, 159–68.

[13] Neuberger T. (2013). A spontán beszéd temporális sajátosságai 6–14 év közötti gyermekeknél. *Anyanyelv-pedagógia* 2013/2. https://www.anyanyelv-pedagogia.hu/cikkek.php?id=4517

Cognitive grouping: Key to Tiberian Hebrew Intonation

Laszlo Hunyadi
University of Debrecen

0. Summary of the talk

The aim of the talk is to contribute to the long-standing debate regarding the original purpose of the system of the 1400 years old Tiberian Masoretic accentuation of the Hebrew Bible representing a complete and highly detailed phonetic annotation, by extending research to an area hitherto having received less attention. It gives special significance to the centuries old suggestion that the system must have merited from direct observations of speech, with Dresher [1] first identifying prosody as the basis of Tiberian accentuation. We will show that the prosodic nature of this system includes the highest level of prosodic structure, i.e. intonation, and that the principal features of its underlying tunes can be captured.

With the obvious lack of direct audio evidence from the time of the Masoretes to validate the reconstruction of intonation, we will rely on an indirect evidence – based on the concept of cognitive grouping (cf. Hunyadi [2]). It will be shown that cognitive grouping, identified as being both universal and modality independent, can serve as an explanatory principle for any system consisting of groups (see (1)). As for speech, it serves as the basic foundation of speech prosody. Assuming that Masoretic accentuation represents speech, by applying these universal principles to the Masoretic grouping structure of a Biblical verse (see (2)), we can thus identify its the hypothetical intonation. We will conclude with presenting AI-generated audio examples of the application of cognitive grouping to the Masoretic annotation of the Biblical texts.

1. Some basic notes on the Tiberian Masoretic accentuation

The Hebrew Bible in its written form dates back to more than two millennia, with the oldest surviving (incomplete) samples found in the caves of Qumran in 1947. The oldest complete copy known today is the Leningrad Codex written in 1008 in Cairo; preceded in time only by the Aleppo Codex (Tiberias, around 920) which, however, was partially destroyed during riots in Aleppo in 1947. The writing of both Codexes is attributed to the Masorete Aaron ben Moses ben Asher from Tiberias, the most prominent member of the family of the grammarians Ben Asher.

The importance of the work of the Ben Asher grammarians was that they were the first who aimed at preserving the hundreds of years of oral tradition of reading the biblical texts (the Hebrew word *masora stands for* tradition) by inventing and applying signs to the originally consonantal text adding the vowels and various consonantal qualities (see examples (3) a-c).

This highly detailed, essentially phonetic transcription was accompanied by notes referring to eventual deviations in spelling of certain words found in other manuscripts, but also including lists of grammatical nature. In addition to the phonetic signs, the inventory of the Masoretes also included more than two dozens of further signs whose purpose has continued to be subject of debate for centuries. The main claims regarding their purpose (cf., Wickes [3], [4]) are:

– they represent phonetic accents (hence their accepted term: Masoretic accents)

Argument: the signs fall on the accented syllable.

However: (a) there are specific signs never falling on the accented syllable; (b) if a word consists of only one syllable, it is redundant to mark the place of its accent; (c) there are (a few) cases where a word has more than one accent marked on the same syllable.

– they represent syntax

Argument: in many cases a pair of accents combine words into a group, corresponding to our modern understanding of syntax.

However: there are many groups which defy the syntax we are aware of today.

– they represent music:

Argument: actually, in synagogal reading practice the signs are followed as musical signs and combine into melodies (hence their other term: cantillation signs)

However: (a) there are verses, which, depending on what occasion they are read (week or festive days), are performed with different melodies; (b) the very same sequences of signs receive different melodic representation in different regions or traditions.

We propose that the above issues regarding the purpose of the Masoretic accents can be resolved by assuming that they represent speech melody, i.e. intonation, allowing at the same time for the perceived phonetic (accentual) and syntactic ambiguity of their system.

2. The application of cognitive tonal grouping to the Masoretic annotation of the Biblical texts

The basic intonational properties of accents are proposed to be as follows:
(a) they represent either pauses or tones (rise, fall, level); (b) they are arranged along a global F0 trend-line, and their sequences are specific to a given pitch range segment of the trend-line.

Examples for the hypothetical reconstruction of Biblical intonation:

(4) declarative: ‘In the beginning God created the heavens and the earth.’ (Gen. 1:1)
(5) focus:‘And when his brothers saw that their father loved him more than any of his brothers’ (Gen. 37:4)
(6) contrastive topic: ‘for the Lord your God, he goes with you’ (Deut. 31:6)
(7) interrogative: ‘Why did You bring harm upon this people? Why did You send me?’ (Ex. 5:22)
(8) syntactic embedding: ‘And your heaven that is over your head shall be bronze’ (Deut. 28:23)

References

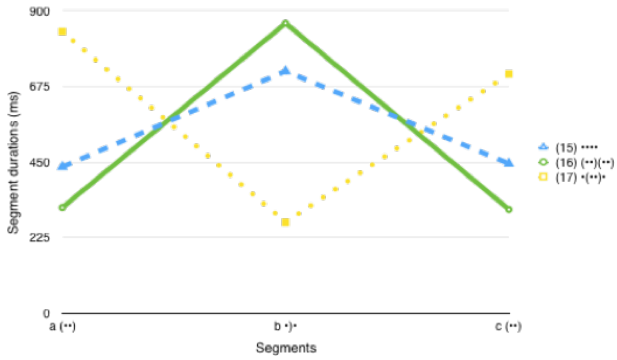
[1] Dresher, B. E. (1994). ‘The Prosodic Basis of the Tiberian Hebrew System of Accents’. *Language Vol. 70, No. 1*, pp. 1-52

[2]Hunyadi, L. (2010). ‘Cognitive grouping and recursion in prosody’. Harry van der Hulst (ed.), *Recursion and Human Language*. Berlin, New York: De Gruyter Mouton. pp. 343-370.

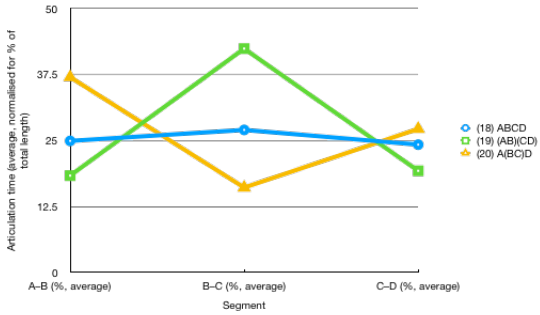
[3] Wickes, W. (1881). *A Treatise on the Accentuation of the three so-called Poetical Books on the Old Testament*. Oxford

[4] Wickes, W. (1887). *A Treatise on the Accentuation of the Twenty-one So-called Prose Books of the Old Testament*. Oxford

(1) temporal grouping of abstract visual vs. tonal patterns
Visual elements: subjects were asked to represent each visual pattern by mouse clicks
Tonal patterns: subjects were asked to read out the sequences of letters; timing was recorded:

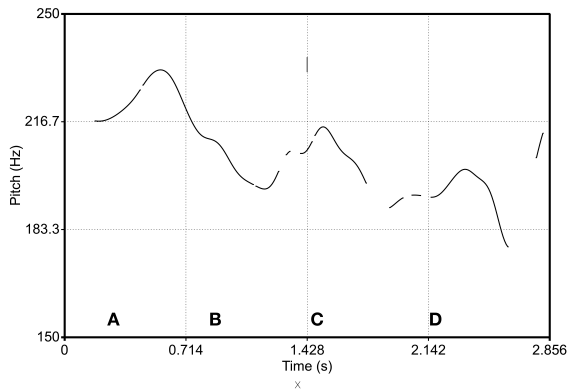


(1a) abstract visual patterns with dots

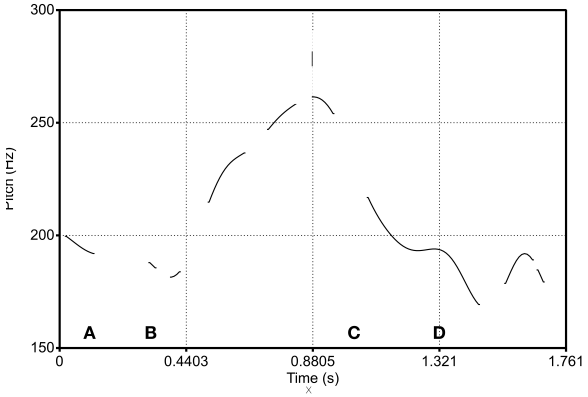


(1b) abstract tonal patterns with letters

(2) tonal grouping of abstract tonal patterns
The tonal representation of the recorded read-outs of sequences of letters was observed:



(2a) abstract tonal pattern (ABCD)



(2b) abstract tonal pattern (AB)(CD)

(3) ‘In the beginning God created the heavens and the earth.’ (Gen. 1:1)

(a) Plain consonantal text:

בראשית ברא אלהים את השמים ואת הארץ׃

(b) Vowels and consonantal features added (blue):

בְּרֵאשִׁית בָּרָא אֱלֹהִים אֵת הַשָּׁמַיִם וְאֵת הָאָרֶץ׃

(c) Accentuation added (red):

בְּרֵאשִׁית בָּרָא אֱלֹהִים אֵת הַשָּׁמַיִם וְאֵת הָאָרֶץ׃

The acoustic properties of Italian obstruents: Results from a recent experiment

Huszthy Bálint
ELTE BTK, Budapest

Background Our main research project, entitled "The Laryngeal Phonology of Italian", aims to reanalyse the obstruent system of Standard Italian within the frameworks of *Laryngeal Realism* and *Relativism* [3], based on *Laboratory Phonology* methods. Previous comprehensive works on Italian phonetics and phonology do not consider the relevance of laryngeal features in reviewing the consonantal system of Italian. However, recent findings in laryngeal phonology suggest that these features play a crucial role in the classification of obstruents and the related phonological phenomena [2].

Data As the first step in this project, we conducted detailed acoustic measurements of the obstruent system of Standard Italian, motivated by the limited amount of existing literature on the subject (cf. [4]). The initial results of this investigation are presented here. The research design involved the recording of 85 target words, selected for each Italian obstruent in any position and context, before every vowel type, as singleton and geminate, in both stressed and unstressed syllables. Six native Italian speakers were recorded in a silent room using a Zoom H4n Pro recorder positioned 20 cm from their mouth. The informants were middle aged (average: 55), three females and three males, three northern speakers (from Genoa, Trieste and Locarno) and three central-southern speakers (from Rome, Bari and Naples). All participants are university level instructors of Italian language and culture in Budapest, and all speak various languages, including their dialect and Standard Italian. They were asked to read the target words aloud in their normal speech tempo, maintaining brief pauses between each word. The data were analysed in Praat, and descriptive statistics were carried out in Excel and R. (In this phase of the work, statistical analysis refers to manually computed mean values of the raw data; inferential statistical testing will be conducted in the next phase of the research.)

General observations Originally we measured three basic acoustic parameters in six stops [p t k b d g], four fricatives [f v s z] and four affricates [ts dz tʃ dʒ]: the duration of the closure or frication phase, the duration of the release phase, and voice onset time (VOT). According to current proposals in the laryngeal literature [1], the release phase is not included in the VOT; in cases of multiple releases, VOT is measured after the last release burst. Additionally, we examined the degree of voicing in initial lenis obstruents. Further laryngeal properties were also observed: multiple releases (up to six bursts) in most fortis stops and affricates, prenasalisation in several initial lenis stops, preaspiration in all fortis geminates; and a few sporadic dialectal features, such as deaffrication in intervocalic [tʃ], partial voicing in intervocalic and post-sonorant fortis stops, fricativisation of intervocalic fortis stops, etc. An example of a target word is shown in Figure 1: *pacco* 'package', as produced by the female speaker from Naples. The initial [p] is clearly unaspirated, with a release phase of 5 ms and a VOT of 3 ms; in contrast, the medial geminated [kː] can be interpreted as both pre- and post-aspirated: the closure phase begins with 56 ms of preaspiration, the closure itself lasts 253 ms, and multiple release bursts are evident (three transients on the wave form and three plosion bars on the sprectrogram). The release phase lasts 23 ms, followed by a VOT of 37 ms; thus, to total duration of the geminate is 369 ms. We now present the overall acoustic values observed for the Italian obstruents, measured in a manner similar to that described above. The main duration values for closure, release and VOT/frication are shown in the table in Figure 2 (in ms), with the first line indicating minimum values, the second line maximum values, the third line rounded mean values and the fourth line the median, respectively.

Results In fortis stops, the release and VOT duration correlate with the place of articulation (as is typical in languages); that is, the more posterior the stop, the longer the phases tend to be: [p] ~3+10 ms » [t] ~5+13 ms » [k] ~14+29 ms (rounded mean values). There is no significant difference in these values between initial and medial, stressed and unstressed, singleton and geminated stops; but they are considerably higher before front vowels, especially [i]. Medial singletons and geminates differ in closure duration (~99 ms vs. ~254 ms), and in the presence of preaspiration, which characterises 100% of the geminates (with ~63 ms at the beginning of the closure), but only 25% of the singletons and to a much lesser extent (~22 ms). We also observed multiple releases in 49% of all occurrences, and partial voicing in medial singletons in 31% among southern speakers. Lenis stops are somewhat shorter in total duration, but have longer release phases, again correlating with place of articulation: [b] ~6 ms » [d] ~11 ms » [g] ~16 ms. Initial lenis stops are clearly prevoiced, exhibiting a negative VOT phase, though closure often begins voiceless: the mean percentage of prevoicing is 93.74%, but we also found several partially voiced lenis stops as well (up to 40% of voice). Another noteworthy finding is the frequent prenasalisation in plain voiced initial stops: the closure phase begins as nasal in 23% of occurrences, averaging 37 ms. The difference between medial singletons and geminates lies only in closure length (~101 ms vs. ~219 ms), with release phases largely overlapping. A particularly interesting finding concerns duration differences between fortis and lenis fricatives and affricates (see the last 8 bars in Figure 3). Fortis fricatives are significantly longer ([f] ~162 ms, [s] ~123 ms) than lenis counterparts ([v] ~95 ms, [z] ~71 ms). Similarly, the fricative phase of fortis affricates ([(t)ʃ] ~99 ms, [(t)s] ~142 ms) is nearly double of that of the lenis ones ([(d)ʒ] ~50 ms, [(d)z] ~70 ms). Furthermore, the fricative parts of lenis affricates were only partially voiced: ~53% voice degree for [(d)ʒ], and ~38% for [(d)z], with many completely voiceless occurrences. This suggests that the laryngeal contrast between fortis and lenis in these obstruents may depend not only on voicing, but also on the duration of frication.

Conclusions Previous literature [5] reported mild aspiration in the fortis series of stops of Standard Italian; however, if we examine the actual VOT values excluding the release burst, these stops appear unaspirated, with extended release phases resulting from frequent multiple bursts. The lenis series is almost entirely prevoiced, so the laryngeal contrast between fortis and lenis primarily resides in the presence or absence of closure voicing. Regarding stops, Italian appears to be a fairly typical voice language. Instead, the duration of frication seems to play a role in the laryngeal contrast of fricatives and affricates. In the near future, we plan to validate these findings through perception experiments, assessing whether frication duration serves as a perceptual cue for fortis/lenis differentiation.

Acknowledgements I am very grateful to my informants, to my colleagues in the research group “Laryngeal variation in synchrony and diachrony”, and to NKFI grant #142498.

References

[1] Abramson, Arthur S. & D.H. Whalen (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics* 63, pp. 75-86.
[2] Beckman, Jill, Michael Jessen & Catherine Ringen (2013). Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics* 49/2, pp. 259-284.
[3] Cyran, Eugeniusz (2011). Laryngeal realism and laryngeal relativism: Two voicing systems in Polish. *Studies in Polish Linguistics* 6: 45-80.
[4] Dian, Angelo, John Hajek & Janet Fletcher (2024). Cross-regional patterns of obstruent voicing and gemination: The case of Roman and Veneto Italian. *Languages* 9: 383.
[5] Huszthy, Bálint (2021). Italian preconsonantal s-voicing is not regressive voice assimilation. *The Linguistic Review* 38/1, pp. 33-63.

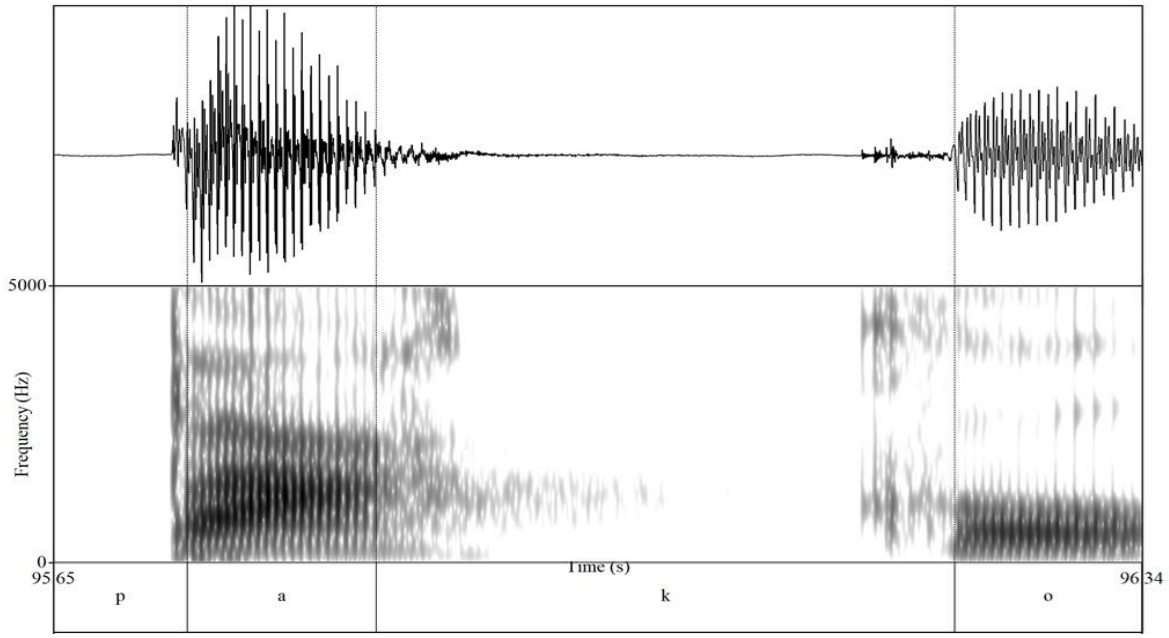


Figure 1: The production of It. *pacco* 'package' with preaspiration and multiple releases

Obstr.	[p]	[t]	[k]	[b]	[d]	[g]	[f]	[v]	[s]	[z]	[tʃ]	[dʒ]	[ts]	[dz]
Clo-sure	78	72	55	45	54	52	0	0	0	0	50	33	12	35
	150	121	118	165	305	197	0	0	0	0	68	130	174	94
	111	95	90	101	101	100	0	0	0	0	61	80	80	67
	107	88	92	106	93	93	0	0	0	0	65	81	54	66
Re-lease	1	2	4	3	4	4	0	0	0	0	1	2	3	4
	8	20	35	18	28	36	0	0	0	0	24	23	13	11
	3	5	14	6	11	16	0	0	0	0	10	7	7	7
	2	4	13	6	11	15	0	0	0	0	11	6	7	7
VOT / Fric.	3	6	8	-23	-47	-70	109	37	93	55	41	23	103	31
	34	39	72	-171	-322	-210	214	178	157	102	168	133	202	158
	10	13	29	-97	-115	-114	162	95	123	71	99	50	142	70
	9	12	27	-102	-115	-113	162	90	118	67	96	46	132	62

Figure 2: Average values of the dataset in ms (lines: min, max, mean, median)

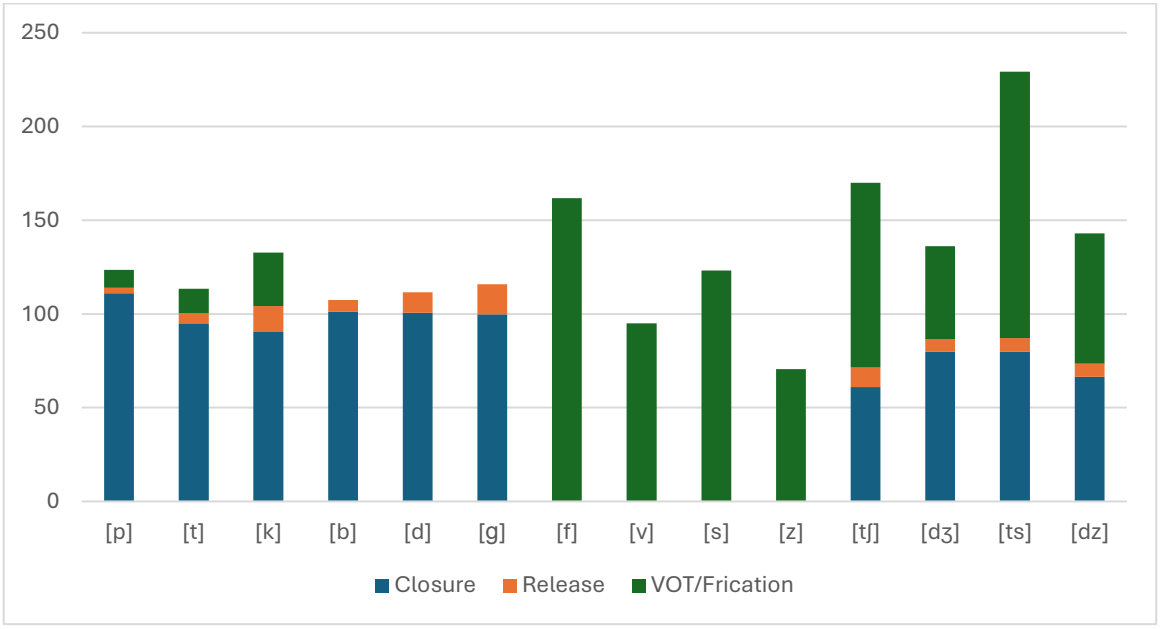


Figure 3: Total duration (ms) of all obstruents by mean values on stacked barplot

Ultrasound-to-Speech Synthesis through Speech Coding

Ibrahim Ibrahimov[1], Csaba Zainkó[1], Gábor Gosztolya[2]

[1]Budapest University of Technology and Economics, Hungary
[2]University of Szeged, Hungary

Speech as a communication method is a vital part of human life, involving several body parts (tongue, lips, etc.) during production. However, for various reasons (speech disorders, environment, etc.), producing audible speech is sometimes not possible or desired. To offer an alternative communication method in these scenarios, Silent Speech Interfaces (SSI) have been introduced. SSI is a technology that targets to recognize or reconstruct speech from articulatory movements. These movements can be captured using various techniques, such as lip videos [4] and ultrasound tongue imaging (UTI, [1]), among other modalities. Particularly, UTI is a promising technique for acquiring detailed information about tongue movement, while also being non-invasive and cost-effective.

Current research on articulatory speech synthesis using UTI primarily focuses on the direct synthesis approach. In this framework, ultrasound tongue image sequences (UTIF) are first mapped to an intermediate acoustic representation, typically a mel-spectrogram, through frame-by-frame learning. The predicted mel-spectrogram is then passed to a neural vocoder to generate the final speech waveform.

For the mapping stage, standard architectures such as 2D convolutional neural networks [1] and advanced models like the Conformer - a convolution-augmented transformer [3] - have been explored. Among these, Conformer-based ultrasound-to-mel mapping has emerged as the state-of-the-art, achieving the highest quality synthesized speech when combined with a HiFi-GAN vocoder. However, this direct synthesis framework is limited by the inherent mismatch between the content-sparse UTIF and the information-rich mel-spectrogram representation. To address this challenge, we propose incorporating UTIF into a speech coding architecture, enabling a more compact and information-balanced representation for speech synthesis.

Speech coding is the process of representing speech signals in a compressed form while preserving perceptual quality and intelligibility. Traditional approaches, such as linear predictive coding (LPC) or code-excited linear prediction (CELP), relied on hand-crafted signal processing techniques to reduce redundancy. Recently, neural-based audio codecs have demonstrated promising results in compressing and reconstructing speech with high perceptual quality. Among these, EnCodec stands out as a state-of-the-art real-time neural codec, offering superior performance by producing high-fidelity audio across a wide range of sample rates and bandwidths [2]. Unlike traditional codecs or feature representations such as mel-spectrograms, EnCodec leverages an encoder–quantizer–decoder architecture to generate compact, discrete latent representations that retain essential perceptual information while drastically reducing redundancy. This makes EnCodec not only more efficient for compression but also particularly advantageous as an intermediate representation for speech synthesis tasks, where balancing compactness and richness of information is crucial.

In this preliminary work, we propose a framework to overcome the sparse content issue of UTIF by mapping it to the latent space of EnCodec before quantization, leveraging its ability to compactly encode speech with high fidelity (Figure 1). To achieve this, we first train a fully convolutional encoder–decoder autoencoder to derive compact latent embeddings from UTIF. These articulatory embeddings are then mapped to the latent speech embeddings using a one-dimensional convolutional neural network (1D-CNN), followed by a long short-term memory (LSTM) layer to effectively capture temporal dependencies in the articulatory-to-acoustic mapping process.

To the best of our knowledge, this is the first attempt to leverage the latent space of a neural speech codec as an intermediate representation for UTI-based speech synthesis. To obtain initial results for this framework, we employed openly available data from the UltraSuite-TaL80 corpus. Specifically, we used the recordings of speaker "01fi", a native English female speaker, for whom paired and synchronized ultrasound–audio samples are provided.

Preliminary visual inspection shows that the obtained speech representation - prior to quantization and decoding - exhibits poor quality (Figure 2). Nevertheless, this limitation does not hinder the development of the framework, as more advanced methods for the mapping stage can be explored, and we plan to continue improving this aspect in future work.

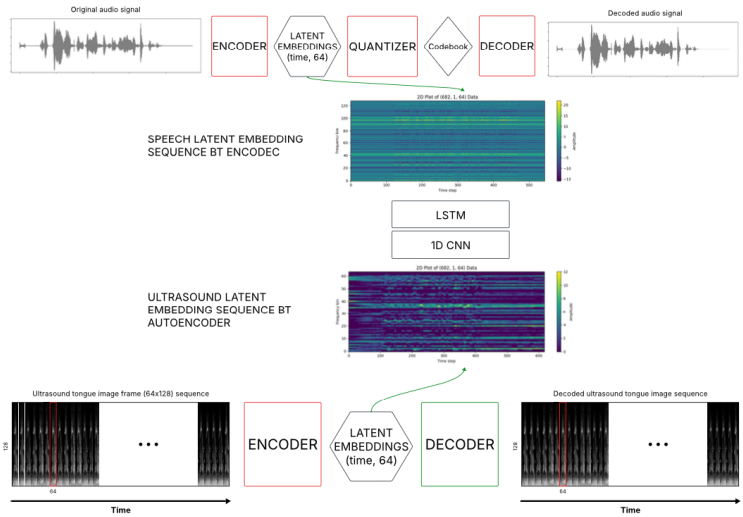


Figure 1: Proposed framework: Ultrasound-to-speech through speech coding.

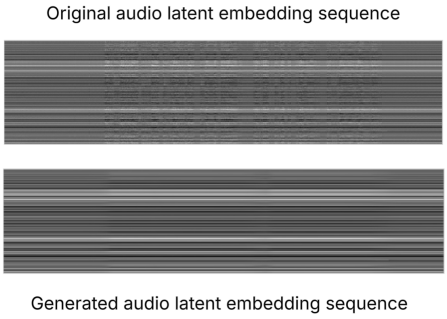


Figure 2: Initial result of preliminary framework. (speaker "01fi", sample "004_xaud")

References

[1] Csapó, T. G. et al. (2020). 'Ultrasound-Based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis'. In: *Proceedings of Interspeech*, pp. 2727–2731. DOI: 10.21437/Interspeech.2020-1031.

[2] Défossez, A. et al. (2023). 'High Fidelity Neural Audio Compression'. In: *Transactions on Machine Learning Research*. Featured Certification, Reproducibility Certification. ISSN: 2835-8856. URL: https://openreview.net/forum?id=ivCd8z8zR2.

[3] Ibrahim Ibrahimov and Csaba Zainkó and Gábor Gosztolya (2025). 'Conformer-based Ultrasound-to-Speech Conversion'. In: *Interspeech 2025*, 5578–5582. DOI: {10.21437/Interspeech.2025-2147}.

[4] Ribeiro, M. S. et al. (2021). 'Silent versus Modal Multi-Speaker Speech Recognition from Ultrasound and Video'. In: *Proceedings of Interspeech*, pp. 641–645. DOI: 10.21437/Interspeech.2021-23.

A magyar anyanyelvű beszélők zárhang-percepciójának vizsgálata kínai szintetizált hangmintákkal

Juhász Kornélia[1,2] és Bartos Huba[1]
[1] ELTE Nyelvtudományi Kutatóközpont
[2] HUN-REN–NYTK Neurofonetikai Kutatócsoport

A kísérlet középpontjában az magyar anyanyelvű beszélők zárhang-percepciójának vizsgálata áll. A magyar nyelvben bilabiális [b, p], alveoláris, [d, t] és veláris [g, k] kifelé pattanó orális explozívák állnak oppozícióban. E fonológiai szembenállást fonetikai szempontból elsősorban a zöngeképzés időzítésével tudjuk jellemezni. Az explozívák zöngésségét a zöngekezdési idő (voice onset time, VOT) értékével tudjuk számszerűsíteni, amely megadja a zöngeképzés kezdetét a zárfeloldás pillanatához viszonyítva [1]. A fonetikai zöngésségi kategóriákon belül – Gósy és Ringen [2] mérései alapján – a magyarban zöngés aspirálatlan ([b] −68,6; [d] −59,4; [g] −52,9 VOTms) és zöngétlen aspirálatlan ([p] 18,4 [t] 20,0 [k] 42,7 VOTms) beszédhangok állnak szemben egymással. Fontos megemlíteni azt is, hogy a képzési hely hatással van a VOT-értékekre, pl. zöngétlen aspirálatlan zárhangok esetében a képzési hely hátrébb mozdulásával növekszik a +VOT értéke. A VOT-értékek mellett a koartikuláció miatt fontos megemlíteni a zárhangot követő magánhangzó elején mért alapfrekvencia (f0) értékét is, ami a zöngétlen beszédhangokat követően magasabban realizálódhat a zöngés explozívákhoz képest [3]. Habár az explozívák fonetikailag akusztikai tulajdonságaik szerint zöngésségi kategóriákba rendezhetők, azonban e beszédhang-megvalósulások észlelése, illetve fonémakategóriába sorolása nyelvspecifikus folyamat. Az explozívákat a magánhangzókhoz viszonyítva kisebb akusztikai variabilitás jellemzi (az összetett akusztikai szerkezetüknek köszönhetően és magánhangzóknál erősebb koartikulációs rezisztenciájukból fakadóan), ezért észlelésük erősen kategorikusnak tekinthető. A kategorikus észlelés esetében a percepciós kategóriahatárnál nem feltételezünk jelentős átmenetet a kategóriák között, hanem az észlelet egy ponton hirtelen zöngésből zöngétlenbe csap át [4].

A kísérlet célja annak a jelenségnek a vizsgálata, hogy a VOT- és az f0-érték manipulálása hatással van-e a magyar beszélők percepciójára, pontosabban a zárhangok zöngésségi kategorizációjára a bilabiális, alveoláris és veláris zárhangpárok oppozíciójában. A szerzők tudomása szerint a magyar anyanyelvűek zárhang-percepciójára vonatkozóan kizárólag Neuberger Tilda (pl. [5]) végzett kísérleteket, a rövid-hosszú (szingleton-gemináta) zárhang-párok észlelését térképezve fel, vagyis a zöngésségi oppozíció még nem került empirikus kísérlet középpontjába. A percepciós kategorizációt kínai szintetizált VOT-kontinuumok előállításával vizsgáltuk [6]. A szintetizált VOT-kontinuum esetében mindhárom képzési helyen automatikusan létrehoztunk egy 10 elemű VOT-skálát, ahol a VOT időtartama egy előre megszabott értékkel fokozatosan növekszik. A VOT-kontinuum szintetizálásakor azonban kizárólag az aspirációs zörej időtartamát manipuláltuk (vagyis a +VOT-értéket, a zörej tulajdonságait nem), abból fakadóan, hogy a –VOT-értékhez tartozó zárszakaszban megjelenő előzönge manipulációja több új változó vizsgálatát is megkövetelné (pl. a zárszakasz hossza, illetve a zárszakaszban megjelenő zönge pozíciója). Ezért e kísérlet alapjául olyan kínai intervokális zárhang-megvalósulásokat használtunk, amelyek habár fonetikailag zöngétlen aspirálatlan szegmentumok, a (kínai nyelvi tapasztalattal nem rendelkező) magyar anyanyelvű beszélők mégis alapvetően zöngés beszédhangnak észlelték őket. A kísérletben manipulált bilabiális képzési helyű zöngétlen megvalósulásokat a 20 magyar anyanyelvű kísérleti személy az esetek 90%-ban azonosította zöngésnek; az alveoláris képzési helyűt 87%-ban, illetve a veláris képzési helyűt pedig 98,4%-ban ítélték meg zöngés zárhangnak [7]. A zárhangot követő magánhangzó kezdetén mérhető f0-varáció tekintetében a +VOT-kontinuumokat további három

kondíció mentén vizsgáltuk: I. Az első kondícióban a +VOT-kontinuum mind a 10 elemében egyaránt ugyanazon (az eredeti) f0-értékek jelennek meg, míg a +VOT értéke fokozatosan növekszik. II. A második kondícióban – hasonlóan az I. kondícióhoz – a +VOT-kontinuum 10 elemében ugyanazon kezdeti f0-érték jelenik meg, azonban ebben az esetben az I. kondícióhoz képest magasabb f0-értékeket rendeltünk. Ha a kezdeti f0-érték hatással van a magyar anyanyelvűek zárhang-percepciójára, akkor az f0 megemelése a zöngétlen irányba tolja az észlelést, vagyis a zöngés-zöngétlen kategóriaváltás az I. kondícióhoz képest korábban (a +VOT-kontinuum 0-hoz közelebbi fázisában) valósul meg. III. A harmadik kondícióban a +VOT-értékekkel együtt fokozatosan növekszik az f0-érték is, ebben az esetben pedig az I. és a II. kondícióhoz képest köztes eredményeket várunk. Összegezve tehát a kísérletben a magyar anyanyelvűek zárhangpercepcióját a három képzési helyen három kondíció mentén vizsgáltuk, tehát összesen 9 +VOT-kontinuumban 90 különböző megvalósulás észlelését elemeztük, ahol a +VOT lépések minden esetben 4,4 ms-mal növekedtek; a bilabiális helyen 2 VOTms-tól, az alveoláris esetében 4 VOTms-tól, illetve a veláris esetében 12 VOTms-tól. A megemelt f0 értékét a szakirodalomban megfigyelt szélsőértékek (0 Hz és 70 Hz) közötti értékben, vagyis 35 Hz-ben határoztuk meg. A kísérletben minden megvalósulás három ismétléssel jelent meg.

A percepciós kísérlet formája két alternatívás kényszerített választási teszt volt a psytoolkitben [8,9]: az adott zárhang megvalósulásához két magyar ortográfiával megjelenített grafémasor volt rendelve (pl. *ábá* vs. *ápá*). A kísérletben 14 naiv magyar anyanyelvű személy vett részt. Az eredmények statisztikai elemzésére binomiális kevert GLMM használtunk az R-ben. A modellben a képzési helyekre, illetve kondíciókra külön GLMM-t írtunk fel, illetve a modellek által prediktált szigmoid görbéket, vagyis a zöngétlen kategória választásának valószínűségét a VOT-lépések függvényében az 1. ábra mutatja be a görbék inflexiós pontjaival. Képzési helytől és kondíciótól függetlenül a VOT-lépések mentén mindig szignifikáns hatás volt megfigyelhető ($p < 0{,}001$), vagyis nem meglepő módon a +VOT emelkedése minden esetben növelte a zöngétlen kategória választásának valószínűségét. Az inflexiós pontokat tekintve mindhárom képzési hely esetében azonos mintázatot figyelhetünk meg: a megemelt f0-értékek esetében (kék színű görbe) az inflexiós pont a VOT-kontinuum alacsonyabb értékei felé tolódik, tehát relatíve alacsonyabb VOT-érték eredményez nagyobb valószínűséggel zöngétlen kategorizációt. Hasonlóképpen a bilabiális és alveoláris képzési helyeken az eredeti f0-érték fenntartása az inflexiós pontot (illetve a teljes zöld színű görbét) a zöngétlen irányba tolta el, vagyis magasabb VOT-értékhez társított alacsonyabb f0-értékek is valószínűen zöngés kategorizációt eredményezhettek (relatíve a másik két kondícióhoz). Azokban az esetekben, amikor a VOT és az f0 együtt változott, az alveoláris és bilabiális képzési helyeken az eredeti, és a megemelt f0-hoz tartozó görbékhez képest köztes értékeket kaptunk. A veláris képzési hely ebben az esetben azonban kivételnek bizonyult: e képzési hely esetében az eredeti f0-lal társított realizációk (zöld görbe) és a VOT-vel együtt változó f0-értékek (piros görbe) inflexiós pontjai (a görbék lefutásával együtt) nagyon hasonlónak mutatkoztak. A három kondíciót tartalmazó statisztikai próba eredményei szerint kizárólag a veláris képzési hely esetében, a megemelt f0-értékkel realizálódó VOT-kontinuum ingerei tértek el szignifikánsan a két másik kondíciótól, míg a többi képzési hely esetében nem találtunk szignifikáns eltérést az egyes kondíciók között. Tehát a bemutatott eredmények az alveoláris és bilabiális képzési helyeken csak statisztikailag nem megerősített eltérést jeleznek. Az eredmények arra utalnak, hogy a veláris képzési helyen az f0 emelése kevésbé egyértelmű kategorizációt indukál, ami – ha angol explozívák eredményeit vesszük alapul [10]– akkor e képzési hely esetében a legkisebb az f0-eltérés a zöngésség szerinti párok között, ami könnyen okozhat interferenciát. E magyarázat magyar nyelvi relevanciája kérdéses és vizsgálatokat igényel. Az eredmények hozzájárulhatnak az idegen nyelvek oktatásához és felhasználhatók lehetnek a beszédfelismerésben és -szintézisben.

Irodalom

[1] Lisker, L.–Abramson, A. 1964. A cross-language study of voicing in initial stops. Acoustical Measurements *Word 20/*3: 384–422.

[2] Gósy, M.–Ringen, C. 2009. Everything you always wanted to know about VOT (in Hungarian). Előadás. *9th International Conference on the Structure of Hungarian*, Budapest, 2009. szeptember 1.

[3] House, A.–Fairbanks, G. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. of the Acoustical Society of America, 25:* 105–113.

[4] Massaro, D. W. – Cohen, M. M. 1983. Categorical or continuous speech perception: A new test, *Speech Communication 2:* 15–35.

[5] Neuberger T. 2021. A rövid és hosszú zöngétlen explozívák észlelése gyermekeknél *és* felnőtteknél. *Beszédtudomány 2/*1: 65–98.

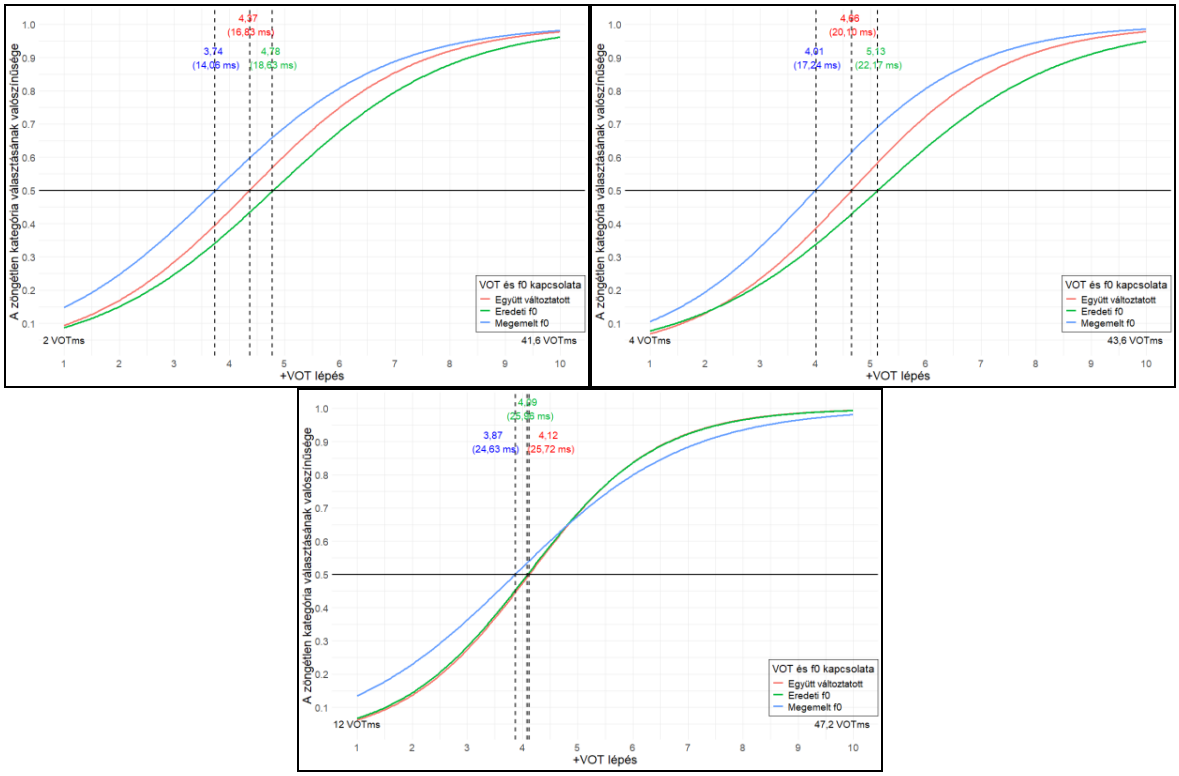
[6] Winn, M. 2020. Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script. *J. of the Acoustical Society of America 147/*2: 852–866

[7] Bartos, H. – Juhász K. 2025. The perception of Mandarin lenis obstruents by naïve Hungarian speakers. Előadás. Chinese as a second language research konferencia. Brassó, 2025. május 30–31.

[8] Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42(4)*, 1096-1104.

[9] Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44(1),* 24-31.

[10]House, A. – Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. of the Acoustical Soc. of Am., 25*/1: 105–13.



1. ábra: A zöngétlen kategória választásának valószínűsége a VOT-lépések függvényében a bilabiális (bal fent), alveoláris (jobb fent) és veláris (alul) képzési helyen a görbék inflexiós pontjaival (zöld=eredeti f0 minden VOT-lépés esetében, kék = megemelt f0 minden VOT-lépésnél, piros = f0 emelése párhuzamosan a VOT-lépésekkel

Temporal characteristics of Mandarin neutral tones in the production of atonal Hungarian learners of Chinese

Kornélia Juhász[1,2], Tekla Etelka Gráczi[1,2], Katalin Mády[1], Fang Hu[3], Shuwen Chen[3], Huba Bartos[1]
[1] ELTE Research Centre for Linguistics,
[2] MTA–HUN-REN NYTK Lendület "Momentum" Neurophonetics Research Group,
[3] Chinese Academy of Social Sciences

The present study focuses on the acquisition of Mandarin Chinese (MC) neutral tones, which are carried by syllables often considered phonologically underspecified, i.e. toneless. In the present acoustic analysis, we aim to investigate the temporal properties of neutral tones in sequences consisting of 3 units appearing in Chinese hortatives in the production of Hungarian native (L1) learners. In MC four lexical tones are contrasted: high level Tone 1 (T1), rising Tone 2 (T2), low falling-rising Tone 3 (T3) and falling Tone 4 (T4). These four tones do not exclusively differ by their F0 curve, their duration serves as a secondary acoustic cue ([2]). The duration of the four MC lexical tones rises in the following order: T4<T2</~T1<T3 [1]. As is mentioned above, in MC there is a fifth tonal value, a neutral tone (NT), and its variable acoustic realization is in the focus of the present paper. Syllables carrying NTs are bound, unstressed lexical morphemes which are often assumed to be phonologically unspecified, since their realization is highly dependent on several different factors, e.g. the value of the preceding full tone, as well as sentence type. Typologically, considering only the relevant cases, neutral tone syllables include grammatical morphemes which are inherently toneless (such as the clause-final particle *le* or the sentence-final hortative particle *ba*), and diminutive kinship terms, in which case the reduplicated second syllable loses its intrinsic full tone and becomes a NT (such as in the case of *māma* 'mother') ([3], [4]). In this analysis the temporal, i.e. durational characteristics of the above-mentioned three neutral tone syllables are investigated in a sequenced manner, utterance-finally. As a result of its unstressed nature, the duration of the NT syllable is reduced compared to the MC full tones, and the magnitude of shortening may vary as a function of the preceding tone ([5]). The mean duration of a neutral tone syllable is found to be longer subsequent to a T3 syllable (about 70% of the preceding T3), whereas subsequent to a T1/T2/T4 the syllables are reduced approximately to 50%–60% of the preceding full tone syllables [6]. Besides, it must be noted that the prosodic position, in particular, phrase-final position, also affects the Mandarin lexical tones' realization: as a general tendency (similarly to the Hungarian L1), phrase-final lengthening can be observed at the utterance boundaries [7, 8, 9]. The motivation of the study is to investigate how speakers with an atonal L1 cope with the tone-dependent and highly variable realizations of neutral tones strung into sequences considering the level of L2 experience, These results, in the future, could be applied as a reference when f0 patterns of the same utterances are compared between groups. Besides, the analysis also provides insight into the acoustic characteristics of the hortative particle *ba*, which is less frequently in the focus of attention compared to e.g., the particle *ma*.

We analysed three adult speaker groups (5 female speakers per group): 1. Hungarians with 1–2 years of learning MC ("Beginners") and 2. Hungarians with 3–4 years of learning MC: upper-intermediate MA students of Chinese ("Advanced learners"), and 3. a control group of Chinese native speakers born and raised in Northern China. We recorded hortative utterances embedded in short dialogues projected on a screen with both Chinese characters and pinyin transcriptions. The target sequences consisted of 4 syllables in an utterance-final position in broad focus sentences. The analysed quadri-syllabic sequences followed the following pattern: T_x + *NT* + *NT* + *NT* (for the recorded sequences see Table 1.). Syllable durations were labelled and

measured in Praat. Since the aim of the experiment was to investigate neutral tone sequences, and in order to adjust the material to L2 learners' vocabulary, we chose those relatives' names which they encounter in the first semester of learning Chinese. This poses problems in the case of *yeye* [jɛjɛ], where segmenting the two syllables is rather subjective. In the acoustic analysis we compared the by-speaker *z*-score normalized relative syllable durations in the sequence (i.e., (syllable duration – speaker's mean) / speaker's SD). In addition, we calculated the duration ratio of the disyllabic kinship terms defined as the percentage of reduction of the 2nd NT syllable relative to the full toned realization in the 1st syllable.

Table 1: The recorded quadri-syllabic sequences (in utterance-final position, in 7/8-syllable-long sentences)

Object		**PRT**	**PRT**
T_x	**N_1**	**N_2**	**N_3**
妈妈 ***māma*** 'mother' 爷爷 ***yéye*** 'grandpa' 奶奶 ***nǎinai*** 'grandma' 妹妹 ***mèimei*** 'sister'		了 ***le*** (particle)	吧 ***ba*** (particle)

We analyzed by-speaker *z*-score normalized relative syllable durations as a function of syllable position by fitting mixed-effects models with random intercept (per speaker) for each speaker group. In each group, syllable position exerted a significant effect on duration ($F(3,12) = 181$, $p < .001$). All three speaker groups produced the same pattern, although they differed significantly in the pattern's realization: production was characterized by shorter durations towards the end of the utterance as the reduction effect on NTs culminated towards the 4th syllable, and there was a lengthening effect affecting the utterance-final syllable (relative to the penultimate one) (Fig. 1). This can be attributed to the general effect of phrase-final lengthening, shaping utterance-final elements both in Chinese and in Hungarian. The phrase-final lengthening effect is significantly longer in both L2 learner groups, compared to the native production. We also analyzed the duration ratio in the kinship terms (i.e., T_x+N_1 in Table 1.), i.e., the reduction of the 2nd syllable relative to the 1st along the four tonal values, within each group, respectively (Fig. 2) by mixed-effects models (Fig. 2). In comparison to previous results, where the reduction rate approximated 50–70% [e.g., 6] relatively to the full tone realization, both native speakers' and L2 learners' reduction rate did not exceed ~ 20 percentage, Fig. 2.). This low reduction rate is not unique in the literature, and possibly induced by relatively long utterances, also in a repeated manner (similar findings were presented in [10]). In native production the full tone's value did not affect the reduction ratio, that is, difference was not found among tonal values in the 1st syllable, possibly due to the high variability of the data. In both L2 learner groups' production, the NT in the kinship term carrying T4 was the least reduced, in particular, to an extent that the NT syllable's duration even surpassed that of the full tone syllable, inducing significant differences from other kinship terms' reduction ratio in both L2 groups. This pattern could be explained by considering the relatively short realization of T4 associated to a relatively complex syllable structure (C+ diphthong), which factors possibly limit the reduction of the NT syllable. These restrictions may apply to T3, as well, where the syllable structure is also complex, but the reduction rate is – according to previous sources [6] – the least among tonal patterns. It must also be pointed out, that if the level of L2 experience is concerned, the duration ratio increases (in other words the reduction rate decreases), i.e., beginners' reduction of the 2nd syllable is less prominent than those of advanced learners in some cases even becomes non-existent. The results of the analysis contribute to the deeper understanding of the relationship between full lexical tones and neutral tones in in the production of L2 learners with atonal L1.

References

[1] Xu, Y. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics,* 25, 61–83.

[2] Zhang, H., Wiener, S. & Holt, L. L. 2022. Adjustment of cue weighting in speech by speakers and listeners: Evidence from amplitude and duration modifications of Mandarin Chinese tone. *The Journal of the Acoustical Society of America, 151,* 992–1005.

[3] Chen, Y. & Y. Xu, 2006. Production of Weak Elements in Speech – Evidence from F0 Patterns of Neutral Tone in Standard Chinese. *Phonetica*, vol. 63, 47–75.

[4] Li, Z. 2003. *The phonetics and phonology of tone mapping in a constraint-based approach*. PhD diss. Cambridge, Mass: MIT.

[5] Tang, P. et al. 2018. Acquisition of weak syllables in tonal languages: Acoustic evidence from neutral tone in Mandarin Chinese. *Journal of Child Language*. 46(1): 24–50.

[6] Cao, J. F. 1986. The Acoustic Cues of Neutral Tone syllable in Standard Chinese," *J. Applied Acoustics*. 5(4), 1–6.

[7] Wu, Y., et al. 2023. Mandarin lexical tone duration: Impact of speech style, word length, syllable position and prosodic position, *Speech Communication*, 146, 45–52.

[8] Markó, A. et al. 2022. Magyar magánhangzók artikulációs és akusztikai jellemzői a fonetikai pozíció függvényében álszavakban. *Általános Nyelvészeti Tanulmányok* 24. 51–79.

[9] White, L., & Mády, K. 2008. The long and the short and the final: Phonological vowel length and prosodic timing in Hungarian. Proc of the 4th Speech Prosody. 363–366.

[10] Vamvakos, G. 2024. Acoustic study of neutral tone syllables produced by intermediate Mandarin Chinese learners. BA thesis, Dalarna University.

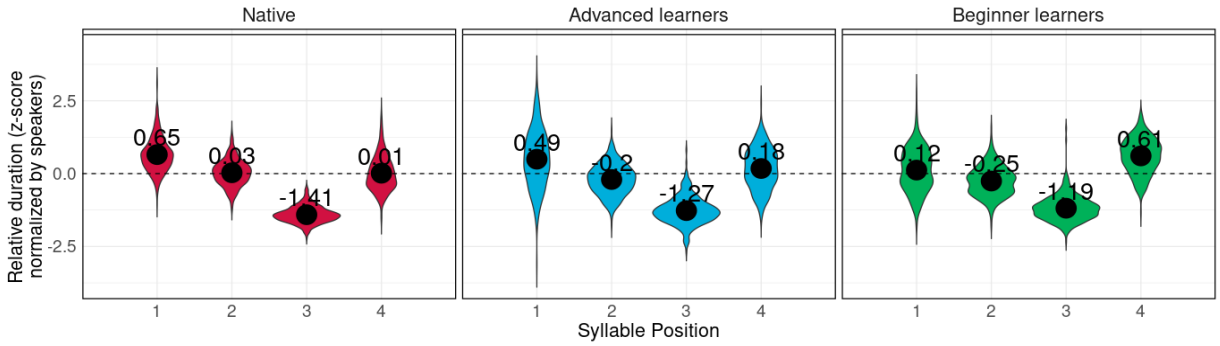


Figure 1: Relative (z-score normalized) duration of the phrase-final quadri-syllabic sequence where the 1st syllable features a full tone, the 2nd, 3rd and 4th syllables are representing neutral tones.

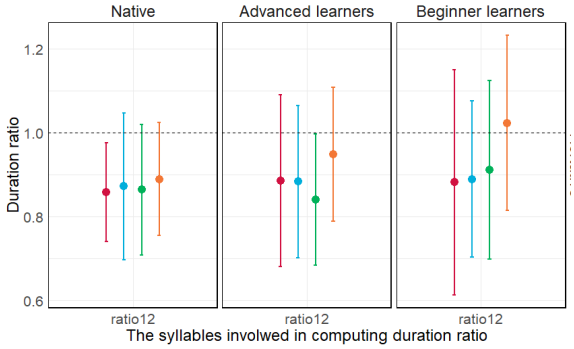


Figure 2: 2nd syllable:1st syllable ratio in the three speaker groups' production where red = T1, blue = T2, green= T3, orange = T4 in the 1st syllable.

Tonal representation of Hungarian yes/no questions and acoustic correlates

Katalin Mády, Uwe D. Reichel, Anna Kohári, Ádám Szalontai

ELTE Research Centre for Linguistics, Budapest

Introduction. Sentence intonation can be described in various ways. A well-known approach is the contour-based model containing dynamic categories such as rise, fall or level. An extensive description of Hungarian intonation in this framework was provided by Varga [6]. Another approach was to decompose the tonal movements into High and Low (H, L) tonal events which are aligned with prominence-bearing syllables and phrase boundaries, resulting in the ToBI annotation system (**To**nes and **B**reak **I**ndices) [5]. Since systems like ToBI focus on specific points of interests in prosodic structure. They integrate structural analysis across different levels like syntax and phonology. Furthermore, they are quantifiable both in terms of labels associated with tonal events as well as acoustic features like fundamental frequency (f0) and intensity (energy). The ToBI system has been adapted to a wide range of languages, which facilitates the comparison between prosodic forms and functions and the prosodic typology across languages. However, annotations rely heavily on previous assumptions on the intonational phonology of the given language.

Intonation annotation. In this paper, we suggest a basis for a systematic and comprehensive description of Hungarian intonation in the tone-based approach. To avoid circularity when developing an annotation system and a phonological description at the same time, the German DIMA (*Deutsche Intonation, Modellierung und Annotation*) system is applied. This pre-phonological, surface-related, theory-neutral annotation system aims at the prosodic representation of the actually produced pitch contour [1]. As with ToBI, annotators are trained to rely on their perceptual impression rather than simply to set marks based on f0 movements. Annotations in DIMA are carried out on three layers. (1) **Phrase** boundaries are strong (%) or weak (-). (2) **Prominence** strength is weak (1), strong (2) or emphatic (3). (3) High (H) and low (L) **tones** are associated with phrase boundaries and prominent syllables. DIMA provides the possibility to mark tonal events (e.g. turning points) without linking them to a phrase boundary or prominence. This is useful when describing variation in the pitch contours that are perceivable but do not necessarily denote different grammatical or pragmatic functions. One such example is given in Fig. 1 showing three possible forms of neutral information-seeking yes/no questions. In ToBI, all three contours are described as L* H L%. In DIMA, the mid (red) contour contains an additional L tone to mark the turning point from level to rise before the high penultimate syllable, whereas the bottom (green) contour contains a high plateau indicated by H on the syllable following the L* pitch accent.

The downstep (!) and upstep (ˆ) symbols indicate that a pitch accent is considerably lower or higher than a previous identical pitch accent within the same utterance. These symbols are also used to mark lower or higher register or larger or smaller pitch range of a phrase compared to the previous one. This feature of DIMA is generalised in our approach: ! and ˆ are used to indicate the relative height of otherwise identical pitch accents of the same

speaker even across utterances. This distinction is useful when we explore intonation patterns of non-neutral pragmatic contents. For example, Hungarian incredulous yes/no questions are characterised by lower f0 and a smaller f0 range [2]. This is captured by the lowered pitch accent !L* as shown in Fig. 2, right panel.

Materials. 146 yes/no questions of 12 native Hungarian speakers (10 f., 2 m.) in the spontaneous Akaka Maptask Corpus [3] were labelled by three trained annotators. A common decision was achieved on consent, if necessary. A total number of 993 tone labels were available for analysis. Label counts were unbalanced: L tones for initial and final phrase boundaries and turning points were the most frequent (33%), followed by H (21%) which is the penultimate f0 peak between the rise and the fall. Out of 251 pitch accents, 196 (78%) were !L, L or ˆL. The prevalence of low pitch accents is not surprising given the rising-falling pattern of yes/no questions.

Prosodic feature extraction. In order to identify acoustic correlates of the manually annotated tone labels, f0 was extracted by autocorrelation with Praat version 6.4.06 and postprocessed with CoPaSul version 1.5.3 [4]. We then extracted standard f0 and energy features (22 and 26, respectively) like median and IQR, as well as 76 local f0 contour features with CoPaSul. For local contour feature extraction, f0 was decomposed into a global and a local component based on the phrase boundary and tone time stamps. All features were extracted within analysis windows of length 200 ms (standard features), respectively 300 ms (contour features), centered on the tone label time stamps.

Analyses. In an initial exploratory analysis we identified all prosodic features that (1) turned out to differ significantly between at least one tone label pair with (2) at least medium effect size. For this purpose we applied Kruskal-Wallis tests with α set to 0.05, and we measured the effect size in terms of Cohen's D, "medium" defined as $D \geq 0.5$. Given the exploratory nature of this analysis and the large amount of features, we did not apply any type-1 error correction. Our results show that among the standard f0 features 15 out of 22 fulfilled both above-mentioned criteria: among the standard energy features 18 out of 26, and among the local f0 contour features 69 out of 76. For two of these identified features, namely f0 and energy median, we show the distribution plots for all twelve tone labels in Fig. 3. Furthermore, the parabolic stylisation coefficient distributions reveal a predominantly falling-rising shape for L tone variants (except of ˆL), and a rising-falling shape for H variants.

Discussion. Our prosodic annotations of spontaneous Hungarian yes/no questions with a slightly modified version of the DIMA system rely on perceptual labelling rather than identifying f0 minima and maxima in the signal. The analysis of various acoustic features support the reliability of these impressionistic labels showing that the six tone heights are indeed distinguishable, and they follow the expected order from low to high. Besides, prominent syllables are characterised by higher energy than their non-prominent equivalents (boundaries and turning points). These results are promising for the future development of an automatic tone prediction tool.

Acknowledgements. This work was funded by the National Research, Development and Innovation Office, grant NKFIH K 135038.

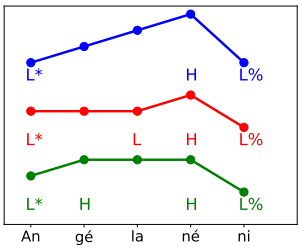


Figure 1: Three contours for the neutral yes/no question 'Aunt Angéla?' (Varga 2002:480).

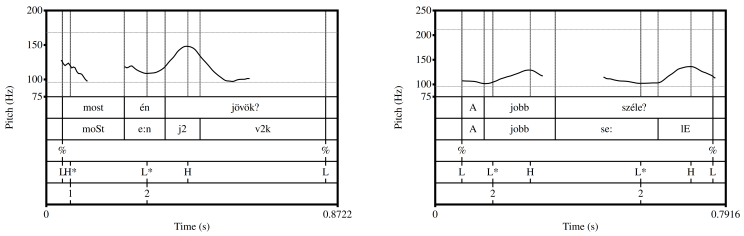


Figure 2: F0 contours and the speaker's f0 range. Left: neutral yes/no question 'Is it my turn?'. Right: incredulous question with lowered pitch and smaller pitch range 'At the right edge?'.

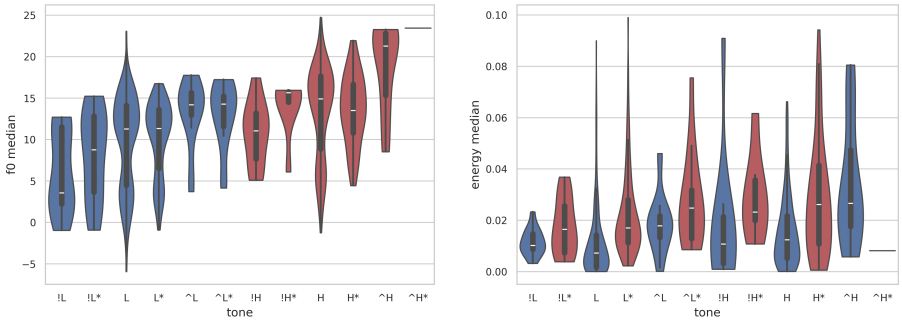


Figure 3: Violin plots for median values for lowered (!) and raised ˆL(ow) and H(igh) tones. The * symbol indicates prominence. Left: f0, right: energy.

References

[1] Kügler, F. et al. (2015). 'DIMA – annotation guidelines for German intonation'. In: *Proc. 18. International Congress on Phonetic Sciences*. Glasgow.

[2] Mády, K. et al. (2023). 'Prosodic cues of distinguishing neutral and non-neutral yes/no questions in Hungarian: the acoustics of surprise'. In: *Proc. 16th ICPhS, Prague*, pp. 1608–1612.

[3] Molnár, C. S. et al. (2023). 'The Akaka Maptask Corpus'. In: *Beszédkutatás – Speech Research Conference*. Budapest, pp. 81–83.

[4] Reichel, U. D. (2016). *CoPaSul Manual – Contour-based parametric and superpositional intonation stylization*. Research Institute for Linguistics, Hungarian Academy of Sciences. Budapest, Hungary.

[5] Silverman, K. et al. (1992). 'ToBI: a standard for labeling English prosody'. In: *Proc. 2nd International Conference on Spoken Language Processing, Banff, Alberta*. Vol. 2, pp. 867–870.

[6] Varga, L. (2002). *Intonation and stress: evidence from Hungarian*. Basingstoke & New York: Palgrave Macmillan.

Az eldöntendő kérdő és a kijelentő mondattípusok prozódiai eltérése Down-szindrómával diagnosztizált fiatal felnőttek beszédében (előtanulmány)

Markó Alexandra[1], Jankovics Julianna[1,2], Pirsel Katalin[1,3,4] és Juhász Kornélia[1,4]
[1]MTA–HUN-REN NYTK Lendület Neurofonetikai Kutatócsoport,
[2]BGéSZC Öveges József Technikum és Szakképző Iskola,
[3]ELTE Nyelvtudományi Kutatóközpont,
[4]ELTE Eötvös Loránd Tudományegyetem

A prozódia akusztikusan az alapfrekvencia, az intenzitás és az időtartam változásaiban nyilvánul meg, amelyek különféle kombinációi számos különböző funkció kifejezésére szolgálnak. Ezek egyike a magyarban az (eldöntendő) kérdés és a kijelentés intonációs megkülönböztetése (pl. *Elmegyünk?* vs. *Elmegyünk.*).
A Down-szindróma az egyik leggyakoribb veleszületett kromoszóma-rendellenesség, amely kognitív (értelmi fogyatékosság), fiziológiai (anatómiai eltérések a beszédképző szervek érintettségével), beszédbeli és nyelvi hiányosságokkal, nehezítettséggel jár együtt. A Down-szindrómával diagnosztizált egyének nyelvi zavarai jelentősen eltérhetnek a különböző nyelvi területeken. A magyar kutatások a fonológiai szintű problémákat jellemzően a szegmentális sajátosságok körében vizsgálták, ahol gyakoriak a beszédhanghibák [1, 2]. A szupraszegmentális tényezők közül a beszédtempó vizsgálatát végezték el gyermekek esetén [1, 2]. Sem a nemzetközi, sem a magyar szakirodalomban nem ismerünk olyan elemzést, amely a prozódiai kontrasztok megvalósítása szempontjából vizsgálja a Down-szindrómával diagnosztizált egyének beszédét.
A prozódia és ezen belül a beszéddallam jelentős szerepet játszik abban, hogy az egyén képes-e percipiálni és produkálni nyelvileg kontrasztív különbségeket. Ezért releváns a prozódiai mintázatok leírása a kommunikációs nehézségekkel küzdő populációkban, mint amilyen a Down-szindrómával diagnosztizált beszélők köre. Korábbi kutatások szerint a prozódiai jellegű kommunikációs nehezítettség gyakori az értelmi fogyatékossággal és a társuló idegrendszeri fejlődési zavarokkal küzdő személyek esetében (vö. pl. [3, 4]).
A prozódiai zavarok sokféleképpen jelentkezhetnek, beleértve a prozódiai mintázatok helyes észlelésének és/vagy előállításának nehézségeit vagy a prozódia nyelvtani, pragmatikai és érzelmi funkcióinak megértésében és/vagy kifejezésében jelentkező nehézségeket. Ezek a károsodások kapcsolódhatnak beszéd-, hallás-, nyelvi vagy kognitív nehézségekhez, és különbözőképpen befolyásolhatják a kommunikációt. Az atipikus expresszív prozódia az érthetőség romlásához és gyenge szociális és kommunikációs kompetenciához vezethet (l. pl. Loveall et al. [3] összefoglalását). A jelen pilotkutatásban Down-szindrómával diagnosztizált beszélők hangfelvételei alapján elemezzük a kijelentő és az eldöntendő kérdő mondattípus három szótagos realizációit.

Módszertan. A vizsgálatban 8 (6 nő és 2 férfi, 23 és 40 év közötti életkorú) Down-szindrómával diagnosztizált (a továbbiakban DS) beszélő és 4 női kontrollbeszélő (19–22 évesek; a továbbiakban KO) mondatbemondásait elemeztük. A mondatok elicitálását többféle módszerrel végeztük (a DS-beszélőknél szituációba ágyazva, képernyőről olvasva stb., a KO-beszélőknél csak képernyőről olvasva), a DS-beszélőknél változó sikerességgel, ennélfogva beszélőnként eltérő mennyiségű adat áll a rendelkezésünkre. A sikeresen megvalósított prozódiai kontrasztot tartalmazó DS-megnyilatkozások kiszűrésére percepciós tesztet végeztünk a Praat (ExperimentMFC) segítségével a szerzők részvételével. Minden megnyilatkozásról döntést hoztunk abban a tekintetben, hogy a prozódiai megvalósulása alapján „kérdő”, „kijelentő”, esetleg „felkiáltó”, vagy egyik csoportba sem tartozik („problémás” – ezek jobbára olyan megnyilatkozások voltak, amelyek felsorolásszerűen, emelkedő dallammal valósultak meg). A percepciós teszt elvégzésére a DS-megnyilatkozások esetében azért volt szükség, mert a mondat

megvalósítása nem mindig történt a kísérletben elicitálni kívánt szándéknak megfelelően. A KO-beszélők esetében a megnyilatkozások e szándék szerint valósultak meg.
A teljes „kérdő" és „kijelentő" megnyilatkozások normalizált időtartamában 1%-onként nyertük ki az f_0-értékeket, majd félhangokká konvertáltuk 50 Hz-es referenciaértékkel. Az f_0-görbéket általánosított additív kevert modellekkel (GAMM) vetettük össze, beszélőnként külön-külön: a modellben az f_0-érték lefutását vizsgáltuk a megnyilatkozás normalizált időtartamán belül, a mondattípus kétszintű parametrikus faktorváltozóval és tokenenként illesztett random smooth funkcióval.

Eredmények. A percepciós teszt alapján mindösszesen 47 három szótagos megnyilatkozás prozódiáját találtuk egyöntetűen a „kérdő" vagy a „kijelentő" csoportba tartozónak (függetlenül attól, hogy mi volt a megadott szándék a kísérletben). A 8 beszélő közül 4 esetében tudjuk vizsgálni a kérdő vs. kijelentő prozódiai kontrasztot (DS004: 2 kérdő vs. 9 kijelentő; DS005: 4 kérdő vs. 3 kijelentő; DS006: 5 kérdő vs. 2 kijelentő; DS008: 2 kérdő vs. 2 kijelentő).
A kérdő és kijelentő funkciót sikeresen megkülönböztető 4 beszélő három szótagos megnyilatkozásain lefuttatott GAMM-ok eredményei a kérdő és kijelentő f_0-görbéket a megnyilatkozás középső szakaszában, azaz a három szótagos megnyilatkozás utolsó előtti szótagjában különböztették meg, ahol az f_0-emelkedést várjuk – hasonlóan a KO-beszélőkhöz (1. ábra). Az emelkedő-ereszkedő kontúron belül mindkét beszélői csoportban nagymértékű variabilitás tapasztalható.

Következtetések. A vizsgálat DS-beszélőinek fele valósította meg sikeresen (legalábbis az itemek egy részében) a kérdő vs. kijelentő prozódiai kontrasztot a kísérleti helyzetben a három szótagos megnyilatkozásokban, más megfogalmazásban a DS-beszélők felénél nem tudtunk kimutatni prozódiai kontrasztot a három szótagos megnyilatkozásokban. (Ez természetesen nem jelenti azt, hogy az egy és két szótagos itemekben sem valósult meg a prozódiai kontraszt.) Arra a kérdésre, hogy mi áll a megkülönböztetés hiányának a hátterében a beszélők felénél, nem adható egyértelmű válasz. Nehézséget jelent ugyanis a kutatási kérdés vizsgálatában, hogy nem áll rendelkezésre jól bevált módszertan a prozódiai kontraszt elicitálására a DS-beszélők körében, illetve, hogy nagyok az egyéni különbségek a populáción belül. Ugyanezen DS-beszélők spontán beszédében nagymértékű perszeveráció és echolália figyelhető meg, ami miatt az egyes realizációk kérdő vagy kijelentő formája nem mindig feleltethető meg az elicitált funkciónak (például egy kijelentő forma helyett kérdő formát valósít meg a beszélő, mert az előző item kérdő funkciójú volt – ezért is döntöttünk úgy, hogy nem a kísérleti szándékot vesszük alapul a funkciók elemzésében, hanem a realizációt, amelyet a percepciós teszt segítségével minősítettünk). A próbavizsgálat tanulságai alapján más módszertanok kipróbálására is szükség van.

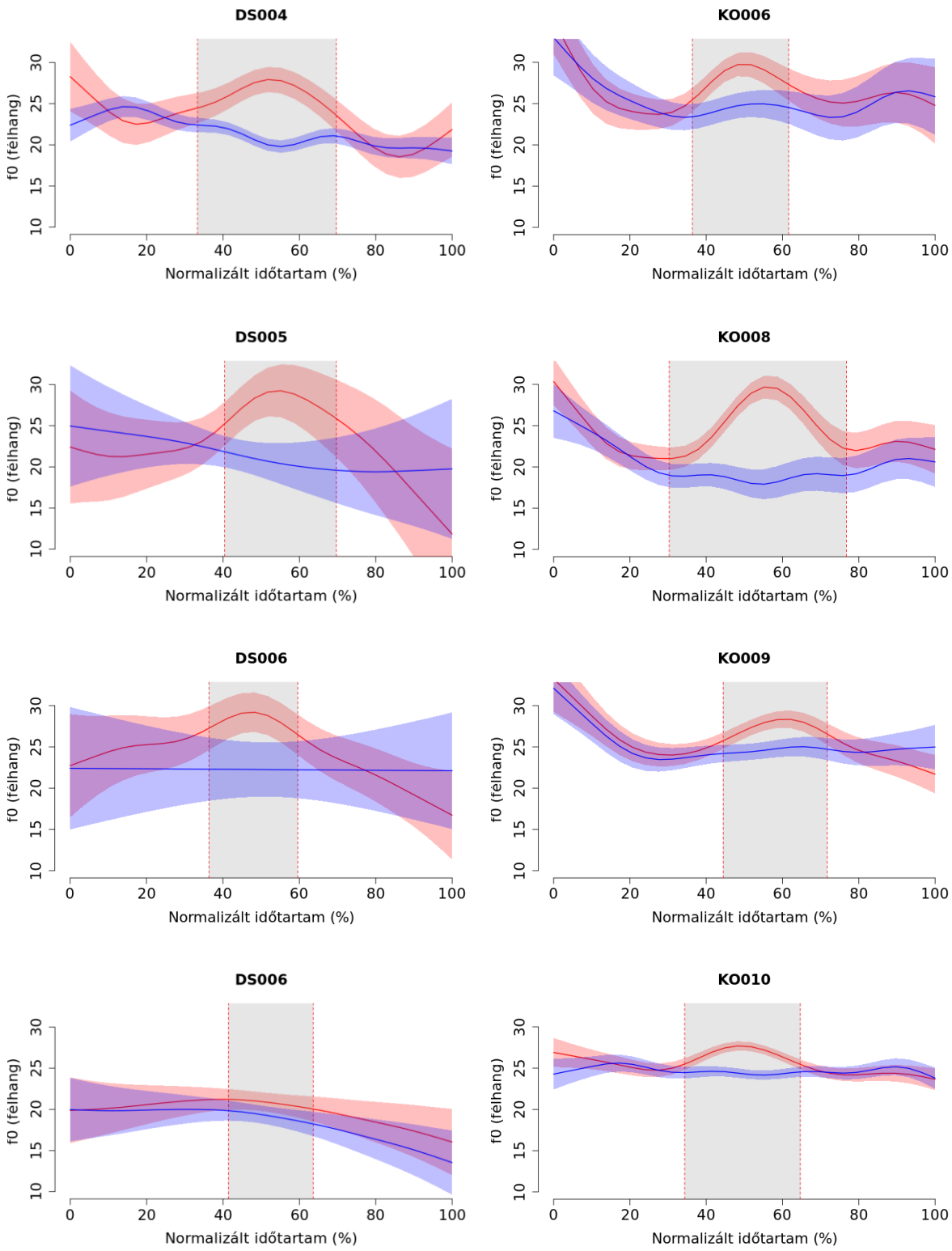
Irodalom

[1] Homor, V. (2008). *Down-szindrómások és más értelmi akadályozottak spontán beszédének összehasonlítása*. Szakdolgozat. ELTE BGGYK, Budapest.

[2] Rohár, A. (2016). 'Down-szindrómás személyek nyelvi képessége'. *Gyógypedagógiai Szemle* 44(3), 195–215. Elérhető: https://epa.oszk.hu/03000/03047/00074/pdf/.

[3] Loveall, S. J., K. Hawthorne & M. Gaines (2021). 'A meta-analysis of prosody in autism, Williams syndrome, and Down syndrome'. *Journal of Communication Disorders* 89, 106055. DOI: https://doi.org/10.1016/j.jcomdis.2020.106055.

[4] Jankovics, J. (2022). *Enyhe értelmi fogyatékossággal élő fiatalok beszédének szupraszegmentális fonetikai elemzése*. PhD-disszertáció. ELTE BTK, Budapest.



1. ábra. A Down-szindrómával diagnosztizált (balra) és a kontroll személyek (jobbra) három szótagú kérdő (piros) és kijelentő (kék) becsült f0-kontúrjai beszélőnként megjelenítve és a megnyilatkozás normalizált időtartamára vetítve. A szürke sáv a mondattípusok közötti szignifikánsan eltérő f_0-szakasz

Descriptions and terminology of speech production phenomena related to fear and anxiety in cuneiform sources of the Ist millennium BC

Adrienn Orosz
Eötvös Loránd University, Budapest

1. Topic Description and Relevance

An increasing number of studies have highlighted the complex effects of fear and anxiety on speech production and the accompanying paralinguistic phenomena (for some areas, see e.g. the bibliographies and research overviews of [1], [2], [3], and [4]). These studies identify the types of errors made during feelings of fear and anxiety, link these slips to specific levels of the speech production process, and discuss the various forms of errors considered pathological as well as their relation to anxiety. Research now also suggests that phenomena in speech production may serve to recognize anxiety and depressive disorders, prompting experts to incorporate speech production analysis into psychiatric screening and monitoring practices [3]. Although this might lead us to believe that the relationship between fear and speech is primarily a modern concern, we have written sources indicating that the people of the ancient Near East also noticed the various manifestations of fear in speech production phenomena, and described them with specific terminology and paraphrases. In my presentation, I aim to identify and show you some of the specific words and composite expressions used to describe these speech production phenomena within first millennium BC cuneiform sources. The overview of these expressions will offer the audience insights into how our predecessors in the ancient Near East understood the effects of fear on speaking, a phenomenon that today lies at the crossroads of psycholinguistics, cognitive linguistics, and linguistic pragmatics.

2. Methodology

To collect the relevant terminology and descriptions, I use text editions (e.g. [5], [6], royal inscriptions, medical texts [7]), databases of published and unpublished cuneiform tablets ([8], ([5], [9]), lexical lists in Sumerian and Akkadian ([10]), as well as modern Akkadian dictionaries ([11], [12], [13], [14], [15], [16]). Beyond the terminological analysis, I also show the textual context and purpose of the texts (e.g., oracles and reports, medical texts, literary works).

3. Results

Based on my research, three types of terminology can be distinguished in the Akkadian sources: terms used specifically to refer to fear-related speech production phenomena (e.g., the verb *parādu* in Gt and Dt stems, and in the form of verbal adjective, *pitruda*), compounds and descriptions in close connection with *parādu* (e.g., *pardiš iddanabbub*; *ēteqim dabābi,* see Fig. 1 for the latter), and terms that refer to a broader category beyond strictly fear-related speech phenomena (e.g., the verb *haṭû* in the Dt stem as compared to *parādu* based on [5] in Fig. 2). I provide examples of these, together with new proposals for their translation, taking into account and reevaluating the results of earlier research.

4. References

[1] Craig, A., Tran, Y. (2006). 'Fear of speaking: chronic anxiety and stammering'. *Advances in Psychiatric Treatment* 12, 63–68, http://apt.rcpsych.org/. DOI: 10.1192/apt.12.1.63.

[2] Hofmann, S. G. et al. (2012). 'Linguistic Correlates of Social Anxiety Disorder'. *Cogn Emot.* 26(4)720–726. DOI: 10.1080/02699931.2011.602048.

[3] Teferra, B. G. et al. (2022). 'Acoustic and Linguistic Features of Impromptu Speech and their Association with Anxiety: Validation Study'. JMIR Mental Health Preprint, https://preprints.jmir.org/preprint/36828.

[4] Zhao, N., Cai, Z.G., Dong, Y. (2023). 'Speech errors in consecutive interpreting: Effects of language proficiency, working memory, and anxiety'. *PLoS ONE* 18(10): e0292718. DOI: 10.1371/journal.pone.0292718.

[5] ORACC / SAAO. The Open Richly Annotated Cuneiform Corpus. State Archives of Assyria Online: https://oracc.museum.upenn.edu/saao/.

[6] SAA 4. Starr, I. (1990). *Queries to the Sungod. Divination and Politics in Sargonid Assyria*. State Archives of Assyria 4. Helsinki: Helsinki University Press.

[7] CT 46 49. Lambert, W. G. and Millard, A. R. (1965). *Cuneiform Texts from Babylonian Tablets in the British Museum. Part XLVI. Babylonian Literary Texts*. London: The Trustees of the British Museum.

[8] CDLI. Cuneiform Digital Library Initiative: https://cdli.earth.

[9] BABMED. Babylonische Medizin: https://www.geschkult.fu-berlin.de/e/babmed/Corpora/index.html.

[10] ORACC / DCCLT. The Open Richly Annotated Cuneiform Corpus. DCCLT: Digital Corpus of Cuneiform Lexical Texts: https://oracc.museum.upenn.edu/dcclt/.

[11] CAD D. Oppenheim, L. A. et al. (eds). (1959). *The Assyrian Dictionary of the Oriental Institute of the University of Chicago. Vol. 3. D.* Chicago (IL): The University of Chicago.

[12] CAD E. Oppenheim, L. A. et al. (eds). (1958). *The Assyrian Dictionary of the Oriental Institute of the University of Chicago. Vol. 4. E.* Chicago (IL): The University of Chicago.

[13] CAD Ḫ. Oppenheim, L. A. et al. (eds). (1956). *The Assyrian Dictionary of the Oriental Institute of the University of Chicago. Vol. 6. Ḫ.* Chicago (IL): The University of Chicago.

[14] CAD P. Roth, M. T. et al. (eds). (2005). *The Assyrian Dictionary of the Oriental Institute of the University of Chicago. Vol. 12. P.* Chicago (IL): The University of Chicago.

[15] AHw. Soden, W. von. (1959–1981). *Akkadisches Handwörterbuch*. Vols. 1–3. Wiesbaden: Harrassowitz.

[16] CDA. Black, J. et al. (eds). (2000). *A Concise Dictionary of Akkadian*. Wiesbaden: Harrassowitz.

5. Figures and illustrations

CT 46 49 i 10: *ēteqim dabābi*

Figure 1.

SAA 4
parādu: only in reference to speech (100 [103]), Starr's translation: "*jumbled*"
haṭû: not only in reference to speech, Starr's translation: "*impaired*" (8), "*faulty*" (4), Ø (2)

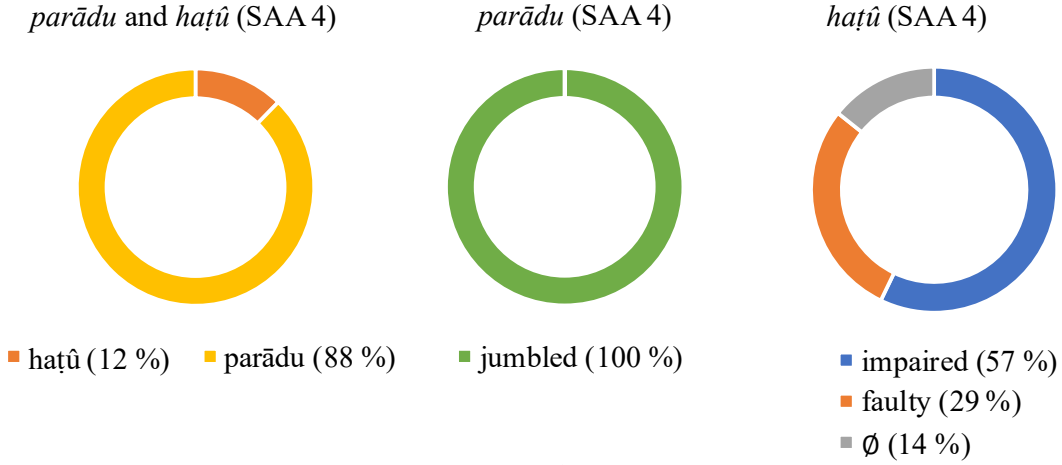


Figure 2.

A glottalizált zöngeminőség észlelése a magyar beszédben

Pirsel Katalin[1,2,3] és Kohári Anna[1]
[1]ELTE Nyelvtudományi Kutatóközpont
[2]ELTE Eötvös Loránd Tudományegyetem
[3]MTA–HUN-REN NYTK Lendület Neurofonetikai Kutatócsoport

Bevezetés. A különböző zöngeminőség-típusokat (glottalizáció, leheletes és modális zönge) a hangszalagok működésének változtatásával hozzuk létre [2]. A zöngeminőség a legtöbb nyelvben, így a magyar nyelvben sem tekinthető fonológiailag jelentésmegkülönböztető jellemzőnek, azaz módosításával a szó lexikai jelentése nem változik [5]. A zöngeminőség-változást azonban nemcsak azokban a nyelvekben észlelik a hallgatók, amelyekben fonológiailag disztinktív jegyként funkcionál. Az amerikai és ausztrál angolban, valamint mandarin kínaiban is kimutatták, hogy a beszélők képesek meghallani a glottalizált és modálisan ejtett magánhangzók közti különbséget, pedig ezen nyelvek fonémarendszerében sem funkcionál a magánhangzók zöngeminősége jelentésmegkülönböztető tulajdonságként [1, 4, 8]. A hangzókon észlelhető zöngeminőségbeli különbségek egyrészt jelölhetik a prozódiai határokat vagy eseményeket, másrészt pragmatikai többletjelentéseket hordozhatnak [2, 5].

A magyar beszédben a glottalizáció gyakran jelenik meg prozódiai határok, főként a frázisvégek jelölőjeként [5]. A glottalizáció percepciója azonban mindeddig kevésbé került a kutatások fókuszába. Markó és munkatársai [7] úttörő munkájukban azt vizsgálták, hogy a glottalizáció megjelenése a beszédben hatással van-e a beszélő tulajdonságainak (pl. barátságosság, okos, agresszív stb.) megítélésére. A tanulmányban nem találtak egyértelmű és konzisztens összefüggést azzal kapcsolatban, hogy a hallgatók másképp értékelték volna a beszélők személyes tulajdonságait a beszédben megjelenő glottalizáció hatására a modálisan ejtett megnyilatkozásokhoz képest. Vizsgálatukban egyrészt hosszú, sok szótagos megnyilatkozásokat használtak, amelyek nehezíthetik a glottalizáció észlelését és elkülönítését a modális megvalósulásoktól. Továbbá ha a hallgatók észleltek is valamilyen eltérést a beszédben, nem jelenti azt, hogy konzisztensen valamilyen tulajdonságokat is társítanak hozzá a magyar beszédben. Ezért vizsgálatunk célja annak feltárása volt, hogy a magyar anyanyelvű beszélők vajon képesek-e elkülöníteni a modális zöngét a glottalizálttól rövid megnyilatkozásokban.

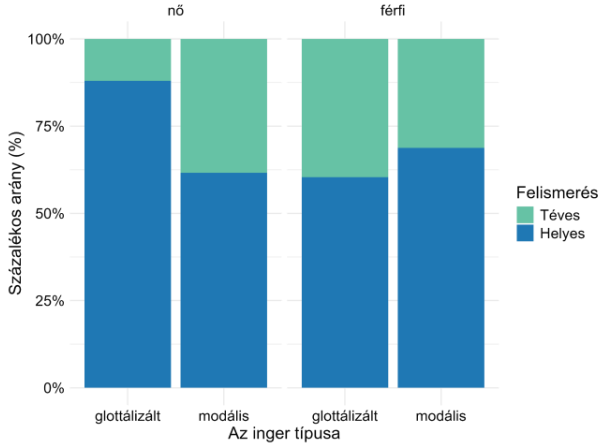
Anyag és módszer. Elemzéseinkhez egy női és egy férfi beszélő anyagát rögzítettük stúdiókörülmények között omnidirekcionális fejmikrofonnal (DPA 4066). Az ingeranyag létrehozásában követtük White és munkatársai [8] módszertanát adaptálva a magyar beszéd sajátosságaira. Olyan egy szótagos mellékneveket kerestünk, amelyek CVC szerkezetűek és mássalhangzóként szonoránsokat vagy zöngés obstruenseket tartalmaznak, valamint olyan főneveket kerestünk, amelyekben a kóda szonoráns vagy nazális volt, és szótagkezdete (onset) a magánhangzótól jól elválasztható zöngétlen obstruens volt. Ezekből az egy szótagos szavakból értelmes, jelentésében összeillő szókapcsolatokat hoztunk létre (pl. *bal fül, mély tál*), amelyben a melléknevet követte a főnév, és a főnévben szereplő magánhangzó változatos volt képzésmódját tekintve. A főnév rím (rhyme) modális részét egy Praat script segítségével úgy manipuláltuk, hogy glottalizált legyen [8]. Mind a modális, mind a glottalizált ingeranyagon manipuláltuk az alapfrekvenciát úgy, hogy a férfi, a női beszélő esetében is 2-2 kategóriát hoztunk létre, egy alacsonyabb és egy magasabb alapfrekvenciával rendelkező ingert. Az ingerek alapfrekvenciájának meghatározásához a magyar beszédben jellemző átlagos értékeket vettük

alapul [3]. A hangfájlok intenzitását egységesen átlagosan 70 dB-re módosítottuk. Az így kapott ingereket a PsyToolKit online felületén keresztül hallgattattuk meg 25 magyar anyanyelvi beszélővel. Először kaptak egy rövid leírást a glottalizációról, majd hallgatható mintafájlokat a jelenségről (hasonlóan [1, 8]). Ezek után azt a kérdést tettük fel nekik, hogy a beszédrészlet tartalmaz-e glottalizációt. Kevert logisztikus regressziós modellt alkalmaztunk, amelyben a független változó a percepciós tesztben adott válasz alapján a felismerés pontossága (az inger zöngeminőségének megfelelő felismerés vs. az inger zöngeminőségétől eltérő válasz) volt. Függő változóként az inger alapfrekvenciájának kategóriáját (magasabb vs. alacsonyabb) és a beszélő nemét (férfi vs. nő), valamint ezek interakcióit adtuk meg. Random hatásként figyelembe vettük az ingerek különböző szószerkezeteit és a hallgatókat. A statisztikát az R 4.5.0 szoftver segítségével hajtottuk végre.

Eredmények. A piloteredmények azt mutatták, hogy a magyar anyanyelvi beszélők is képesek felismerni a glottalizált szótagokat és elkülöníteni a modálisoktól. A hallgatók a glottalizációt tartalmazó ingerek 74,2%-ában ismerték fel, míg az esetek 25,8%-ban tévesen jelölték, hogy a meghallgatott szótagok nem tartalmaznak semmilyen glottalizációt. A csak modális zöngét tartalmazó ingereknél 65,2% volt a helyes felismerés, és a hallgatók 34,8%-ban tévesen glottalizáltként azonosították. A Mcnemar-teszt alapján a hallgatók nem véletlenszerűen válaszoltak, hanem valóban képesek voltak percepciójuk alapján a glottalizáció felismerésére és a modális zöngétől való elkülönítésre ($p < 0,001$). Megvizsgáltuk az észlelés pontosságára ható tényezőket is. Az inger zöngeminőségének típusa, az inger beszélőjének neme és az inger alapfrekvenciája (magasabb vs. alacsonyabb) is hatással volt ($p \leq 0,003$) a zöngeminőség felismerésének pontosságára a logisztikus regressziós modell eredményei alapján. Ugyanakkor ezek a ható tényezők egymással interakcióban voltak (1. ábra). A Tukey-féle post hoc teszt alapján a glottalizált zöngét tartalmazó inger felismerése nagyobb arányban volt sikeres a női beszélő ejtésében, mint a férfi beszélőében, míg a modális zönge felismerése a férfi beszélő ejtésében bizonyult könnyebbnek a hallgatók számára ($p \leq 0,003$). Másfelől a glottalizáció felismerésének pontossága nagyobb volt alacsonyabb alapfrekvencia mellett, mint magas alapfrekvencia esetében, modális zönge esetében viszont magasabb alapfrekvencia esetében találták el gyakrabban az inger zöngeminőségét ($p \leq 0,02$). A felismerés pontossága az alacsony és a magas alapfrekvenciájú ingernél is a női beszélő ejtésében volt magasabb a férfi beszélőhöz képest ($p \leq 0,02$). A ható tényezők hármas interakciója nem mutatott szignifikáns eltérést.

Következtetések. Eredményeink szerint annak ellenére, hogy a magyarban a zöngeminőség nem fonológiailag disztinktív jegy, magyar anyanyelvi hallgatók is képesek felismerni a glottalizációt és elkülöníteni a modális zöngétől kontrollált körülmények között. Az eredmények azt mutatták, hogy az alacsonyabb alapfrekvenciájú ingerek torzíthatják a hallgatók észlelését, mivel ezek esetében nagyobb valószínűséggel glottalizált zöngekategóriába sorolták az ingert akkor is, amikor valójában csak modális zöngét hallottak. Hasonló eredményt találtak ausztrál angolban White és munkatársai [8]. Ugyanakkor a hallgatók az elvárásaikat a beszélő nemének hangterjedelméhez képesek igazítani, hiszen női beszélő ejtésében magasabb arányban ismerték fel a glottalizációt, mint férfi beszélő esetében. Elképzelhető, hogy a nők körében gyakoribb glottalizációs produkció [6] a hallgatókban olyan elvárást alakít ki, amely női beszélők esetében fokozza a glottalizáció észlelésének valószínűségét. Az alapfrekvencia és a beszélőre jellemző hangterjedelem összefüggéseinek megértése még további vizsgálatokat igényelnek több beszélő és több kategória bevonásával, ugyanakkor az eredmények egyértelműek a tekintetben, hogy a magyar beszédben a glottalizáció tudatosan is észlelhető akusztikai jelenség.

Köszönetnyilvánítás. A vizsgálatot az MTA Bolyai János Kutatási Ösztöndíja (BO/00160/25/1) támogatta.



1. ábra. Külön a női és külön a férfi beszélő ejtésében a glottalizált és modális zöngével rendelkező hangingerek felismerési aránya a hallgató által

Irodalom

[1] Davidson, L. (2019). Perceptual coherence of creaky voice qualities. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1961–1965). Australasian Speech Science and Technology Association.

[2] Garellek, M. (2022). Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics*, 94, 101155.

[3] Gráczi, T. E., Krepsz, V., Markó, A., Huszár, A., & Száraz, B. (2019). Az f0-jellemzők felolvasásban és spontán beszédben. *Alkalmazott Nyelvtudomány*, *19*(2).

[4] Li, A., Lai, W., & Kuang, J. (2022). How do listeners identify creak? The effects of pitch range, prosodic position and creak locality in Mandarin. In *Proceedings of Speech Prosody* (Vol. 2022, pp. 480–484).

[5] Markó, A. (2013). *Az irreguláris zönge funkciói a magyar beszédben*. ELTE Eötvös Kiadó.

[6] Markó, A. (2014). Frequency of glottalization in Hungarian read and spontaneous speech. In *Proceedings of the 10th International Seminar on Speech Production (ISSP)* (pp. 269–272).

[7] Markó, A., Csapó, T. G., & Takács, K. (2017). Listeners' evaluation of voice quality in Hungarian speakers. *Beszédkutatás, 25*, 55–66.

[8] White, H., Penney, J., Gibson, A., Szakay, A., & Cox, F. (2024). Influence of pitch and speaker gender on perception of creaky voice. *Journal of Phonetics, 102,* 101293.

Érzelmek felismerése prozódiából tipikus és atipikus fejlődésben

Svindt Veronika
ELTE Nyelvtudományi Kutatóközpont
MTA-HUN-REN Neurofonetikai Kutatócsoport

Bevezetés. A szociális kommunikáció kevésbé kutatott, noha a hétköznapi kommunikációs helyzetekben való boldogulást jelentős mértékben meghatározó eleme az érzelmi prozódia felismerésére és produkciójára való képesség. Számos olyan kommunikációs helyzet van, amelyben a szó szerinti jelentés önmagában nem ad kapaszkodót egy-egy megnyilatkozás megértésére vonatkozóan, mert a beszélő által szándékozott jelentéstartalom megértéséhez – többek között – a beszéd bizonyos szupraszegmentális tényezőit is megfelelően kell értelmeznünk. Tipikusan ilyen információ például a jellegzetes ironikus intonáció, de ide tartozik az érzelmek kifejezése a prozódia segítségével is. Amennyiben a hallgató nem képes észlelni és/vagy azonosítani a prozódia által hordozott többletjelentést, a beszélő kommunikációs szándéka fel nem ismert marad, amely félreértéshez vezethet.

Korábbi kutatások azt mutatják, hogy bizonyos életkor előtt a gyermekek nem képesek a prozódiából következtetéseket levonni az egyén attitűdjeire, illetve érzelmi állapotaira vonatkozóan [1, 4, 6, 7]. Ez a képesség tipikus fejlődés esetén viszonylag gyors ütemben fejlődik. Bizonyos fejlődési zavarokban (pl. ASD, ADHD), amelyekben a szocioemocionális kogníció korlátozott vagy megkésett fejlődése gyakran fennáll, a prozódia által hordozott többletjelentések megértésének nehezítettségével is számolnunk kell [2, 3, 9, 5].

Jelen vizsgálat célja annak alaposabb megismerése, hogy a tipikus és atipikus fejlődés során a gyermekek milyen életkorban válnak képessé az alapérzelmek (öröm, harag, félelem, meglepődés, szomorúság, undor) prozódiából történő felismerésére, illetve segítséget jelent-e számukra, ha a prozódia mellett a kontextus is segíti a megértést.

Módszertan. A vizsgálat két mondat–kép párosítási feladatból állt. Az első, baseline feladatban a gyermekek valamely érzelmet expliciten megnevező mondatokat hallottak, melyekhez az adott érzelmet megjelenítő arcot kellett kiválasztaniuk. A második feladatban az érzelmi jelentéstartalmat a prozódia hordozta; a feladat szintén az adott érzelemhez tartozó arc kiválasztása volt. E feladat két kondícióban 8-8 mondatot tartalmazott. Az első kondícióban az érzelmeket csak prozódia alapján kellett felismerni, egyszerű mondatszerkezetben (*„Ez egy X.”*). A második kondícióban az érzelemfelismerést szemantikai segítség is támogatta, a mondatok szerkezete ennek megfelelően *„Ez egy [valamilyen] X.”* volt. Mindkét kondícióban a 6 alapérzelem mellett 2-2, érzelmileg neutrális mondatprozódiát és arckifejezést tartalmazó item is szerepelt a feladatban.

Az arckifejezések az *Amsterdam Dynamic Facial Expression Set (ADFES)* [9] kutatási célokra létrehozott, validált eszköztárából származtak. A mondatokat tapasztalt, gyakorló színészek (1 férfi, 1 nő) mondták föl hanghordozóra a megfelelő érzelmi intonációval.

Résztvevők. A vizsgálatban összesen 160 gyermek vett részt, 103 fő tipikus fejlődésű (átlag életkor: 7;6 év) és 57 fő atipikus fejlődésű (átlag életkor: 9 év) gyermek. Az atipikus csoportba az alábbi diagnózisokkal kerültek a gyermekek: 28 fő ADHD, 11 fő autizmusspektrum zavar (ASD), 10 fő ASD+ADHD kettős diagnózis, és 8 fő a sajátos magyar jogi kategóriából: beilleszkedési, tanulási és magatartási nehézség (BTMN) státusszal rendelkező gyermekek.

A vizsgálatban való részvétel tájékoztatott beleegyezés után, a szülő írásos hozzájárulásával, önkéntes alapon, megfelelő etikai engedélyek birtokában történt. Az atipikus csoportba tartozó gyermekek érvényes gyermekpszichiátriai diagnózissal, illetve szakértői bizottság által megállappított BTMN státusszal rendelkeztek. A statisztikai elemzéshez az életkori, és a tipikus vs. atipikus csoportok összehasonlítására Mann-Whitney U tesztet, ill. Kruskal-Wallis tesztet alkalmaztunk.

Eredmények

*a) Tipikus fejlődés. A*z eredmények azt mutatják, hogy az expliciten kimondott érzelmeket már az 5 évesek is megbízhatóan felismerik. A prozódia+szemantikai segítség kondícióban az 5;0–5;11 éves korosztály szignifikánsan alacsonyabb teljesítményt ért el az összes többi korcsoporttal összehasonlítva. Hét éves kortól fölfele a csoportok egymástól nem különböztek, ami arra utal, hogy amennyiben az érzelmi prozódiát valamilyen szemantikai segítség is támogatja, a gyerekek körülbelül iskoláskortól képesek azt megbízhatóan azonosítani. A csak prozódia általi érzelemfelismerés bizonyult a legnehezebbnek tipikus fejlődésben is, itt plafonhatást csak a 9 évesnél idősebb gyermekeknél látunk.

b) Atipikus fejlődés. A tipikus és atipikus fejlődésű csoport összehasonlításába csak a 6 évesnél idősebb gyermekeket vontuk be (tipikus: 78 fő; atipikus: 52 fő). Az eredmények azt mutatják, hogy a baseline kondícióban, vagyis ahol az érzelmet az expliciten kifejezett szövegből kellett azonosítani, nem volt különbség a tipikus és az atipikus fejlődésű csoport között (U=1880,5, n=130, p=,434). A prozódiafelismerési feladatban a két kondíciót összesítve az atipikus fejődésű csoport szignifikánsan alacsonyabb teljesítményt ért el a tipikus fejlődésű gyermekekhez képest (U=919,5, n=129, p <,001). Abban a kondícióban azonban, ahol a prozódia mellett szemantikai segítség is rendelkezésre állt, nem volt szignifikáns különbség a két csoport teljesítménye között (U=1688, n=129, p = ,149). Ha az adott érzelmet csak a prozódia alapján kellett felismernie a gyermekeknek, az atipikus csoport szignifikánsan alacsonyabb teljesítményt nyújtott, mint a tipikus fejlődésűek (U=1317, n=129, p <,001).

Az egyes atipikus fejlődésű csoportokat külön-külön vizsgálva azt találtuk, hogy amennyiben az érzelmet csak a prozódiából kellett felismerni, akkor az ADHD és a BTMN csoport nem különbözött a tipikus fejlődésű csoporttól (ADHD: H(4)=13,802, p=,837; BTMN: H(4)=28,362, p=,325), az ASD és az ASD+ADHD csoport azonban szignifikánsan alacsonyabb teljesítményt nyújtott (ASD: H(4)=35,432, p=0,49; ASD+ADHD: H(4)=36,042, p=,042). A tipikus fejlődésű csoport átlagához képest 1,5 szórásnyival alacsonyabb teljesítményt nyújtó gyerekek aránya nagyon eltérő az egyes csoportokban. A tipikus fejlődésű gyermekek 90%-a átlagos eredményt nyújtott a feladatban. Ezzel szemben az ADHD gyermekek 41%-ának, az ASD gyermekek 64%-ának, az ASD+ADHD kettős diagnózisú gyermekek 80%-ának okozott nehézséget az érzelmek prozódiából történő felismerése.

Következtetések. Az eredmények azt mutatják, hogy azokban az esetekben, ahol egy alapérzelmet expliciten kimondott szövegből kell azonosítani, a tipikus és az atipikus fejlődésű csoport között nincsen teljesítménybeli különbség, ami arra utal, hogy az alapérzelmek azonosítása önmagában nem jelent problémát az atipikus fejlődésű csoportnak sem. Hasonló a helyzet abban az esetben, ha az érzelmi prozódia megértését egy szemantikai elem támogatja. A csak szupraszegmentális csatornán keresztül érkező információ feldolgozása azonban kihívást jelenthet a vizsgált atipikus fejlődésű csoportokban még 10 éves kor körül is.

Szakirodalom

[1] Aguert, M., Laval, V., Lacroix, A., Gil, S., & Le Bigot, L. (2013) Inferring emotions from speech prosody: Not so easy at age five. *PLoS ONE* 8(12): e83657.

[2] Ashwin, C., Chapman, E., Colle, L., & Baron-Cohen, S. (2006) Impaired recognition of negative basic emotions in autism: A test of the amygdala theory. *Social Neuroscience* 1: 349–363.

[3] Balconi, M., Amenta, S., & Ferrari, C. (2012) Emotional decoding in facial expression, scripts and videos: A comparison between normal, autistic and Asperger children. *Research in Autism Spectrum Disorders* 6: 193–203.

[4] Chronaki, G., Hadwin, J. A., Garner, M., Maurage, P., & Sonuga-Barke, E. J. (2015) The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. *The British Journal of Developmental Psychology 33*(2): 218–236.

[5] Greco, C., Romani, M., Berardi, A., De Vita, G., Galeoto, G., Giovannone, F., Vigliante, M., & Sogos, C. (2021) Morphing task: The emotion recognition process in children with Attention Deficit Hyperactivity Disorder and Autism Spectrum Disorder. *International Journal of Environmental Research and Public Health 18*(24): 13273.

[6] Morton, J. B., & Trehub, S. E. (2001) Children's understanding of emotion in speech. *Child Development 72*(3): 834–843.

[7] Sauter, D. A., Panattoni, C., & Happé, F. (2013) Children's recognition of emotions from vocal cues. *The British Journal of Developmental Psychology*, *31*(Pt 1): 97–113.

[8] van den Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011) Moving Faces, Looking Places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES*). American Psychological Association,* 11(4): 907–920. https://aice.uva.nl/research-tools/adfes-stimulus-set/adfes-stimulus-set.html?cb

[9] Yoshimatsu, Y., Umino, A., & Dammeyer, J. (2016) Characteristics of the Understanding and Expression of Emotional Prosody among Children with Autism Spectrum Disorder. *Autism - Open Access* 6(4) doi: 10.4172/2165-7890.1000185

Szorongás és hezitálás spanyol mint idegen nyelven: egy nyelvi és kommunikációs tréning néhány eredménye

Szánthó Zsuzsa[1,2,3]
[1]ELTE Eötvös Loránd Tudományegyetem
[2]MTA–HUN-REN Lendület Neurofonetikai Kutatócsoport
[3]ELTE Nyelvtudományi Kutatóközpont

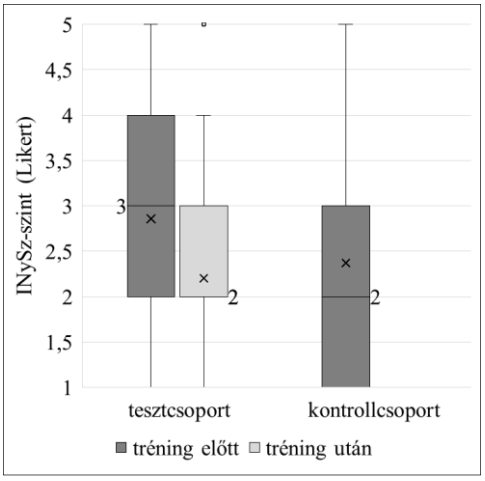
A jelen vizsgálatban magyarajkú spanyol nyelvtanulók körében vizsgáltam egy nyelvi és kommunikációs tréning hatását az idegen nyelvi szorongásra, valamint a spanyol nyelvű megszólalás egyes jellemzőire. A tréning az általános megközelítés szerint kis létszámú (5–20 fő), tapasztalati tanuláson alapuló csoportos képzés, amely hatékonyan alkalmazható az idegennyelv-oktatásban is, ugyanis javítja a résztvevők teljesítményét a nyelvhasználati helyzetekben [5]. A tréningmódszer a beszédkészség fejlesztésére, illetve ennek részeként a kommunikációs stratégiák oktatására szintén alkalmas lehet, és utóbbiba az idegen nyelven való hezitálás nyelvspecifikus formái is beletartoznak [3]. A hezitálás leggyakoribb formái magánhangzószerűek, nyelvenként változatos hangszínnel: a magyarban a leggyakoribb az [ə]-szerű [6], míg a spanyolban az [e]-szerű hezitálás [9]. A beszédben előforduló hezitálások további megvalósulási formája a nyújtás, amely egyes hangok „extrém” hosszú ejtését jelenti [4], [2]. A spanyolban leggyakrabban egyes magánhangzók és magánhangzószerű hangok hosszabbodhatnak meg [8]. Egy spanyol nyelvtanulók körében végzett kutatás szerint a magyar anyanyelvűek spanyol megszólalásaiban előforduló nyújtások nem a spanyol mintát követik: a vizsgálatba bevont spanyolok leginkább az [a] hangot nyújtották, míg a magyarok az [i] hangot [1]. Az idegennyelv-használat egyik gyakran vizsgált területe az idegen nyelvvel és annak használatával összefüggő nyelvtanulói attitűd. Ezek szerint a kutatások szerint az idegen nyelven való megszólalás egyik legjellemzőbb gátló jellegű jelensége a megszólalástól való félelem és aggódás, amit a szakirodalom idegen nyelvi szorongásnak (INySz) nevez [7]. A jelen vizsgálat kérdései a következők. Egy spanyolt tanuló diákoknak szóló nyelvi tréning önismereti és idegennyelvi gyakorlatai alkalmasak-e arra, hogy csökkentsék a résztvevők idegen nyelvi szorongását? Ezek a gyakorlatok elősegítik-e az idegen nyelvre jellemző kommunikációs stratégiáknak a célnyelvre jellemzőbb használatát, különös tekintettel a magánhangzószerű hezitálásokra és nyújtásokra? A vizsgálat hipotézisei a következők: a nyelvi és kommunikációs tréningen való részvétel hatására a tréning előtt mértekhez képest a tréning után 1) a résztvevők INySz-szintje alacsonyabb; a résztvevők spanyol nyelvű megnyilatkozásaiban 2) az [ə]-szerű hezitálások gyakorisága alacsonyabb; 3) az [e]-szerű hezitálások gyakorisága magasabb; és 4) a nyújtások inkább jelennek meg a spanyolban jellemző helyeken (az ott jellemzően érintett magánhangzókon).

A vizsgálatba 14 magyar anyanyelvű, legalább B2-es szintű (KER) spanyol nyelvtudással rendelkező nőt vontam be. Közülük 7 részt vett a tréningen (tesztcsoport), 7 nem vett részt rajta (kontrollcsoport). A kísérleti személyek INySz-szintjét ötfokozatú Likert-skálás attitűdelemzéssel [10], [11] mértem úgy, hogy tesztcsoportban a mérést a tréning előtt és után is elvégeztem. Az elemzés alapján a kísérleti személyeket további két csoportra, INySz-t mutatókra és INySz-t nem mutatókra osztottam: a tesztszemélyek közül 2 INySz-t mutató és 5 INySz-t nem mutató volt, míg a kontrollszemélyek kizárólag INySz-t nem mutatók voltak. A tréning hat alkalomból állt, amely egyéni és csoportos formában váltakozva önismereti, kommunikációs és spanyol nyelvi gyakorlatokból épült fel. A vizsgálat korpusza a kísérleti személyekkel csendes szobában készült összesen 35 db magyar és spanyol nyelvű interjú volt. A tesztszemélyekkel két spanyol nyelvű interjú készült (egy-egy a tréning előtt és után), a kontrollszemélyekkel egy spanyol nyelvű interjú. Mindkét csoporttal felvettünk egy magyar

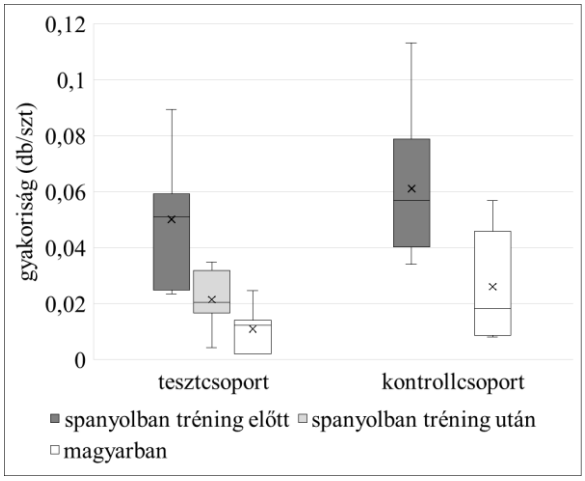
nyelvű interjút is. A hangfelvételeket két szempontból elemeztem: 1) megvizsgáltam a beszélők magyar és spanyol nyelvű megszólalásaiban megjelenő [ə]- és [e]-szerű hezitálások gyakoriságát, és összevetettem ezeket csoportonként; valamint 2) megvizsgáltam azt, hogy a tesztszemélyek spanyol megszólalásaikban mely hangokat nyújtották és milyen gyakorisággal a tréning előtt és után. A nyújtások vizsgálatakor a beszédhangok „normális" és „extrém hosszú" időtartamú megvalósulásai közötti különbségtételt percepciós teszttel végeztem el 5 fő spanyolul beszélő, beszédtudományokban jártas tesztelő bevonásával [2].

Az idegen nyelvi szorongással kapcsolatos attitűdelemzés szerint a tesztcsoportban az INySz-szint a tréning után a tréning előtt mértekhez képest csökkent, és elérte az INySz-t nem mutató kontrollcsoportban talált értéket: a tesztcsoportban az INySz-szint mediánja a tréning előtt 3, a tréning után 2 volt; míg a kontrollcsoportban 2 (1. ábra). A magánhangzószerű hezitálások gyakoriságának vizsgálata szerint az [ə]-szerű hezitálás gyakorisága a tesztcsoportban és a kontrollcsoportban a spanyolban közel megegyezett, a tesztcsoportban 0,051 db/szt, a kontrollcsoportban 0,057 db/szt volt a medián értéke. A tréning végére azonban a tesztcsoportban kevesebb mint a felére csökkent, 0,02 db/szt volt (2. ábra). A spanyolra jellemző [e]-szerű hezitálást a spanyolban a kísérleti személyek csak a tesztcsoportban, és csak a tréning után ejtettek, azonban ott is csak kis mértékben fordult elő: aránya az összes magánhangzószerű hezitálás 2,4%-a volt. A tréning után a spanyolban a tesztcsoportban a nyújtások gyakorisága összességében kis mértékben nőtt: a medián a tréning előtt 0,047 db/szt, a tréning után 0,053 db/szt volt. Ez a növekedés ugyanakkor nem a spanyol beszédben jellemzően nyújtott magánhangzót, az [a] hangot érintette, ugyanis az ejtett [a] hangok 5%-át nyújtották a tréning előtt és 3%-át a tréning után. Nőtt ugyanakkor a nyújtva ejtett [e], [i] és [o] magánhangzók aránya: az összes [e] hang 3,5%-át nyújtották a résztvevők a tréning előtt és 3,9%-át a tréning után, az összes [i] hang 4,5%-át nyújtották a tréning előtt és 5,9%-át a tréning után, valamint az összes [o] hang 7,1%-át nyújtották a tréning előtt és 7,9%-át a tréning után.

A vizsgálat eredményei alátámasztották az első hipotézist, hiszen a tréning előtt mértekhez képest a tréning után a tréning résztvevőinek INySz-szintje az INySz-t nem mutató kontrollcsoport INySz-szintjére csökkent. Az eredmények megerősítették a második és a harmadik hipotézist is, ugyanis a tréning után a résztvevők spanyol megszólalásaiban az [ə]-szerű hezitálás gyakorisága alacsonyabb volt, és körükben – bár kis mértékben, de – nőtt az [e]-szerű hezitálás gyakorisága. Bár a tesztcsoportban a tréning hatására a nyújtás gyakorisága összességében növekedett, az eredmények a negyedik hipotézist nem támasztották alá, ugyanis a nyújtás aránya a spanyolban jellemző helyen, azaz az [a] magánhangzón a tréning után a tréning előtt mértekhez képest csökkent, míg a spanyolban kevésbé jellemzően nyújtott [e], [i] és [o] nyújtásának gyakorisága nőtt. Mindezek alapján arra következtethetünk, hogy a tréningmódszer már egy rövidebb időtartamú foglalkozássorozatban is hozzájárulhat az idegennyelv-elsajátítás eredményességéhez, ugyanis elősegítheti a célnyelvvel kapcsolatos negatív attitűdök mérséklését, valamint hozzájárulhat egyes hezitálási formák célnyelvre jellemzőbb használatához.



1. ábra. Az INySz-szint a teszt- és kontrollcsoportban



2. ábra. Az [ə]-szerű hezitálás gyakorisága

Irodalom

[1] Baditzné Pálvölgyi K. (2019). ‘A megnyúlások és kitöltött szünetek mintázatai küszöbszinten álló magyar ajkú spanyol nyelvtanulók spontán beszédprodukciójában’. *Alkalmazott Nyelvtudomány*, *19*(2), 1–14. old.

[2] Deme A. (2012). ‘Magánhangzónyújtások gyermekek spontán beszédében’. *VI. Alkalmazott Nyelvészeti Doktoranduszkonferencia: Budapest*, 2012. 02. 03., 24–39. old.

[3] Dörnyei, Z. (1995). ‘On the Teachability of Communication Strategies’. *TESOL Quarterly*, *29*(1), 55–85. old.

[4] Gósy M. (2002). ‘A megakadásjelenségek eredete a spontán beszédtervezés folyamatában’. *Magyar Nyelvőr*, *126*(2), 192–204. old.

[5] Hilliard, A. (2014). ‘Spoken Grammar and Its Role in the English Language Classroom’. *English Teaching Forum*, *52*(4), 2–13. old.

[6] Horváth V. (2014). *Hezitációs jelenségek a magyar beszédben*. Budapest: ELTE Eötvös Kiadó.

[7] Horwitz, E. K., M. B. Horwitz & J. Cope (1986). ‘Foreign Language Classroom Anxiety’. *The Modern Language Journal*, *70*(2), 125–132. old.

[8] Machuca, M. J., J. Llisterri & A. Ríos (2015). ‘Las pausas sonoras y los alargamientos en español: Un estudio preliminar’. *Normas*, *5*, 81–96. old.

[9] Machuca, M. J. (2024). ‘An Acoustic Study on the Use of Fillers in Spanish as a Foreign Language Acquisition’. *Second Language Acquisition—Learning Theories and Recent Approaches*. Szerk. Maqbool, T. & L. Y. Lang. IntechOpen.

[10] MacIntyre, P. D. & R. C. Gardner (1994). ‘The Subtle Effects of Language Anxiety on Cognitive Processing in the Second Language’. *Language Learning. A Journal of Research in Language Studies*, *44*(2), 283–305. old.

[11] Piniel, K. (2024). *Investigating foreign language anxiety: Lessons for research into individual differences*. Cham: Springer Nature.

Ha csak egy dupla feketét szeretnél, de a hangodat nem érti senki…
– A személyre szabott gépi beszédelőállítás új lehetőségei hirtelen fellépő beszédzavarok esetén

Tóth László[1], Székely Éva[2] és Mihajlik Péter [3,4]
[1] Szegedi Tudományegyetem, Informatikai Intézet, Szeged
[2] Division of Speech, Music and Hearing
KTH Royal Institute of Technology, Stockholm
[3] ELTE- Nyelvtudományi Kutatóközpont, Fonetikai Csoport, Budapest
[4] Budapesti Műszaki és Gazdaságtudományi Egyetem, Villamosmérnöki és Informatikai Kar, Távközlési és Mesterséges Intelligencia Tanszék, Budapest

E-mail: tothl@inf.u-szeged.hu, szekely@kth.se, mihajlik.peter@nytud.hu

A hirtelen beszédképesség-vesztést elszenvedő egyének számára a személyre szabott szintetikus hang nem csupán kommunikációs eszköz, hanem az identitás folytonosságának megőrzését is szolgálhatja. Míg a beszédbankolás (tervezett hangminták felvétele és tárolása várható beszédkészségvesztés esetére) lehetősége felkészülési időt igényel, a stroke-on átesett személyeknek gyakran csupán meglévő hangfelvételekre kell támaszkodniuk. De vajon egy egyetemi előadásokon alapuló TTS (Text To Speech) hang használható-e hétköznapi interakciókban, például pékségben történő rendelés során?

A szövegfelolvasó rendszerek fejlődése az utóbbi években robbanásszerű volt: a korábban jellemzően 4–100 órányi felvételt igénylő egyéni TTS-modellek helyét átvették a több tízezer órás nyelvi diverzitást lefedő alapmodellek [1], amelyek akár néhány másodperces referenciahang alapján is képesek beszélőspecifikus szintézisre. E technológiai áttörések különösen ígéretesek azok számára, akik váratlanul veszítik el beszédképességüket, de alkalmazhatóságuk a kiegészítő és alternatív kommunikáció (AAC: Augmentative and Alternative Communication) személyre szabásában még feltárásra vár.

Ebben az előadásban egy együtt-tervezésen (co-design) alapuló esettanulmányt mutatunk be egy olyan kutatási résztvevővel, aki stroke előtt beszédbankolási célú mondatokat és spontán előadásokat is rögzített – gépi tanulással és beszédtechnológiával foglalkozó professzori munkája részeként. Ez az egyedi adatállomány lehetővé tette különféle beszédelérhetőségi forgatókönyvek és beszélőadaptációs stratégiák szisztematikus összehasonlítását, beleértve azokat az eseteket is, amikor alig vagy egyáltalán nem áll rendelkezésre az adott személy korábbi természetes hangja.

A keletkező szintetikus hangokat minőség, beszélőhasonlóság, AAC-kompatibilitás és többnyelvűség szempontjából értékeltük. Eredményeink szerint akár minimális személyes felvétellel is elérhető kiváló TTS-minőség, míg körülbelül egy óra spontán beszéd jelentősen növeli a beszélőazonosság élményét és a hétköznapi interakciókra való alkalmasságot.

A hallgatói értékelések mellett kvalitatív keretrendszert is alkalmaztunk, amely a kutatási résztvevő szakértői és érintetti (AAC-felhasználói) perspektíváit is integrálja [2]. Ennek révén egy újfajta interakciós megközelítést is kipróbáltunk, ahol a beszédzavaros (dizartriás) hang mint

prozódiai prompt szerepelt a személyre szabott TTS-rendszerben – lehetőséget adva a felhasználónak a hangsúlyozás kontrollálására és ezáltal az üzenet átadásának alakítására.

Az eredmények alapján megvitatjuk, hogy miként alakítják a nagy nyelvi és beszédmodellek a percepciós és interakciós aspektusokat a hangrekonstrukció során, valamint hogy milyen módon lehet ezeket a rendszereket az egyéni kommunikációs szükségletek szolgálatába állítani.

Irodalom

[1] Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., Weber, J. (2024) XTTS: 'A Massively Multilingual Zero-Shot Text-to-Speech Model'. Proc. Interspeech 2024, 4978-4982, doi: 10.21437/Interspeech.2024-2016

[2] Dietz, A., Wallace, S. E., & Weissling, K. (2020). Revisiting the role of augmentative and alternative communication in aphasia rehabilitation. *American Journal of Speech-Language Pathology*, 29(2), 909–913.

A zárt *ë* realizációja a magyar nyelvjárásokban

Vargha Fruzsina Sára[1], Kocsis Zsuzsanna[1], Szabó Gergely[1,2,3]
[1]Nyelvtudományi Kutatóközpont
[2]Eötvös Loránd Tudományegyetem
[3]Károli Gáspár Református Egyetem

A magyar nyelvjárások többségének magánhangzó-rendszerében kétféle *e* fonémát találunk: egy nyíltabbat, amelyet a magyar dialektológiai hagyományban *e*-vel és egy zártabbat, amelyet *ë*-vel jelölünk. A magyar helyesírásban nincs meg ez a megkülönböztetés. Imre Samu A magyar nyelvjárások rendszere ([1]) című monográfiájában részletes kimutatásokat, térképeket készített az egyes magánhangzók hangszínrealizációiról A magyar nyelvjárások atlasza (a továbbiakban MNyA.) adatai alapján. A zárt *ë* esetében azonban kimutatások helyett az alábbi megállapításokat tette, szembe állítva az atlasz jellemzően homogén hangjelölési gyakorlatát a saját megfigyeléseivel: „Nem mutatnak számottevő különbséget a hangszín szempontjából az *ë* fonéma realizációi sem. Mindazokon a területeken, ahol önálló hangeszköz, elsöprő többséggel az *ë* jel a fővariáns. A Dunántúlon egészen ritka változatként fel-felbukkan az *ë̗,* az északi részeken némileg gyakoribb az *e̗*. Szubjektív benyomásaim alapján azonban úgy vélem, hogy e két területen az *ë* fonéma helyi általános realizációja között ennél némileg nagyobb a különbség. [...] Mindenesetre célszerű lenne ezt a kérdést megfelelő gépi mérésekkel tisztázni ([1]: 272.)”. Jelen előadásunkban a zárt *ë* ejtésének térbeli variabilitását, valamit az *e* és a zárt *ë* egymáshoz képesti realizációját vizsgáljuk akusztikai fonetikai eszközökkel, külön kitérve a fonémamegkülönböztetés földrajzi határán található településekre.

Kutatásunkat a MNyA. hangfelvételei alapján végezzük. A hangfelvételek 1960 és 1964 között készültek, összesen 352 kutatóponton. A kutatás alapját a Magyar nyelvjárási hangoskönyvekben (MNyHK. I–XVI.) megjelent, hanggal összekapcsolt hangfelvételek adják, a korpuszt saját lejegyzéseinkkel bővítjük, hogy minél teljesebb képet tudjunk adni a hangfelvételek kutatóponthálózata által lefedett területről. A teljes nyelvterületre érvényes kutatást nem végezhetünk, mivel a MNyA. ellenőrző gyűjtése során a határon túli területek közül kizárólag Csehszlovákiában készültek interjúk.

A kutatás módszertani hátterét a Bihalbocs nyelvészeti technológiái biztosítják ([2]). A magánhangzó-minőségek vizsgálatához formánsmérést végzünk a Praat Burg-algoritmusa használatával, az F1 és F2 értékeket és a magánhangzó időtartamát az interjúk hanggal összekapcsolt lejegyzéseiben az egyes magánhangzó-realizációkhoz kötötten rögzítjük. Méréseinket minden esetben ellenőrizzük. A kutatás során minden kiválasztott kutatóponton egy adatközlő anyagát dolgozzuk föl, és végzünk méréseket. Méréseink alapján – az interjúk átlagos időtartamából adódóan – elsősorban a gyakrabban ejtett magánhangzók minősége vizsgálható. Az egyes beszélők magánhangzóit az akusztikai térben ábrázoljuk, láthatóvá téve a magánhangzók ejtése közti összefüggéseket. A különböző beszélők ejtésmódjának összevethetősége érdekében a nyers Hz-értékeket normalizáljuk, illetve átskálázzuk a hosszú *í* átlagos F1 és F2 értékéhez, vagyis az adott beszélő F1-minimumához és F2-maximumához viszonyítva ([3][4][5]).

Mivel a magyar dialektológiában szokásos lejegyzői gyakorlatból kiindulva a magánhangzók ejtésében megmutatkozó különbségeket leginkább egy zárt–nyílt tengely mentén tudjuk megragadni, elemzésünkben elsősorban az F1 értékekre koncentrálunk. A dialektológiai kutatások gyakorlatához igazodva, összhangban az F1 és az artikulációs nyíltság közti erős korrelációval, az *e* és *ë* hangoknál mért F1 értékekből következtetünk a magánhangzók minőségének különbségére. A szakirodalommal való kompatibilitás fenntartása érdekében részben artikulációs terminológiát használunk akár percepción alapuló fonetikus lejegyzésből, akár akusztikai értékekből következő minőségekről van szó. A zárt *ë* 58 kutatóponton mért F1 értékeit térképezve azt látjuk, hogy a Dunántúl jelentős részén mértünk magasabb értékeket a keletebbre fekvő kutatópontokhoz viszonyítva. Megfigyelhető továbbá, hogy a Palócföld keleti szélén található négy kutatópont esetében szintén magasabb F1 értékeket mértünk.

A térképes kimutatást kiegészítve az egyes beszélőkhöz tartozó értékeknek az akusztikai térben való ábrázolásával a következőket mondhatjuk. A dunántúli beszélők ejtésmódjára jellemző, hogy azoknál, akiknél az *e* F1 értéke magasabb, vagyis nyíltabb az *e* realizációja, gyakran mértünk magasabb F1 értékeket a zárt *ë* esetében is. A palóc nyelvjárású beszélők esetében mindkét hang, az *e* és a zárt *ë* realizációja is zártabb, kisebb köztük az euklideszi távolság.

Külön vizsgáltuk azokat a kutatópontokat – a Palócföld keleti szélén –, ahol a zárt ë realizációja nyíltabbnak tűnik, mint a nyugatabbra található palóc nyelvjárású településeken. Külön megnéztük ezeken a kutatópontokon a MNyA. lejegyzéseit is, ahol azt találjuk, hogy a legkeletibb vizsgált település, Baktakék esetében az atlaszadatokban is tükröződik, hogy nem egyértelmű a nyíltabb *e* és a zárt *ë* közti megkülönböztetés. Ezeken a kutatópontokon a két fonéma összeolvadását, illetve ennek utolsó fázisát valószínűsíthetjük.

Eredményeink arra engednek következtetni, hogy Imre Samu „szubjektív benyomásai" helyesek voltak, a zárt *ë* fonéma realizációiban jellegzetes térbeli különbségeket találunk. A MNyA. e magánhangzó vonatkozásában szinte teljesen egységes hangjelölése ellentétben áll a jelen kutatás eredményeivel, így kimondhatjuk, hogy a MNyA. lejegyzése téves képet ad a zárt *ë* ejtéséről. Míg az egyetlen *e* fonémát ismerő magyar nyelvváltozatok *e*-jétől a Dunától keletre, illetve a nyelvterület északi részén elhelyezkedő nyelvjárásokban a zárt *ë* valóban zártabban (a viszonyítási *é* minőségénél csak egy árnyalattal nyíltabban) realizálódik, addig a Dunántúl jó részén ez nincs így. Ezeken a kutatópontokon a zárt *ë* minősége jellegzetesen nyíltabb, hasonló az egyetlen *e* fonémát ismerő magyar nyelvváltozatok *e*-jéhez.

[1] Imre Samu 1971. *A mai magyar nyelvjárások rendszere*. Budapest: Akadémiai Kiadó.

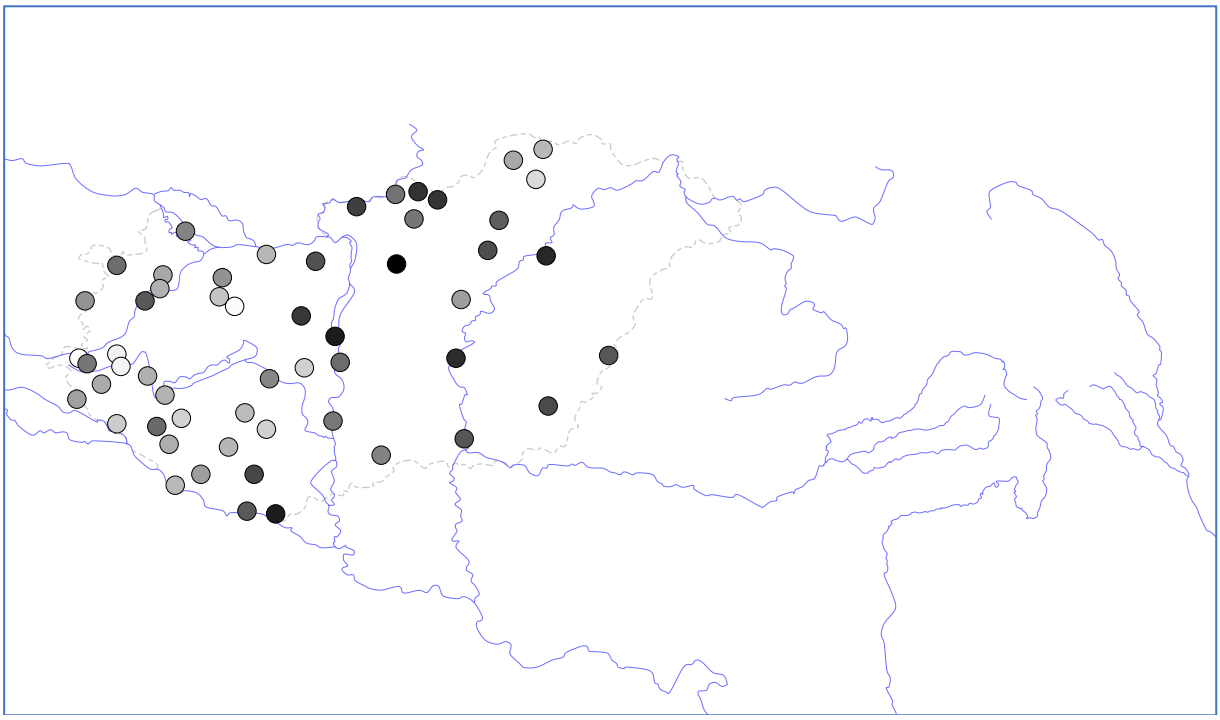
[2] Vékás Domokos 2007. Számítógépes dialektológia. In Guttmann Miklós – Molnár Zoltán (szerk.): *V. Dialektológiai Szimpozion. Szombathely, 2007. augusztus 22–24.* Szombathely: Berzsenyi Dániel Főiskola. 289–293.

[3] Kocsis Zsuzsanna – Szabó Gergely – Vargha Fruzsina Sára 2024. Dialectometric and acoustic research on Hungarian vowel pronunciation. In: Nicolae Saramandu et al. (eds.) *Proceedings of the Xth Congress of the International Society for Dialectology and Geolinguistics.* Alessandria: Edizzione dell'Orso. 377–386.

[4] Vargha Fruzsina Sára – Kocsis Zsuzsanna – Szabó Gergely 2025. Magyar nyelvjárási magánhangzók az akusztikai térben. In: Bodó Csanád – Szabó Gergely (szerk.): *A*

szociolingvisztikai kutatás társadalmi hatásai. Budapest: ELTE Eötvös Kiadó. 325–339.

[5] Kocsis Zsuzsanna (Benyújtva) A nyelvjárási *ö* és *ë* közötti átmeneti hangok akusztikai vizsgálata. *Magyar Nyelv.*



A zárt *ë* normalizált F1 értékei a MNyA. 58 kutatópontján. Minél magasabb az F1 értéke, annál világosabb egy adott kutatópont, minél alacsonyabb, annál sötétebb. (Maximum érték: 648 Hz Rábagyarmat; minimum érték: 417 Hz Hévízgyörk)

Methodological and Technical Challenges in Research Involving EMA on Typical and Disordered Speech

Anna Wojnarowska-Borowiec[1] and Hanna W. Kasperek[2]

[1] University of Warsaw, Poland
[2] Adam Mickiewicz Univeristy, Poland

This contribution outlines the methodological framework and practical challenges encountered during a series of data collection sessions using electromagnetic articulography (*EMA; Carstens AG501*) within an on-going research project investigating articulatory and functional parameters in both typical speakers and individuals with speech sound disorders (SSDs). The study combines EMA with acoustic analysis and aims not only to collect high-quality phonetic data, but also to develop a standardized and replicable protocol for EMA-based research in both clinical and experimental settings. Methodological decisions and technical procedures are presented in light of recent literature on EMA (e.g., [1, 2, 3, 4, 5]).

The presentation will introduce the detailed protocol for EMA sessions, including:

- the development of comprehensive metadata questionnaires covering anatomical, developmental, and phonetic information;
- the design of linguistic materials for elicited speech and tasks aimed at spontaneous speech production;
- the preparation of informed consent forms and participant documentation in accordance with Bioethics Committee standards;
- a standardized list of photographic and video documentation (e.g., sensor placement, body posture, oral cavity images);
- precise wiring and sensor attachment schemes;
- disinfection procedures for tools and workstations;
- instructions for sensor mounting on intraoral and extraoral structures;
- and a standardized calibration protocol using a biteplate and head correction to minimize motion artifacts.

A structured session flow was also developed, comprising: signal testing, recording of standard stimuli (e.g., isolated sounds, syllables); elicited utterances (scenarios targeting realizations of specific words within phrase containers and selected phrase-level phenomena, particularly articulatory correlates of speech rhythm); spontaneous speech, control recordings; and final documentation. Each recording session was conducted by a team including an engineer, a phonetic-clinical specialist and supporting personnel, to ensure both high-quality data and participant comfort.

The presentation will also address the most common technical challenges encountered during the recordings—such as sensor displacement, movement artifacts, and signal interference—and will describe the team's practical solutions aligned with international methodological standards.

Importantly, this contribution highlights the need for greater transparency and standardization in EMA-based research. Challenges related to participant diversity, equipment precision, and data interpretation emphasize the importance of reporting practical details that are often absent from official publications. The integration of multimodal data sources and adherence to reproducible protocols are essential for the development of articulatory databases for typical speech as well as for supporting clinical and cross-linguistic applications of EMA in the future.

While most published studies tend to be concise and outcome-oriented, this presentation aims to shed light on the **process** of resource creation by presenting the challenges and insights encountered during actual data collection sessions — including organizational, technical, and participant-related aspects — which are often omitted from such publications. By sharing these insights, the project contributes to broader discussions on best practices in experimental phonetics and clinical speech research.

References

[1] Fuentes, R. et al. (2017). 'Systematic standardized and individualized assessment of masticatory cycles using electromagnetic 3D articulography and computer scripts'. In: *BioMed Research International* 2017.1, p. 7134389.

[2] Lorenc, A. et al. (2023). 'Articulatory and acoustic variation in Polish palatalised retroflexes compared with plain ones'. In: *Journal of Phonetics* 96, p. 101181.

[3] Mik, Ł. et al. (2018). 'Fusing the electromagnetic articulograph, high-speed video cameras and a 16-channel microphone array for speech analysis'. In: *Bulletin of the Polish Academy of Sciences. Technical Sciences* 66.3.

[4] Rebernik, T. et al. (2021). 'A review of data collection practices using electromagnetic articulography'. In: *Laboratory Phonology* 12.1, p. 6.

[5] Wang, Y. K. et al. (2013). 'Model-based identification of motion sensor placement for tracking retraction and elongation of the tongue'. In: *Biomechanics and modeling in mechanobiology* 12.2, pp. 383–399.