Speech Research conference

Hungarian Research Centre for Linguistics

Budapest, 23–24. February 2023

Beszédkutatás konferencia

Nyelvtudományi Kutatóközpont

Budapest, 2023. február 23–24.

Organisers/Szervezők:

Gráczi, Tekla Etelka
Horváth, Viktória
Juhász, Kornélia
Kohári, Anna
Krepsz, Valéria
Mády, Katalin

DOI: 10.18135/BeszKutKonf.2023

Scientific committee/Tudományos bizottság

Asztalos, Anikó
Baditzné Pálvölgyi, Kata
Baló, András Márton
Balogné Bérces, Katalin
Belz, Malte
Beňuš, Štefan
Bóna, Judit
Bučar Shigemori, Lia Saki
Csapó, Tamás Gábor
Deme, Andrea
Fejes, László
G. Kiss, Zoltan
Gaál, Zoltán Kristóf
Gocsál, Ákos
Gyarmathy, Dorottya
Hunyadi, László
Huszár, Anna
Jankovics, Julianna
Jordanidisz, Ágnes
Káldi, Tamás
Kallio, Heini
Kis, Orsolya
Liker, Marko
Maitz, Péter
Markó, Alexandra
Mihajlik, Péter
Mus, Nikolett
Németh, Zsuzsanna
Németh T., Enikő
Neuberger, Tilda
Rebrus, Péter
Salminen, Tapani
Sass, Bálint
Szalontai, Ádám
Szekrényes, István
Sztahó, Dávid
Trouvain, Jürgen
Varga, Marianna
Varga, Zoltán
Vicsi, Klára
White, Laurence
Zainkó, Csaba

Contents

Tartalomjegyzék

Invited talks/Meghívott előadások

Mondatfajták és beszédaktusok a beszélt nyelvben: Esettanulmány egy székely nyelvi jelenségről

Gyuris Beáta

Nyelvtudományi Kutatóközpont, Budapest

Eötvös Loránd Tudományegyetem, Budapest

Az előadás célja, hogy bemutassa és modellezze a székely dialektusban használt *ejsze* pragmatikai jelölő ([5]) formai és jelentéstani tulajdonságait, [4] alapján.

Elsőként amellett érvelek, hogy az *ejsze* nem járul hozzá a kijelentő mondatok igazságfeltételeihez, csak a használati feltételeikhez (*use conditions,* [2]), amit például az (1a) és (1b) elfogadhatósága közötti kontraszt mutat az alábbi kontextusban:

(1) Egy játék keretében *A* golyót húz egy olyan dobozból, amelyben egy fekete és kilenc fehér golyó van. Tudja, hogy melyik színből hány golyó van a dobozban, már kezében a kihúzott golyó, de még nem nézte meg a színét. Ezt mondja magában:
   a. # *Ejsze* egy fehéret húztam ki.
   b. Lehetséges, hogy egy fehéret húztam ki.

Az alábbi példák azt sugallják, hogy az *ejsze* egyaránt jelölheti azt, hogy a beszélő a kijelentő mondat által jelölt *p* propozíciót lehetségesnek ((2)–(3)), nagyon valószínűnek ((4)), nyilvánvalónak ((4)–(5)), vagy köztudottnak ((6)) tartja:

(2) Ejsze mégsem nyitják meg a bányát Verespatakon. Korodi mai kolozsvári sajtótájékoztatóján fejezte ki kételyeit azzal kapcsolatban, hogy a Rosia Montana Gold Corporation (RMGC) megkaphatná a szükséges engedélyeket a verespataki ciántechnológiás aranybányászáshoz.[1]

(3) A nővér kikíséri a beteget a rendelőből egy kisebb beavatkozás után.
Nővér: Szüksége van további segítségre?
Beteg: *Ejsze* megleszek!

(4) – De szép gombok! *Ejsze* arany! – koppantotta meg az egyiket, éppen az ásós vitéz szíve felett. (Magyar Nemzeti Szövegtár, [9])

(5) *A* meglát egy új jeggyűrűt *B* ujján. Ezt mondja neki:
Ejsze eljegyeztek! (Bende-Farkas Ágnes, p.c.)

[1] https://foter.ro/cikk/20140515_*ejsze_megsem_nyitjak_meg_a_banyat_verespatakon/

(6) [A részeg ember puskával akar patkányt irtani.]
Ejsze olyan részeg vagy, hogy célozni se tudsz. ([6])

A továbbiakban azonban amellett fogok érvelni, hogy lehetséges egy egységes jelentést rendelni ehhez a pragmatikai jelölőhöz, ahelyett, hogy többjelentésű kifejezésnek tekintenénk. A javasolt interpretáció szerint (amely [1]-nek a német *wohl*-ra vonatkozó javaslatát veszi át) az *ejsze* a fenti mondatokban *inferenciális evidenciális* kifejezésnek tekintendő. [1] nyomán azt mondjuk, hogy az ilyen kifejezés azt jelzi, hogy a beszélő információja a mondat jelöletének igazságáról a saját *felülbírálható* (visszavonható, nem-monoton, *defeasible*) *következtetésének* eredménye, amely a rendelkezésére álló információkonn, köztük a kontextusbeli evidenciákon alapul. Az előadás azt is bemutatja, hogy a fenti használati feltételeket hogyan lehet formálisan megragadni.

Az előadás második része az *ejszé*-nek a kérdő mondatokkal ill. kérdő beszédaktusokkal való kompatibilitását vizsgálja. (7)-(10) alapján megmutatjuk, hogy az *ejsze* sem eldöntendő kérdő főmondatokban, sem alárendelő tagmondatokban, sem többszörös emelkedő-eső (/\) intonációjú kijelentő mondatokban ([3]) nem fordulhat elő:

(7) *A* a barátját kérdezi: (**Ejsze*) félsz-e a vizsgától?

(8) *A* a barátját kérdezi: (**Ejsze*) félsz a vizsgától /\ ?

(9) Mari azt kérdezte Palitól, hogy (**ejsze*) fél-e (**ejsze*) a vizsgától (**ejsze*).

(10) *A* a barátját kérdezi: (**Ejsze*) félsz /\ a vizsgától /\ ?

A fenti példák egyértelmű agrammatikussága ellenére egyes beszélők elfogadhatónak tartják a (11)-(12)-hez hasonló függő eldöntendő kérdő tagmondatokat. Az előadásban lehetséges magyarázatokat vázolunk fel az ellentét feloldására.

(11) Mari azt kérdezte Palitól, hogy (ejsze) fél (ejsze) a vizsgától (ejsze).

(12) Ejsze félsz a vizsgától? – kérdezte Mari Palitól.

Végül megvizsgáljuk az *ejsze* szerepét a (13)-(14)-hez hasonló kiegészítendő kérdő mondatokban. Javaslatunk szerint ezekben a példákban *illokúciós evidenciális* kifejezésként (l. [7]) viselkedik.

(13) Ejsze kinél vótál tegnap? ([10])

(14) Ejsze, ezt meg honnan a fenéből tudod? (Magyar Nemzeti Szövegtár, [8]), idézi [10].)

Hivatkozások

[1] Eckardt, R. (2020). „Conjectural questions: The case of German verb-final *wohl* questions". *Semantics and Pragmatics* 13, 1–54. old.

[2] Gutzmann, D. (2015). *Use-conditional meaning. Studies in multidimensional semantics*. Oxford: Oxford University Press.

[3] Gyuris, B. (2019). „Thoughts on the semantics and pragmatics of rising declaratives in English and rise-fall declaratives in Hungarian". *K + K = 120. Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*. Szerk. Gyuris, B., K. Mády & G. Recski. MTA Nyelvtudományi Intézet, 247–280. old.

[4] Gyuris, B. (megj. előtt). „The semantics of *ejsze* in the Székely dialect of Hungarian". *Journal of Uralic Linguistics*.

[5] Gärtner, H.-M. & B. Gyuris (2012). „Pragmatic markers in Hungarian: Some introductory remarks". *Acta Linguistica Hungarica* 59(4), 387–426. old.

[6] Ittzés, N., szerk. (2016). *A magyar nyelv nagyszótára VI.* Budapest: MTA Nyelvtudományi Intézet.

[7] Murray, S. E. (2017). *The semantics of evidentials*. Oxford: Oxford University Press.

[8] Oravecz, C., T. Váradi & B. Sass (2014). „The Hungarian gigaword corpus".

[9] Oravecz, C., T. Váradi & B. Sass (2014). „The Hungarian Gigaword Corpus". *LREC 2014 - Ninth International Conference on Language Resources and Evalutation*. Szerk. Calzolari, N. és tsai. Lisbon: European Language Resources Association (ELRA), 1719–1723. old.

[10] Tuzson, E. (2022). *Az ejsze szó használata Csíkszenttamáson*. MA-szakdolgozat, Babeș-Bolyai Tudományegyetem, Kolozsvár.

Global kinematic patterns of tongue movement in typical and atypical speech

Marko Liker
University of Zagreb, Croatia
mliker@ffzg.hr

When studying speech, it is important to keep in mind that the primary speech organ, the tongue, is a muscular hydrostat [11]. Muscular hydrostats, such as the octopus, are not frequently found in nature. Their main characteristic is that they do not have a skeleton or other rigid structure. Instead they are mainly composed of muscular tissue. Their movements are therefore rather complex to control and this complexity must be taken into account when studying tongue movements during speech. At least two important consequences of this complexity should be considered: 1. the tongue moves in such a way that deforms its whole midsagittal cross section, rather than individual points along its surface [7], and 2. the tongue has a potentially unlimited number of patterns in which it can move, because it is not restricted by the typical six degrees of freedom in movements of rigid objects [15]. This makes tongue movements very hard to investigate and very challenging to treat when they are impaired. It would, therefore, be very useful to identify a set of basic global kinematic patterns of tongue movements during speech, which would be relatively easy to identify and quantify, and then in turn to investigate the relevance of those patterns for the quantification of other speech phenomena, such as coarticulation and speech intelligibility. Previous investigations have shown that there are probably four such basic kinematic patters of tongue movement during speech: pivoting, arching, shifting and bracing (locking) [4, 6, 7, 8, 10, 14, 16]. These patterns are currently largely under-investigated and most of them have yet to be confirmed in different languages. The focus here will be on tongue pivoting and tongue bracing gestures with evidence from Croatian.

Tongue pivoting is the pattern during which the tongue is anchored at one point along the midsagittal cross section, while posterior and anterior to the anchor point the tongue moves in opposite directions. It occurs when gestures of two segments produce maximal movements in two constriction areas and minimal movement between them [8]. It is hypothesized that its purpose is to make the acoustic signal articulatorily transparent [7]. It is currently unclear whether these patterns occur in all languages and how they relate to coarticulation and speech intelligibility.
Tongue bracing can be described as active mechanical support of the tongue during speech, which is achieved when the tongue comes into contact with a rigid surface of the vocal tract [4]. Although initial research of this phenomenon focused on lateral bracing in specific sounds [2, 3, 9, 10, 12], subsequent investigations showed that the tongue is actively and continuously braced during speech [1, 4]. These studies also showed that lateral bracing was the default bracing pattern and that central bracing would take over when lateral bracing is blocked [4, 14]. The results from these studies support the claim that tongue bracing is one of the global kinematic patterns of tongue movement during speech. It is currently unknown what happens when typical bracing patterns are interrupted and how these pervasive and continuous bracing patterns relate to coarticulation and speech intelligibility.

From the short review presented above it is clear that there are at least three relevant research problems connected with tongue pivoting and tongue bracing: 1. How widespread and how variable are these phenomena (in terms of different languages and in terms of different speech sounds), 2. What happens when they are interrupted (in typical vs. atypical speech)?, and 3. How are these phenomena connected with coarticulation and speech intelligibility (in typical and atypical speech)?

I shall address these research questions by presenting ultrasonic tongue imaging (UTI) and electropalatographic (EPG) data from Croatian typical adult speakers and their atypical counterparts who are prelingually deaf and use cochlear implants (CI). Speech material is obtained from the CROCO corpus containing acoustic, EPG and UTI recordings of 10 typical speakers and seven CI users [13], all recorded in a communicative situation which facilitates natural speech in a dialogue (i.e. a variation of the map task). The resulting quasi-spontaneous speech provides an opportunity to analyze speech patterns in a more natural communicative situation than reading sentence lists. The rationale for including CI users, alongside their typically speaking counterparts in this investigation is threefold: 1. studying clinical population can reveal aspects of speech production which could have relevant implications for speech production theory, 2. it can help us understand what happens when certain speech processes become interrupted in the continuum between typical and atypical speech, and 3. it can advance our approach to clinical treatment of atypical speech [5].

UTI data will be used to show how pivot patterns can be captured in terms of pivot anchor and pivot strength. The analysis will show that pivot patterns can occur not only in transitions between sounds, but also within a single speech sound. The comparison between typical and atypical speakers will reveal a possible connection between pivot patterns and speech intelligibility. EPG data will be used to show how the coordination between lateral and central bracing can be quantified. The data will show the difference in bracing pattern coordination between typical speakers and CI users. It will be shown that bracing coordination patterns differ in CI users with different clinical data (e.g. age at implantation, speech audiogram results) and with different levels of speech intelligibility scores. I shall also show how differences in the coordination of lateral and central bracing between typical and CI speakers can be related to differences in coarticulatory processes.

Several implications of these analyses will be discussed (e.g. could these patterns be used to differentiate between language-specific and biomechanically-universal aspects of speech?, could these patterns be useful in speech impairment diagnosis?, could these patterns be targeted with specifically designed speech exercises in order to improve overall intelligibility?).

References

[1] Bressmann, T., Flowers, H., Wong, W., Irish, J. C. (2010). ‘Coronal view ultrasound imaging of movement in different segments of the tongue during paced recital: Findings from four normal speakers and a speaker with partial glossectomy’. *Clinical Linguistics & Phonetics*, 24(8), pp. 589–601.

[2] Fuchs, S., Perrier, P., Geng, C., Mooshammer, C. (2006). 'What role does the palate play in speech motor control? Insights from tongue kinematics for German alveolar obstruents'. In: Harrington, J., Tabain, M. (eds.), *Speech production: Models, phonetics processes, and techniques.* New York: Psychology Press, pp 149-164.

[3] Gibbon, F. E., Yuen, I., Lee, A., & Adams, L. (2007). 'Normal adult speakers' tongue palate contact patterns for oral and nasal stops'. Advances in Speech-Language Pathology 9, 82–89.

[4] Gick, B., Allen, B., Roewer-Després, F., Stavness, I. (2017). 'Speaking tongues are actively braced'. *Journal of Speech, Language, and Hearing Research,* 60(3), pp. 494-506.

[5] Hardcastle, B., Tjaden, K. (2008). 'Coarticulation and speech impairment'. In: Ball, M.J., Perkins, M.R., Müller, N., Howard, S. (eds), *Handbook of Phonetic Sciences,* Blackwell Publishing Ltd., pp. 506-520.

[6] Honda, K., Takano, S., Takemoto, H. (2010). 'Effects of side cavities and tongue stabilization: Possible extensions of the quantal theory'. Journal of Phonetics 38, 33-43.

[7] Iskarous, K. (2005). 'Patterns of tongue movement'. *Journal of Phonetics,* 33(4), pp. 363-381.

[8] Kim, B., Tiede, M.K., Whalen, D.H. (2019). 'Evidence for Pivots in Tongue Movement for Diphthongs'. In: Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia, pp 3666-3670.

[9] Kochetov, A., Colantoni, L. (2011). 'Coronal place contrasts in Argentine and Cuban Spanish: An electropalatographic study'. Journal of the International Phonetic Association 41(3), pp. 313–342.

[10] Lee, A., Gibbon, F.E., Oebels, J. (2015). 'Lateral bracing of the tongue during the onset phase of alveolar stops: an EPG study'. *Clinical Linguistics and Phonetics,* 29(3), pp. 236-245.

[11] Levine W.S., Torcaso C.E., Stone M. (2005). 'Controlling the Shape of a Muscular Hydrostat: A Tongue or Tentacle*'. In: P. Dayawansa W., Lindquist A., Zhou Y. (eds): *New Directions and Applications in Control Theory.* Lect. Notes Control, vol 321. Springer, Berlin, Heidelberg, pp 205-222.

[12] Liker, M., Gibbon, F.E. (2015). 'Place of articulation of anterior nasal versus oral stops in Croatian'. *Journal of the International Phonetic Association,* 45 (1), 35-54.

[13] Liker, M., Vidović Zorić, A., Zharkova, N., Gibbon, F. (2019). 'Ultrasound analysis of postalveolar and palatal affricates in Croatian: a case of neutralisation', In: Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia,* pp. 3666-3670.

[14] Liu, Y., Tong, F., de Boer, G., Gick, B. (2022). 'Lateral tongue bracing as a universal postural basis for speech'. *Journal of the International Phonetic Association,* 1-16.

[15] Rosenbaum, D.A. (2010). *'Human Motor Control'.* Elsevier: Burlington, San Diego, London.

[16] Stone, M. (1991). 'Toward a model of three-dimensional tongue movement'. *Journal of Phonetics,* 19, pp. 309-320.

Results in automatic speaker verification for forensic purposes on the ForVoice120+ dataset

Dávid Sztahó
Budapest University of Technology and Economics, Budapest

Speaker identification (SI) and verification (SV) (together: speaker recognition) have been studied for a long time and have a still growing literature [1]. In speaker identification the task is to identify an unknown speaker from a set of already known speakers: find the speaker who sounds closest to the test sample. Closed-set (or in-set) scenario is when all speakers within a given set are known. On the other hand, open-set (or out-of-set) speaker identification is when the set of known speakers may not contain the potential test subject. In speaker verification, we verify if a speaker is who he or she claims to be by comparing two speech samples/utterances and evaluating if the speakers of the two samples match. This is traditionally done - in general forensic voice comparison practice - by comparing the test sample to a sample or samples of the given speaker and a universal background model [2]. Another way of comparing if a speaker pair is of same origin is to score the pairs as 'same' or 'different' and create a model accordingly. It follows from this definition that technically, forensic speaker comparison belongs to the speaker verification scheme, although the 'known' speaker is not specifically known in this case. A voice sample may be associated with an assumed speaker. The goal is to verify if this speaker's identity (suspect) matches another unknown speaker's identity (offender). Often, the goal of the forensic voice comparison is also to verify if two unknown speaker's identity is the same. In practice, this verification is done with the same method.

There is a paradigm shift in forensic sciences and practice [3]–[5] that makes the automatic and semi-automatic evaluation of evidence available using various modalities and kinds of measurements (such as DNA, fingerprint). This new paradigm, called likelihood-ratio (LR) framework, supports a processing pipeline easily computable across multiple evidence types. Forensic voice comparison, where speaker recognition techniques are adapted to the requirements of the framework, is such a field, where this new paradigm can be adapted. It produces a probability ratio of same and different speakers for an evidence.

There are traditional (non-automatic) methods for forensic voice comparisons, such as perceptive analysis, during which, as the name suggests, an expert reaches a final decision after listening to the recordings. In spectrographic comparison, the expert observes similarities and differences in the spectrogram of the recordings. This method has received many criticisms and is no longer used today. In the acoustic-phonetic approach, the expert mainly analyzes the characteristics observed during the previous two methods, but now they are quantified. Such characteristics can be the pitch (fundamental frequency) and the formants describing the vocal tract. The aim is to compare the variability of the measured acoustic characteristics and make a decision by using statistical means.

A common characteristic of automatic speaker identification and verification techniques is that they require as little user intervention as possible. In addition to specifying the audio recordings, only a few parameters of the system must/can be set, the rest are done automatically. During the development of the techniques, several procedures were developed. Of these, speaker embeddings based on deep learning are considered the most effective at the present time. Automatic methods started with applying GMM [2], GMM-UBM [6] approaches, in which acoustic features were modelled by combination of multiple Gauss distributions and

applying model for that represented all speakers other than the target speaker (called universal background models). Then, i-vectors [7] became the state-of-the-art, extracting a feature vector representing a speaker (through factor analysis of the superimposed parameters of a trained GMM (called GMM supervector). Today's state-of-the-art speaker verification and identification technologies are mainly based on deep learning methods made available by the appearance of huge speech corpora. They contain novel classification and feature extraction methods, such as d-vectors [8], j-vectors [9], x-vectors [10] and ECAPA-TDNN networks [11].

In order to be able to perform experiments within the framework of LR, a database is required that meets the basic assumptions of the new paradigm [12]: i) samples that differ in time from each speaker must be recorded (similarity modeling), ii) it must include many speakers, preferably representative of the population (to model typicality), iii) voice samples recorded in different ways must be recorded (to compensate for so-called channel mismatch, e.g. telephone or studio quality), iv) different speech types from one speaker must be recorded due to differences in speaking style to compensate for differences appearing even within the speaker (speech style mismatch compensation). In accordance with all these challenges, we designed the ForVoice120+ database, which is unique among currently available domestic databases. The work was funded by project no. FK128615, which has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FK_18 funding scheme.

The developed ForVoice120+ database is structured according to the following objectives. Our goal is to create a database that (i) fits the criteria formulated in the new paradigm shift introduced in forensics, so that the analyzes carried out on it can be important cornerstones for the development and evaluation of forensic voice comparison systems, for establishing new, individual acoustic-phonetic characteristics, (ii) annotated, thus providing the possibility for professionals to carry out acoustic, phonetic, linguistic and other speech technology research, with particular regard to the characteristics of the individual speech of the speaker, furthermore iii) to have a basic database on which new research directions can be implemented in phonetics and speech technology and provide a basis for answering research questions that could not be studied in the Hungarian language before (e.g. systematic analysis of variance within and between speakers over a time, etc.).

Most of the pretrained deep speaker embeddings are trained on English corpora because it is easily accessible. Thus, language dependency is a key factor in automatic forensic voice comparison, especially when the target language is linguistically quite different. There are numerous commercial systems available, but their models are trained on a different language (mostly English) than the target language. In the case of a low-resource language, developing a corpus for forensic purposes containing enough speakers to train deep learning models is costly. We found that the model pre-trained on a different language but on a corpus with a vast number of speakers can be used on samples with language mismatch. The best equal error rate with speaker enrollment was 1%. The effect of sample durations and speaking styles were also examined. It was found that the longer the duration of the sample in question the better the performance is. Also, if multiple recordings are present from the suspect, there is no real difference if various speaking styles are applied.

Emotional conditions play a significant role in forensic voice comparison and speaker verification systems. When emotion is present in speech, the verification's performance will deteriorate. We found that both applied deep learning models (x-vector and ECAPA-TDNN) are sensitive to the emotional content of the speech samples. They performed worse when the

trials were composed of different emotional samples. A 2% decrease in EER was found if different emotions were present compared to the case when the same emotions were applied. If the emotional content on the recording is not neutral, but it is the same in both recordings of a trial, the performance deterioration is not present.

For the rest of the ForVoice120+ project, we will conduct speaker verification using twins' voices to see how they deteriorate the performance of such systems. Also, a book with proposed title 'Demonstration of deep learning based automatic forensic voice comparison on the ForVoice120+ speech database' will be completed to summarize our results and technology background for speech technology and forensic experts.

References

[1] R. M. Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Computers & Electrical Engineering*, vol. 90, p. 107005, 2021, doi: 10.1016/j.compeleceng.2021.107005.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995, doi: 10.1109/89.365379.

[3] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice*, vol. 51, no. 3, pp. 91–98, 2011, doi: 10.1016/j.scijus.2011.03.002.

[4] G. S. Morrison, "Forensic voice comparison and the paradigm shift," *Science & Justice*, vol. 49, no. 4, pp. 298–308, 2009, doi: 10.1016/j.scijus.2009.09.002.

[5] M. J. Saks and J. J. Koehler, "The coming paradigm shift in forensic identification science," *Science*, vol. 309, no. 5736, pp. 892–895, 2005, doi: 10.1126/science.1111565.

[6] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, Sep. 1997, pp. 963–966. doi: 10.21437/Eurospeech.1997-337.

[7] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Tenth Annual conference of the international speech communication association*, 2009.

[8] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4052–4056. doi: 10.1109/ICASSP.2014.6854363.

[9] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in *Sixteenth annual conference of the international speech communication association*, 2015.

[10] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification.,” in *Interspeech*, 2017, vol. 2017, pp. 999–1003.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” in *Interspeech 2020*, Oct. 2020, pp. 3830–3834. doi: 10.21437/Interspeech.2020-2650.

[12] G. S. Morrison, P. Rose, and C. Zhang, “Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice,” *Australian Journal of Forensic Sciences*, vol. 44, no. 2, pp. 155–167, 2012, doi: 10.1080/00450618.2011.630412.

Conference presentations/Konferenciaelőadások

Phonological transfer of voicing in multilingual speech

Zsuzsanna Bárkányi[1], and Zoltán G. Kiss[2]
[1] The Open University; Hungarian Research Centre for Linguistics
[2] Eötvös Loránd University, Budapest

The aim of the present paper is to contribute to the growing body of research on how previously acquired languages impact on the acquisition of laryngeal features, more specifically the phonetics and phonology of voicing in third or subsequent languages (L3). Such results might inform theories of phonological learning.

Past decades have seen an increase in research on the acquisition of third and successive languages mostly by adult and heritage speakers. However, most of these studies focus on morphosyntactic phenomena, and do not agree on which theoretical account explains best the attested transfer or the lack of it, i.e. whether it is the native language that has a privileged role and thus serves as a source of transfer (e.g. [5]); or rather L2 that is acquired later – often in adulthood – and as such acquisition is cognitively more similar to L3 ([1]); or it is the typologically more similar language ([9]), and due to cognitive economy when sufficient similarity is found, it is transferred wholesale. It could also be the case that both (all) previously acquired languages are available for transfer. According to [2], all previously acquired languages are available for transfer but only in case it is facilitative. On the other hand, [11] and [10] advocate for both positive and negative transfer. They also claim that transfer occurs property-by-property rather than wholesale depending on which aspects of L1 or L2 are perceived to be more similar and can be both positive and negative. The modeling of similarity, complexity, frequency, etc. is not straightforward, but by allowing a wide array of optionality, the models fail to predict much and cannot be falsified.

Researchers also advocate that it is more beneficial to study the initial stages of language acquisition since as learners progress, they might learn the relevant grammatical structures and simply make fewer errors ([8]). This, however, might not be the case in the acquisition of the phonetic/phonological system of a foreign language. Fossilisation occurs earlier and ultimate attainment is generally less native-like in this domain, the belief that native-like pronunciation when acquisition occurs after puberty is common (e.g. [13]). Also, data on the nature of the driving forces behind phonetic and phonological transfer are less abundant, and similarly to other linguistic levels, no prevalent model has been identified, neither is the link between perception and production clearly defined. Conducting research in the field is not an easy task since a multitude of language external variables should not only be controlled for but also defined and measured including age, length of study, context of study (e.g. immersion, heritage language), type of instruction (or lack of it), metalinguistic awareness, etc.

The specific research questions we aim to explore in the present study are the following:

- What is the role of previous linguistic experience in the acquisition of L3 voicing features?
- Which speech learning model(s) can account for the observed cross-linguistic influences?

As a first step, we carried out a systematic literature review of the advances and state of research on phonetic/phonological transfer in L3/L*n* acquisition in the past five years. Peer-reviewed publications that appeared in a journal, book or dissertation were included. Out of 39 such studies, 16 deal with the acquisition of laryngeal features and they fall into three thematic groups: (i) the implementation of voicing contrast in stops; (ii) voice stop lenition in Spanish and (iii) word-final devoicing (or contrast preservation) in obstruents.

The vast majority of studies that focus on the phonetic realization of stops, usually by measuring Voice Onset Time in stops. This is understandable as VOT production received much attention in L3 literature earlier (e.g. [13]) and in the study of L2 speech as well. This is probably because non-target-like VOT realisations have been found to correlate with the degree of foreign accent. While these studies tested different groups of trilingual speakers (e.g. heritage speakers, beginner L3 learners, advanced L3 learners) and employed different methodologies (reading, picture naming; monolingual sessions, bilingual sessions, etc.) to find out whether learners created separate phonetic categories in their languages, very few studies actually have attempted to account for the results within a specific theoretical model. Most of these studies are production studies, their results are inconsistent and are hard to compare (mostly due to differences in the methodology used). Those few studies that examine perception as well or link production to perception (e.g. [6]) claim that there is no clear link between production and perception, for instance, L3 learners could perceive the new acoustic feature voicing lead, but found it difficult to produce L3 voiced stops, and voiced and voiceless stops do not behave in the same way. The predictions of Flege's [4] Speech Learning Model also seem to be born out in L3 phonology, i.e. that different but very similar L1 (and/or L2) categories are difficult to acquire in production and can lead to confusion in perception.

The exploration of allophonic alternations – like voiced stops in L3 Spanish – (e.g. [7]) presents evidence in support of the Typological Primacy Model (Rothman 2011). Studies on "feature changing" phonological processes like word-final devoicing (e.g. [12]). The authors claim that final obstruent devoicing is hard to overcome, however, the role of L1 and L2 is not clear, neither is the link between production and perception ([3]). These studies didn't find a uniform transfer of phonological features and advocate for property-by-property transfer. While we do agree that transfer is not uniform and there is an intricate interaction between different language-specific inventories, phonological processes and prosodic hierarchies, as well as production and perception, the question is whether every single phonological process is leaned in a different way or we can make meaningful generalisations if we focus on a well-defined set of phenomena like post-lexical laryngeal processes that do not create novel segments. Good candidates for such endeavor are regressive voicing assimilation, pre-sonorant voicing and word-final devoicing.

In this presentation methodological suggestions will be made on how to tease apart the predictions of some of the current speech learning models in relation to dynamic voicing processes.

References

[1] Bardel C. & Sánchez L. (2017). 'The L2 status factor hypothesis revisited: The role of metalinguistic knowledge, working memory, attention and noticing in third language learning'. In Angelovska T. & Hahn A. (eds) L*3 syntactic transfer: Models, new developments and implications*. Amsterdam: John Benjamins, pp. 85–102.

[2] Berkes É. & Flynn S. (2012). 'Multilingualism: New perspectives on syntactic development.' In: Ritchie WC and Bhatia TK (eds) *The handbook of bilingualism and multilingualism*. 2nd edition. Chichester: John Wiley and Sons, pp. 137–67.

[3] Domene C. (2021). *Beyond transfer? The acquisition of an L3 phonology by Turkish– German bilinguals*. Unpublished doctoral dissertation, Universität Würzburg, Würzburg, Germany.

[4] Flege, J. (1995). 'Second language speech learning: Theory, findings, and problems.' In W. Strange (ed.) *Speech Perception and Linguistic Experience*: *Issues in Cross-Linguistic Research*. Baltimore: York Press, pp. 233–277.

[5] Hermas A. (2015). 'The categorization of the relative complementizer phrase in third-language English: A feature re-assembly account'. *International Journal of Bilingualism* 19: 587– 607.

[6] Liu, J., & Lin, J. (2021). 'A Cross-Linguistic Study of L3 Phonological Acquisition of Stop Contrasts'. *SAGE Open*, 11(1). https://doi.org/10.1177/2158244020985510

[7] Lloyd-Smith, A. (2021). 'Perceived foreign accent in L3 English: the effects of heritage language use'. *International Journal of Multilingualism*.
DOI: 10.1080/14790718.2021.1957899

[8] Puig-Mayenco, E., González Alonso J. & Rothman J. (2020). 'A systematic review of transfer studies in third language acquisition'. *Second Language Research* 36(1): 31–64.

[9] Rothman J. (2015). 'Linguistic and cognitive motivations for the Typological Primacy Model (TPM) of third language (L3) transfer: Timing of acquisition and proficiency considered'. *Bilingualism: Language and Cognition* 18(2): 179–90.

[10] Slabakova R. (2017). 'The scalpel model of third language acquisition'. *International Journal of Bilingualism* 21: 651–66.

[11] Westergaard M, Mitrofanova N, Mykhaylyk R. (2017) 'Crosslinguistic influence in the acquisition of a third language: The Linguistic Proximity Model'. *International Journal of Bilingualism* 21: 666–82.

[12] Wrembel, Magdalena, Ulrike Gut, Romana Kopečková, & Anna Balas. (2020). 'Cross-Linguistic Interactions in Third Language Acquisition: Evidence from Multi-Feature Analysis of Speech Perception' *Languages* 5, no. 4: 52.

[13] Wunder, E. M. (2011). 'Crosslinguistic Influence in Multilingual Language Acquisition: Phonology in Third or Additional Language Acquisition'. In G. De Angelis and JM. Dewaele (eds.) *New trends in Crosslinguistic Influence and Multilingualism Research*. Bristol, Blue Ridge Summit: Multilingual Matters, 2011, pp. 105-128.

Influence of Language Familiarity on Speaker Age Estimations

Tadayyon Chahartagh, Nafiseh[1], Gocsál Ákos[2,3]
[1]Alzahra University, Tehran, Iran
[2] University of Pécs, Faculty of Music and Visual Arts, Institute of Music, Pécs
[3] Hungarian Research Centre for Linguistics, Budapest

Previous research on speaker age estimation from voice has demonstrated patterns related to the accuracy of estimations. Most of the studies have shown that young adults are usually believed to be older than they are, the age estimations of middle-aged speakers are usually fairly accurate, while older speakers are usually thought to be younger than their calendar age [10]. Other factors that possibly influence accuracy, such as listeners' own age [6] or smoking [12] have also been investigated. Several studies have found that listeners' native language can also affect their age estimation accuracies. Jiao, Watson [4] and Nagao Nagao and Kewley-Port [7] showed that listeners were more accurate in estimating ages of speakers of their own native languages than when they heard speakers of other languages. Our objective is, therefore, to investigate if the same phenomenon, reported in the literature, exists with Iranian listeners who hear voices in their native language and a language, which is unknown to them. The present paper is a pilot study of a larger research project that aims to study the influence of language familiarity in the estimation of speaker age by Iranian and Hungarian listeners. To date, no similar research has been carried out between Hungarian and Azerbaijanian Turkish. The advantage of using these languages is that the foreign language will most likely be very unfamiliar for the listeners, so not only the possible influence of linguistic contents, but also their previous experiences with the languages or their speakers can be ruled out.

Turkish is a member of Altaic branch of Ural-Altaic family of languages. Azerbaijanian Turkish (or Azerbaijani or Azeri), is a member of the southwestern (or Oğuz) group of Turkic languages [1, 3] and has the largest number of speakers among non-Persian population of Iran [2]. Urmia, where the data were collected, is the capital of West-Azerbaijan province, located in the north-west of Iran. The majority of Urmia's population are Azerbaijanian Turks but there are also Kurdish, Armenian, and Assyrian minorities. Most of the people can also speak Persian as the official language of the country. Turkish is an agglutinating language in which the grammatical elements are attached together in a way that can be segmented easily. Word order in this language is pragmatically controlled and is fluid relative to, for example, English [11]. Azerbaijanian Turkish has 9 vowels and 26 consonants [5].

In this study, 20-30 seconds long spontaneous speech samples were played to listeners, who were asked to estimate the age of the speakers in years. Speakers included 5 males and 5 females per language, all natives of Hungarian and Azerbaijanian Turkish. The speakers were selected from age ranges 20-30, 30-40, 40-50, 50-60, and 60-70 years, one male and female from each age group and each language. The Hungarian samples were selected from BEA spontaneous speech database, from the "interview" or "argument" task of the BEA recording protocol [8]. Turkish speakers were selected using friend-of-a-friend technique [9] from Tabriz, East-Azerbaijan, Iran. Due to the social situation dominant in Iran during the data collection process, the speakers were asked to record themselves using their mobile phones. They were instructed to sit in a silent place and keep their phone 20-25 centimeters distant from their mouth. The speakers' were asked to talk about one of their pleasant memories. The voice files were sent to the authors via e-mail

or Telegram. None of the Turkish or Hungarian speakers smoked or had hearing or speaking impairment.

The listeners included 39 students of medicine at Urmia University of Medical Sciences (UMSU), Urmia, West-Azerbaijan, Iran (age range: 20-31), all of them native speakers of Turkish. None of the participants had prior phonetic training and none of them reported any kind of hearing impairment or previous medical treatment that may have influence their auditory process. Excerpts of 25-30 seconds were selected from the middle of each voice file between major prosodic boundaries. The excerpts were randomized, coded, and added to a playlist by the researchers prior to the experiment and were played in a silent classroom at the Faculty of Medicine of UMSU. Listeners were first familiarized with the task by playing three sound samples to them, then they were asked to record their age estimations on an answer sheet in years. Age estimation data were entered into a table in Excel 365 and SPSS 27. Statistical analysis included the calculation of Pearson's correlation coefficients and fitting a linear mixed model.

Correlation was found between calendar age and mean estimated age ($r = .745$, $p < .001$). Calculating *r* for Hungarian speakers ($r = .677$, $p < .032$) and Turkish speakers ($r = .843$, $p < .01$) suggested more consistent age estimations when listeners heard speakers of their native language. Next, accuracy data (i.e. the difference of calendar age and estimated age) were calculated for each estimation. A linear mixed model was fit on the data, with accuracy as dependent variable, and speaker age (in two groups, young and old), speaker gender, and speaker nationality as fixed factors and listener as the random factor. The model with the lowest AIC value was finally selected. Significant interaction was found between speaker gender and speaker language (estimate = -14.602, std. error = 2.933, df = 734, t = -4.944, $p < .001$). An inspection of the estimated marginal means revealed that that Hungarian males were more likely to be believed younger (EMM = -7.821, std. error = 1.209, df = 150.305) than the Turkish speaking males (EMM = -.269, std. error = 1.209, df = 150.305), while the Turkish speaking females (EMM = -.842, std. error = 1.209, df = 150.305), were believed to be younger than Turkish speaking males. EMM for Hungarian female speakers was -2.451 (std. error = 1.209, df = 150.305). The three-way interaction was also significant (estimate = 17,320, std. error = 3.956, df = 734, t = 4.378, $p < .001$) and the inspection of the estimated marginal means revealed the prominance of age underestimation of older Hungarian males (EMM = -15.436, std. error = 1.679, df = 409.559).

These findings support the previous results that language familiarity affects speaker age estimations, however, the significance of the interaction terms highlights that this effect varies across different conditions. Further studies should address the role of acoustic parameters, and possibly age stereotypes associated to certain sets or intervals of those parameters, which may explain the differences found here. One might believe that the perception of a biological feature is universal across cultures. The differences suggest that social perception, even the estimation or judgments related to biological features of humans seem to be influenced by cultural factors, e.g. the native language of the speaker. It is believed that understanding the role of those factors may help better understand human communication and social relations.

References

[1] Brown, K., & Ogilvie, S. (2010). *Concise Encyclopedia of Languages of the World*. Elsevier.

[2] Crystal, D. (2010). *The Cambridge encyclopedia of language a dictionary of language*. Cambridge: Cambridge University Press.

[3] Heijat, J. (2001). *Seiri Dar Tarix-e Zaban Va Lahjeha-ye Torki (A survey of the History of Turkic languages and dialects)*. Tehran: Nashr-e Peykan.

[4] Jiao, D., Watson, V., Wong, S. G.-J., Gnevsheva, K., & Nixon, J. S. (2019). Age estimation in foreign-accented speech by non-native speakers of English. *Speech Communication*, 106, 118–126. https://doi.org/10.1016/j.specom.2018.12.005

[5] Mokari, P. G., & Werner, S. (2017). *Azerbaijani*. Journal of the International Phonetic Association, 47(2), 207–212.

[6] Moyse, E. (2014). Age Estimation from Faces and Voices: A Review. *Psychologica Belgica*, 54(3), 255–265. https://doi.org/10.5334/pb.aq

[7] Nagao, K., & Kewley-Port, D. (2005, oct). The Effect of Language Familiarity on Age Perception. *International Research Conference on Aging and Speech Communication, Bloomington, Indiana*.

[8] Neuberger, T., Gyarmathy, D., Gráczi, T. E., Horváth, V., Gósy, M., & Beke, A. (2014). Development of a large spontaneous speech database of agglutinative Hungarian language. In *International Conference on Text, Speech, and Dialogue* (p. 424–431). Cham: Springer.

[9] Podesva, R. J., & Sharma, D. (2014). *Research methods in linguistics*. Cambridge: Cambridge University Press.

[10] Skoog Waller, S., Eriksson, M., & Sörqvist, P. (2015). Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology*, 6(978). https://doi.org/10.3389/fpsyg.2015.00978

[11] Slobin, D. I., & Zimmer, K. (1986). *Studies in Turkish Linguistics*. Amsterdam/Philadelphia: J. Benjamins Publishing Company.

[12] Winkler, R. (2010). Influence of smoking on the accuracy of age estimations from voice. In S. Fuchs, P. Hoole, C. Mooshammer, & M. Żygis (eds.), *Between the regular and the particular in speech and language* (p. 77–96). Frankfurt/M.: Peter Lang.

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. FK-128814.

Az artikuláció, a beszéd és az agyi jelek elemzése nyelvultrahang és elektroenkefalográf használatával

Csapó Tamás Gábor[1], Arthur Frigyes Viktor[1],
Nagy Péter[2,3], Boncz Ádám[3]

[1] BME, Távközlési és Médiainformatikai Tanszék
[2] BME, Méréstechnika és Információs Rendszerek Tanszék
[3] TTK, Kognitív Idegtudományi és Pszichológiai Intézet,
Hang- és Beszédészlelési Kutatócsoport, Budapest

Az agy-számítógép interfészek (Brain-Computer Interface, BCI) lehetővé tehetik a számítógépek közvetlen, fizikai aktivitás nélküli vezérlését. A beszéd BCI-k esetén a hang kimenet, azaz egyfajta némabeszéd-interfész kidolgozásá a cél az agyi jelek alapján. Az idegrendszeri jel rögzítési módok közül a BCI számára az elektroenkefalográfia (EEG) lehet a legmegfelelőbb, mivel lényegesen kisebb kockázattal jár, mint az invazív módszerek. Nemzetközi szinten végeztek már kezdeti kutatásokat EEG és beszéd alapú BCI kidolgozására [2, 7], azonban ez még nem eredményezett jól érthető szintetizált beszédet. Az artikulációs mozgáshoz kapcsolódó információ segíthetne ebben [1], azonban egy kivételtől eltekintve [9] eddig csak beszédből származtatott adatokon keresztül vizsgálták az agyi jelekkel párhuzamosan az artikulációt [1, 6, 8].

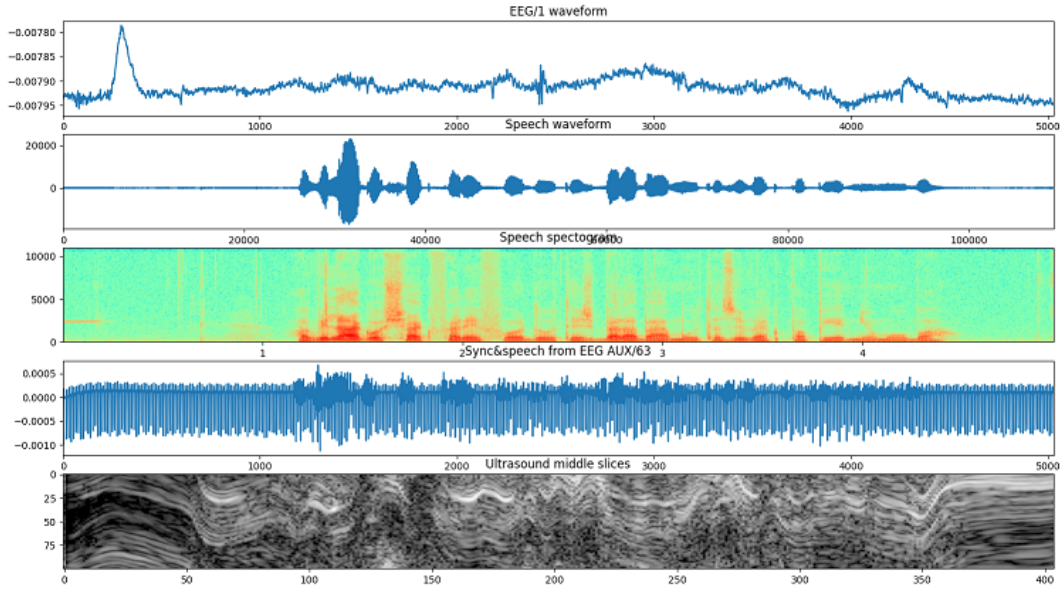
A jelen kutatásban az agyi jelek elemzését nyelvultrahang-alapú artikulációs adatokkal bővítjük, hogy jobb minőségű beszéd BCI-t tudjunk majd létrehozni. A felvételek az ELKH TTK egyik csendes szobájában készültek. Az EEG jelet 64 csatornás Brain Products actiCHamp típusú erősítővel rögzítettük (részletek: [5]). A nyelv középvonalának (szagittális) mozgását a „Micro" rendszerrel rögzítettük (AAA v220.02 szoftver, Articulate Instruments Ltd.) egy 2–4 MHz frekvenciájú, 64 elemű, 20 mm sugarú konvex ultrahangvizsgálófejjel, 82 fps sebességgel, és rögzítő sisakot is alkalmaztunk. A fém sisakot az EEG érzékelők fölé helyeztük el úgy, hogy az eszközök lehetőleg ne zavarják egymást. A beszédet egy Beyerdynamic TG H56c tan omnidirekcionális kondenzátor mikrofonnal vettük fel, és M-Audio M-Track 2x2 USB-s külső hangkártyával digitalizáltuk, 44 100 Hz-en. A hangkártya kimenetét (amely a „Micro" ultrahang „frame sync" szinkronizáló jelét és a mikrofonból származó beszédjelet együttesen tartalmazza) rákötöttük az EEG AUX csatornájára – így az agyi és artikulációs jeleket külön számítógépeken rögzítettük, de mégis tudjuk utólag szinkronizálni az adatokat (ld. 1. ábra, részletesen az előadásban). Eddig három magyar anyanyelvű férfi beszélővel rögzítettünk 3 x 10 percnyi felvételt: az adatközlők a PPBA adatbázisból származó mondatokat olvastak fel, a szöveg megjelenését követően azonnal, azaz a szöveg komolyabb értelmezése nélkül.

A kutatás kezdeti fázisában egy egyszerű kísérletet végeztünk: az EEG-vel mért agyi jelből mély neuronhálóval az artikulációs mozgásra vonatkozó információt (nyelvultrahang képek) prediktáltunk. Teljesen kapcsolt mély neurális hálózatot (fully connected deep neural network, FC-DNN) tanítottunk, melynek során a Hilbert-transzformált EEG bemenetből [10] nyelvultrahang képeket predikáltuk. Kísérleteink során egy 5 rejtett réteges, rétegenként 1000 neuront tartalmazó neuronháló struktúrát használtunk (hasonlóan az első nyelvultrahang-beszéd szintézis tanulmányunkhoz, [3]). Az eredmények szerint az EEG és nyelvultrahang közötti kapcsolat kimutatható [5]. Az eredmények ábrázolásához minden nyelvultrahang képből kivágtuk a középső függőleges vonalat (kb. ez megfelel a nyelv közepének), és ezen vonal időbeli változását ábrázoltuk, egy spektrogramhoz (hang-

színképhez) hasonlóan. A 2. ábra ennek eredményét mutatja: felül a beszédhez tartozó spektogram, középen az ugyanezen bemondáshoz tartozó nyelvultrahang középvonal időbeli változása, alul pedig a DNN által prediktált nyelvultrahang középvonal látható. Az a) mel-spektrogram és a b) artikulációs mozgás közötti hasonlóság egyértelműen észrevehető: a beszédben lévő formánsmozgások, és a nyelv függőleges mozgása nagyjából kivehető az ábrákon. A c) DNN-predikált nyelvultrahang középvonalon viszont a nyelv mozgása nem látszik, azaz a DNN nem tudta megtanulni az EEG és a nyelvultrahang közötti összefüggést. Ugyanakkor valamilyen információ mégis látszik a DNN-predikált képeken: a 195. időpillanat végén az egyik mondatnak vége van, és a következő elkezdődik, ami a b) eredeti nyelvultrahangon jól kivehető, és a c) becsült nyelvultrahangon is látszik.

A jelen előadásban ismertetett multimodális (EEG, nyelvultrahang és beszéd) analízis és szintézis révén túlmutatunk a legkorszerűbb nemzetközi trendeken. A kutatás hosszútávú célja, hogy hozzájáruljunk a beszéd alapú agy-számítógép interfészekhez. Az eredmények potenciálisan alkalmazhatók lehetnek a mozgássérült személyek rehabilitációs eszközeiként. A jövőben tervezzük a pszicho-neurolingivsztikai szempontokat is figyelembe venni.

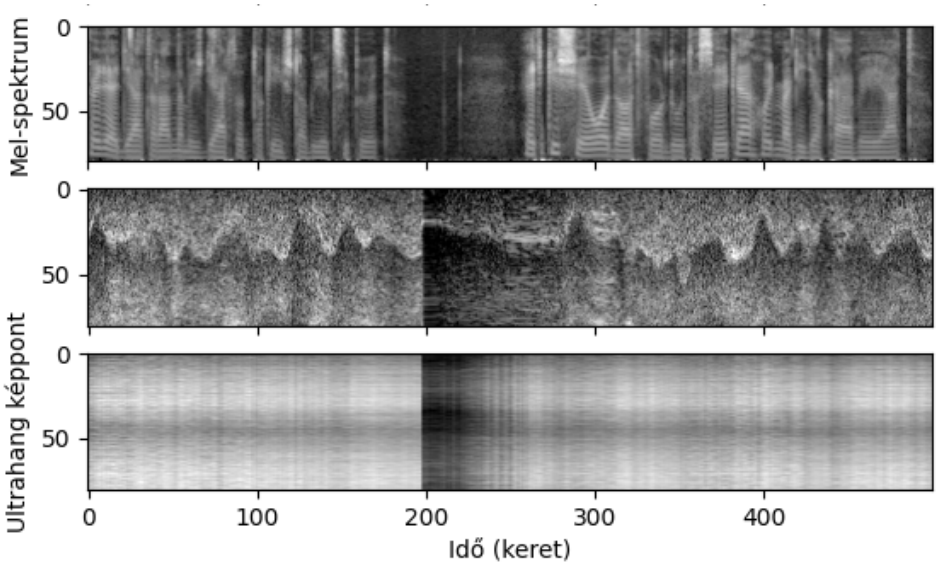
A kutatást az OTKA FK 142163 projekt finanszírozta [4]. Csapó Tamás Gábor kutatásait az MTA Bolyai János kutatói ösztöndíja, valamint az Új Nemzeti Kiválóság Program Bolyai+ (ÚNKP-22-5-BME-316) pályázata támogatta.



1. ábra. Példa a szinkronizált EEG, beszéd, és nyelvultrahang felvételre. a) EEG / 1. csatorna, b) beszédjel, c) beszéd spektrogram, d) ultrahang szinkornizálójel és beszédjel (EEG AUX-on), e) nyelvultrahang képek középső vonalának időbeli változása („kimogram").

Hivatkozások

[1] Anumanchipalli, G. K., J. Chartier & E. F. Chang (2019). „Speech synthesis from neural decoding of spoken sentences". *Nature* 568.7753, 493–498. old. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1119-1.



2. ábra. Demonstrációs minta: a) eredeti beszédminta 80-dimenziós mel-spektrogramja, b) eredeti nyelvultrahang felvétel középvonalának időbeli változása („kimogram"), c) az EEG-ből becsült nyelvultrahang középvonalának időbeli változása.

[2] Arthur, F. V. & T. G. Csapó (2022). „Deep learning alapú agyi jel feldolgozás és beszédszintézis előkészítő munkálatai". *MSZNY 2022*. online, 185–198. old.

[3] Csapó, T. G. és tsai. (2017). „Beszédszintézis ultrahangos artikulációs felvételekből mély neuronhálók segítségével". *MSZNY 2017*, 181–192. old.

[4] Csapó, T. G. és tsai. (2022). *OTKA FK-22, Analysis of articulation and brain signals for speech-based brain-computer interfaces*. URL: https://www.researchgate.net/project/Analysis-of-articulation-and-brain-signals-for-speech-based-brain-computer-interfaces.

[5] — (2023). „A beszéd artikulációs mozgásának predikciója agyi jel alapján - kezdeti eredmények". *MSZNY 2023*, 357–368. old.

[6] Favero, P. és tsai. (2022). „Mapping Acoustics to Articulatory Gestures in Dutch: Relating Speech Gestures, Acoustics and Neural Data". *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 802–806. old. ISBN: 9781728127828. DOI: 10.1109/EMBC48229.2022.9871909.

[7] Krishna, G. és tsai. (2020). „Speech Synthesis Using EEG". *Proc. ICASSP*. online, 1235–1238. old. ISBN: 9781509066315. DOI: 10.1109/ICASSP40776.2020.9053340. arXiv: 2002.12756.

[8] Le Godais, G. (2022). „Decoding speech from brain activity using linear methods". Dissz. Université Grenoble Alpes.

[9] Lesaja, S. és tsai. (2019). „Decoding Lip Movements during Continuous Speech using Electrocorticography". *International IEEE/EMBS Conference on Neural Engineering, NER*. San Francisco, CA, USA, 522–525. old. ISBN: 9781538679210. DOI: 10.1109/NER.2019.8716914.

[10] Verwoert, M. és tsai. (2022). „Dataset of Speech Production in intracranial Electroencephalography". *Scientific Data 2022 9:1* 9.1, 1–9. old. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01542-9.

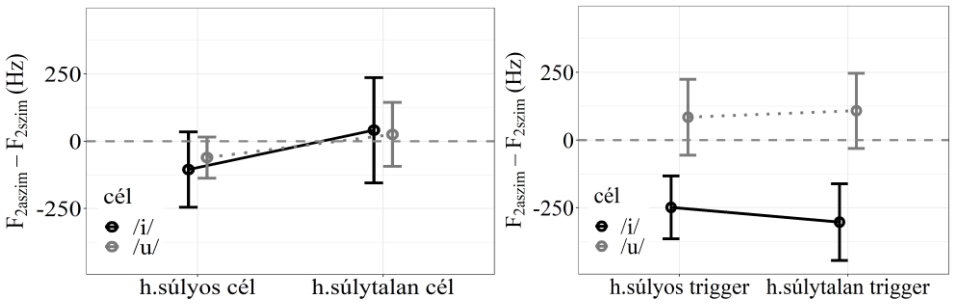
A hangsúly hatása a magánhangzók koartikulációs ellenállására és agressziójára

Deme Andrea[1,2], Bartók Márton[2], Gráczi Tekla Etelka[3,2], Csapó Tamás Gábor[4,2],
Juhász Kornélia[1,2,3] és Markó Alexandra[2]
[1]Eötvös Loránd Tudományegyetem, Budapest
[2]MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport, Budapest
[3]Nyelvtudományi Kutatóközpont, Budapest
[4]Budapesti Műszaki és Gazdaságtudományi Egyetem, Budapest

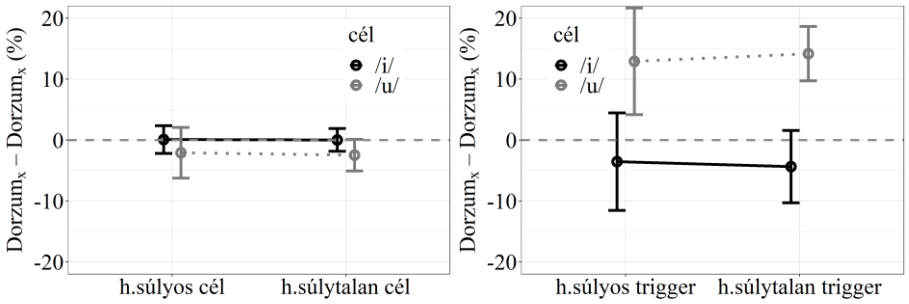
Magánhangzók közötti koartikulációnak a szakirodalom azt a folyamatot tekinti, amikor nem a szomszédos beszédhangok, hanem a hangsorban egymástól távolabb álló magánhangzók hatnak egymásra egy/több közbeeső mássalhangzón (és akár több további magánhangzón/szótagon) keresztül [7]. Az ennek a hatásnak az eredményeként előálló változatosság mértékét bizonyos tényezők befolyásolhatják. Vannak olyan beszédhangok, amelyek eleve ellenállóbbak vagy rezisztensebbek a koartikulációs hatásokkal szemben [8], de vannak olyan fonetikai pozíciók is, amelyek növelhetik a rezisztenciát azért, mert „erősítik" a szegmentumot [2]. Ilyen lehet például a hangsúlyos helyzet [1, 5]. Ezzel párhuzamosan azt is feltételezik, hogy az „erősebb" szegmentumok koartikulációs agresszivitása – azaz a hatásuk a környező beszédhangokra – is nagyobb. Ennek a feltevésnek (tudomásunk szerint) a magánhangzók közötti kölcsönhatásokat illetően csak egy átfogóbb kísérletes ellenőrzése létezik: itt bár a megnövekedett rezisztenciát kimutatták, az agresszió megnövekedését nem támasztották alá az eredmények a beszéd artikulációs vetületében angol nyelvű beszédben [1]. A jelen kutatásban azt a kérdést vizsgáltuk meg, hogy a magyarban, ahol a prozódia szerepe a kiemelésben korlátozottabb lehet [6], kimutatható-e megnövekedett koartikulációs rezisztencia és agresszió (egymással együtt járóan) mondathangsúlyos szótagokban. A jelen kísérletben – az idézett korábbi vizsgálathoz [1] hasonlóan – álszavakban vizsgáltuk a kérdést, de attól eltérően az artikulációt és akusztikumot párhuzamosan vizsgáltuk (ráadásul nem egy, hanem két mérőszám segítségével). Megjegyezzük, korábbi vizsgálatok valódi szavakban alátámasztották azt, hogy a hangsúly növeli a magánhangzók koartikulációs ellenállását az artikulációban és az akusztikumban egyaránt (ott a beszédhangok koartikulációs agressziója az anyag jellegzetességei miatt nem volt elemezhető) [3, 4].

A jelen vizsgálatban /pVpVpVpV/ szerkezetű szavakat elemeztünk, ahol a magánhangzók az /u/ és /i/ voltak. A kísérletben 9 nő vett részt, a célmagánhangzókat akusztikailag (az F_2 frekvenciaértéke szerint) és artikulációsan (a vízszintes nyelvtesti helyzet szerint) elemeztük, ez utóbbit elektromágneses artikulográfiával. A koartikulációs hatásokat kétféleképpen számszerűsítettük: a neutrális és koartikuláló helyzetű magánhangzók minőségbeli különbségével (azaz az egyes mérőszámok különbségeként; a továbbiakban *távolság*), illetve a neutrális és koartikuláló helyzetű magánhangzók együttes változatosságával (azaz az egyes mérőszámok kontextusok közötti relatív szórásaként; a továbbiakban *szóródás*). Vizsgáltuk a célhangok változatosságát a hangsúly hatására (a rezisztencia elemzésére), illetve az indukáló magánhangzó hangsúlyosságának hatására (az agresszió elemzésére). (Mivel a vizsgált szavakat egyesével mutattuk be a résztvevőknek, akiket arra kértünk, hogy azokat „mondatszerűen" olvassák fel, azt várhattuk, hogy – hasonlóan az egyszavas közlésekhez – ezekben az esetekben az első szótagon jelenik meg a közlésegységszintű vagy mondatszintű hangsúly, és ezt a várakozást a kísérletvezetők percepciója megerősítette.)

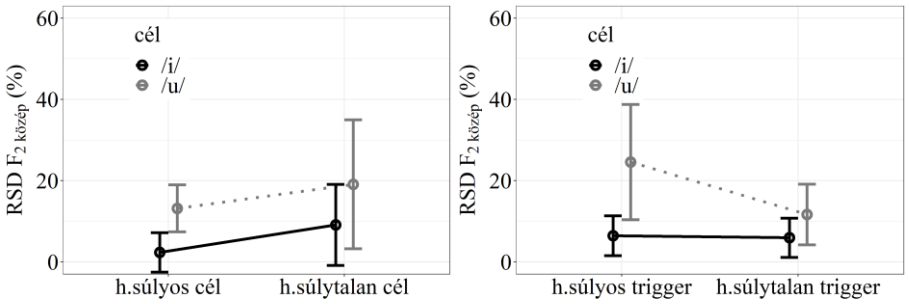
Az eredmények azt mutatták, hogy az artikulációban és az akusztikumban megmutatkozó tendenciák egyeznek: a vizsgált álszavakban a hangsúly a magánhangzóknak sem a koartikulációs ellenállását, sem pedig az agresszióját nem növelte meg egyik doménben sem (1–4. ábra). Ez ellentmond korábbi valódi szavas kísérletek eredményeinek [3, 4], amiből arra következtetünk, hogy a rezisztencianövekmény hiánya az álszavaknak köszönhető: ezek ejtése hiperartikulált lehet, ami pedig erősítést vonhat magával [8]. Mivel azonban a más nyelvekre kapott adatok ugyancsak álszavas kísérletekből származnak, azt is fel kell tételeznünk, hogy az eredmények a magyar nyelv sajátosságaival is összefüggnek, ahol a prozódia kisebb súllyal alkalmazódik a kiemelésben [6], tehát kevésbé képes a szegmentumok erősítésére is. Az eredmények fontos adalékkal szolgálnak a folyamatos beszéd leírásának nyelvfüggő sajátosságaihoz.



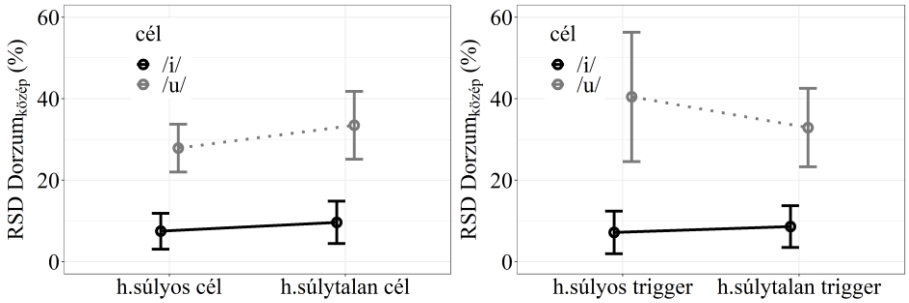
1. ábra: Akusztikai távolság: rezisztencia (balra) és agresszió (jobbra)



2. ábra: Artikulációs távolság: rezisztencia (balra) és agresszió (jobbra)



3. ábra: Akusztikai szóródás: rezisztencia (balra) és agresszió (jobbra)



4. ábra: Artikulációs szóródás: rezisztencia (balra) és agresszió (jobbra)

Irodalom

[1] Cho, T. (2004). ’Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English’. *Journal of Phonetics* 32, 141–176. old.

[2] de Jong, K. J. (1995). ’The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation’. *Journal of the Acoustical Society of America,* 97, 491–504. old.

[3] Deme A., Bartók M., Gráczi T. E., Csapó T. G., & Markó A. (2019). ’A mondathangsúly hatása a magánhangzók megvalósulásának változatosságára’. *Nyelvtudományi Közlemények 115.* 199–232. old.

[4] Deme, A., Bartók, M., Csapó, T. G., Gráczi, T. E., & Markó, A. (é.n.). ’The effect of pitch-accent on the acoustic and articulatory variability of vowels: a real-word EMA study’. Kézirat.

[5] Fowler, C. S. (1981). ’Production and perception of coarticulation among stressed and unstressed vowels’. *Journal of Speech and Hearing Research* 24, 127–139. old.

[6] Mády, K., & Kleber, F. (2010). ’Variation of pitch accent patterns in Hungarian. *Proceedings of Speech Prosody 2010’.* http://speechprosody2010.illinois.edu/papers/100924.pdf [letöltés: 2022. július 11.]

[7] Öhmann, S. E. G. (1966). ’Coarticulation in VCV utterances: spectrographic measurements’. *Journal of the Acoustical Society of America* 39, 151–168. old.

[8] Recasens, D., & Espinosa, A. (2009). ’An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan’. *Journal of the Acoustical Society of America* 125, 2288–2298. old.

[9] Scarborough, R. (2012). ’Lexical similarity and speech production: Neighbourhoods for nonwords’. *Lingua* 122, 164–176. old.

Indulatszóval kezdődő függetlenedett mellékmondatok a mai beszélt és írott beszélt magyar nyelvben: léteznek?

Dér Csilla Ilona[1] [2]
[1] Károli Gáspár Református Egyetem, Budapest
[2] Nyelvtudományi Kutatóközpont, Budapest

Egy korábbi kutatás [1] tanúsága szerint egyes *hogy* kötőszós függetlenedett (inszubordinált) mellékmondatok [2] keletkezésében az indulatszóval (*ah, jaj, óh~oh*) álló változatok meghatározó szerepet játszhattak. A mai magyar köznyelvben már nem használt, pozitív kívánságokat kifejező konstrukciók (1) kizárólag indulatszókkal jelentkeztek a 18. század végétől kezdve, de azok általánosak voltak az átkokban/negatív kívánságokban (2) is:

(1) *SARA. Él? igazán él? –* ***Oh, hogy*** *még sokáig éljen! hogy igen szerencsés, örömökben gazdag napokat éljen!* ***Oh, hogy*** *Isten az ő életéhez toldja az én napjaim' felét!* (Lessing: Miss Sara Sampson. Fordítá Kazinczy Ferenc, Buda, 1842)
(2) ***Oh hogy*** *szakadna rám Kárpát ' nehéz hegye. 's feszűlne égig fáradt testem, ne kellene többé futnom e' világon.* (MTSz, Vörösmarty Mihály: Összes művei 6. Drámák 1. Zsigmond, 1823: 228)

Jelen vizsgálat fő célja annak tisztázása, hogy az ilyen, indulatszóval kezdődő függetlenedett mellékmondatok milyen mértékben és formákban elterjedtek a mai magyar spontán és félspontán beszélt és az írott beszélt nyelvben, különösen óhajtó partikulával álló nem inszubordinált megoldások variánsaiként. Brdarné Szabó korábbi kutatása szerint [3] magyar anyanyelvű egyetemista adatközlők angol és német függetlenedett feltételes mellékmondatok fordításaként csak az utóbbiakat (pl. (6)) tartották elfogadhatónak, az inszubordinált verziókat (pl. (5)) nem:

(3) ***If*** *we could get rid of him!*
(4) ***Wenn*** *wir ihn (doch) bloß loswerden könnten!*
(5) ***Ha** meg tudnánk szabadulni tőle!
(6) **Bárcsak** meg tudnánk szabadulni tőle!

Előzetes korpuszvizsgálatok [4] alapján azonban úgy tűnik, hogy léteznek az (5)-öshöz hasonló alakulatok, csak indulatszóval társulva:

(7) *Ömlöttek a sorok, készült a vers, töprengés és javítgatás nélkül.* ***Ó ha*** *még egyszer így mehetne!* (MNSz2, doc#1376, szépirodalom)
(8) ***jaj ha*** *nekem lenne egy ilyen csöppségem……* (MNSz2, doc#2886, személyes-közösségi)

Az elemzést mindkét függetlenedett mellékmondati típusra: a *hogy* kötőszóval álló mondatrészkifejtő, illetve a *ha, hogyha* kötőszóval induló feltételes mellékmondatokra elvégeztem: ezek összes, *jaj, oh~óh, ó, ah* indulatszóval kezdődő változatát kigyűjtöttem két korpuszból: az MNSz2 beszélt nyelvi (sajtó) és írott beszélt nyelvi (személyes: közösségi média és vitafórum) alkorpuszaiból, illetve a BEA [5] száz spontán társalgást tartalmazó alkorpuszából.

Az eredmények azt mutatják, hogy a BEA-társalgások alig tartalmaznak ilyen mellékmondatokat: kizárólag a *jaj*-jal induló *hogy*-változatra (9) volt összesen öt adat, a többire egy sem:

(9) *én is a ugye mindenki e amikor így meghallják hogy kutyát lakásban tartok akkor mindenki* ***jaj hogy ez izé*** *az milyen hogy állatkínzás meg mit tudom én*

Az MNSz2 teljes anyagában azonban gyakoribbak voltak az indulatszóval kezdődő inszubordinált mellékmondatok, és nagy változatosságot mutattak: az összesen 120 találatból csak az *ah* + *hogyha* és *oh* + *hogyha* kombinációk hiányoztak. Ezek közül a beszélt és írott beszélt nyelvi alkorpuszokból összesen 30 találat származott (a *hogyhá*-val kezdődő mondatokat leszámítva minden variánsra volt példa), ezekben az *ó+ha* dominált, akárcsak a teljes anyagban.
A korpuszelemzéseket egy kérdőíves kutatás egészítette ki [6], amelyben több korosztály (14–30, 30–60, 60 felettiek) esetében vizsgáltam az indulatszóval, valamint a nélkül álló függetlenedett (korábban mellékmondat volt) és önálló (nem volt mellékmondat), illetve az óhajtó partikulával (*bár/csak/*) kezdődő önálló változatok elfogadhatóságát és funkcióit. Ez a vizsgálat még folyamatban van – az előzetes eredmények szerint (38 kitöltő) a beszélői megítélések rendkívül nagy egyéni változatosságot mutatnak, de közös vonásuk, hogy a *hogyhá*-val álló változatokat az anyanyelvi beszélők nem találják jellemzőnek a mai magyar nyelvhasználatra.
A vizsgálatok feltételezésemet részben megerősítették, részben nem, mivel bár az indulatszóval álló konstrukciók ma is használatosak az írott beszélt nyelvben, a beszélt nyelvi (al)korpuszokból szinte teljesen hiányoznak. Fontos eredmény azonban, hogy a szépirodalomban, főként a versekben jóval gyakoribbak, mint más MNSz2-műfajokban, ami egybecseng azzal az eredménnyel [7], hogy a diskurzusjelölővel (*hát, szóval*) álló *hogy*-változatok is mennyiségileg igen számottevőek az irodalom nyelvében.

A kutatást az Nemzeti, Kulturális, Fejlesztési és Innovációs Hivatal K-128810 számú pályázata és a Bolyai János Kutatási Ösztöndíj támogatta.

Irodalom

[1] Dér, Cs. I. (2022). ‘*Ó, hogy azt tsak a’ Szerelmek ’S a’ lágy szellök hajtsák!* Az indulatszókat követő, illetve a nélkülük álló hogy kötőszós inszubordinált mondatok megjelenéséről’. *A nyelvtörténeti kutatások újabb eredményei XI.* Szerk. Forgács T. & M. Németh & B. Sinkovics. Szeged: Szegedi Tudományegyetem Magyar Nyelvészeti Tanszék, 45–57. old.

[2] Evans, N. (2007). ‘Insubordination and its uses’. *Finiteness. Theoretical and empirical foundations.* Szerk. Nikolaeva, I. Oxford: Oxford University Press, 366–431. old.

[3] Brdarné Szabó, R. (2006). ‘Stand-alone dependent clauses functioning as independent speech acts: A crosslinguistic comparison’. *The metaphors of sixty: Papers presented on the occasion of the 60th birthday of Zoltán Kövecses.* Szerk. Benczes, R. & Sz. Csábi. Eötvös Loránd University: Department of American Studies. 84−95. old.

[4] MNSz2 = Magyar Nemzeti Szövegtár 2. változat http://clara.nytud.hu/mnsz2-dev/

[5] BEA = Magyar Spontán Beszéd Adatbázis. http://www.nytud.hu/adatb/bea/index.html

[6] A kérdőív internetes lelőhelye: https://docs.google.com/forms/d/e/1FAIpQLScQ419oNl0C5xiHlrNzYToLMn0Yo2Rx3xjNypArtLzEvd-ywg/viewform?vc=0&c=0&w=1&flr=0

[7] Dér, Cs. I. (2022). '*Háthogy esetleg netalántán azt gondolnátok rólam…:* A *szóval* és a *hát* diskurzusjelölővel álló *hogy* kötőszós mellékmondatokról'. *Jelentés és Nyelvhasználat* 9.1: 43–57. old.

A diskurzusjelölők mint egyéni beszédjellemzők felhasználhatósága az igazságügyi nyelvész szakértői munkában

Főző Eszter[1], Gráczi Tekla Etelka[2] és Markó Alexandra[1,2]
[1] NBSZ Szakértői Intézet
[2] Nyelvtudományi Kutatóközpont

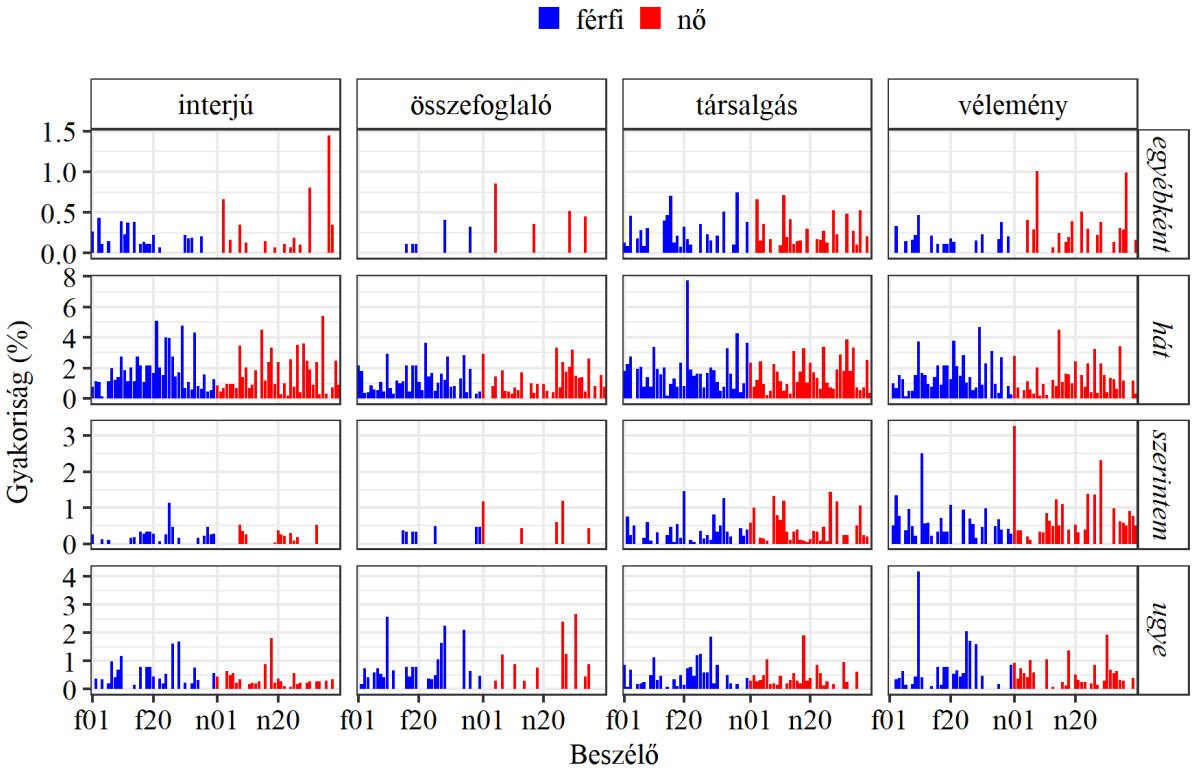
A diskurzusjelölőket olyan nyelvi elemek funkcionális csoportjaként tartják számon, amelyek a diskurzus szervezésében vesznek részt: diskurzusszegmenseket kötnek össze, illetve pragmatikai viszonyokat jelölnek. A diskurzusjelölők általában nem befolyásolják a megnyilatkozás igazságfeltételeit, ugyanakkor esetenként rendelkezhetnek emocionális és expresszív funkcióval. Elsősorban beszélt nyelvi elemeknek tartják őket, de az írott nyelvben, illetve az írott beszéltnyelviség szövegtípusaiban is előfordulnak bizonyos jellemző diskurzusjelölők [2].

A diskurzusjelölők használata és az egyéni nyelvi viselkedés (idiolektus) közötti összefüggések lehetősége korábbi tanulmányokban is felmerült már különféle szempontokból (pl. [7], [1]; [3], [11], [10]). Ugyanakkor nincs tudomásunk olyan kutatásról, amely ezt a kérdést több beszélő esetében, nagy anyagon, részletekbe menően vizsgálta volna, annak ellenére, hogy egy felmérésben 36 igazságügyi szakértőként működő válaszadó (4 kontinens 13 országából) 76%-ban válaszolta azt, hogy a diskurzusjelölőket (és más társalgási jellemzőket) használják az igazságügyi fonetikai összehasonlításban [4]. A diskurzusjelölők használata és az egyéni nyelvi viselkedés összefüggése a szövegtípusokon keresztül is megfigyelhető, aminek ugyancsak jelentősége van a kriminalisztikai beszélő- és szerzőazonosításban (vö. pl. [13]). A diskurzusjelölők fajtái, arányai és szerepei visszatükrözik a szövegtípus formalitását, stílusát, valamint legfőbb jellegzetességeit ([9], [6]).

A jelen kutatásban a diskurzusjelölők elfordulásait kvantitatív (gyakoriságok a különféle beszédtípusokban) és kvalitatív (jellemző kontextusok, diskurzustársulások, pozíciók) szempontból elemezzük annak érdekében, hogy egyéni mintázatokat tárhassunk fel. A diskurzusjelölők kiválasztásában két szempont vezérelt. Egyrészt olyan elemeket választottunk, amelyek a magyar szakirodalom alapján gyakorinak mutatkoznak ([3], [6], [10]). Másrészt, mivel nagy beszédanyagot elemeztünk automatikus módszerrel, alapvető szempont volt az, hogy a kiválasztott elemeknek ne legyenek gyakori homofón (nem diskurzusjelölői funkcióban jelentkező) megfelelői, amelyek megzavarják az adatokat. Mindezek alapján négy diskurzusjelölő, a *hát,* az *egyébként,* a *szerintem* és az *ugye* használatát vizsgáljuk. A korpuszunkat a BEA adatbázis [5] spontán és félspontán anyagából válogattuk, a vizsgált beszédfeladatok az interjú, a vélemény, a társalgás és a két tartalmi összefoglaló voltak. A diskurzusjelölők automatikus azonosítása a hangfelvételek leiratán [8] futtatott szkripttel történt. Az életkor tekintetében a kriminalisztikai szempontból releváns intervallumba ([14], [12]) tartozó 78 beszélőt (39 nő, 39 férfi) válogattunk ki, és nemenként 13-13 fős csoportokra osztottuk őket a következő életkori sávok szerint: a 21–25, a 28–39 és a 41–59 éves korúak csoportjába. A végzettség szerint a 78 beszélő közül 32 középfokú, 46 felsőfokú végzettségű volt.

A lejegyzett szövegekből kinyertük az egyes diskurzusjelölők előfordulásának számát és a teljes szószámot. Az előzetes eredmények alapján összesen 194531 szóból 4301-et, azaz 2,21%-ot tett ki a négy vizsgált diskurzusjelölő. Gyakori volt, hogy egy adott diskurzusjelölő nem fordult elő néhány beszélőnél egy-egy beszédtípusban. A *hát* előfordulása volt a leggyakoribb: szinte minden beszélőnél minden beszédtípusban megjelent, és nagyobb arányban, mint a többi diskurzusjelölő.

A legkevesebb előfordulást az *egyébként* esetében kaptuk, a beszélők közül is kevesebben használták. A négy diskurzusjelölőnek a szószámhoz viszonyított gyakoriságát beszédfeladatonként és beszélőnként az 1. ábra összegzi. A diskurzusjelölők előfordulási gyakorisága a legtöbb esetben gyenge vagy közepesen erős korrelációt mutatott az adott beszédminta összes szószámával, ami arra utal, hogy a szöveg hossza valamilyen mértékben befolyásolja a vizsgált diskurzusjelölők felhasználhatóságát a nyelvész szakértői elemzésben, ami egybevág az általános gyakorlati tapasztalattal. Másrészt a beszédfeladatok között a *hát* minden esetben és az *ugye* majdnem minden esetben szignifikáns korrelációt mutatott. Ez a korreláció gyenge, legfeljebb közepes erősségű, de azt jelzi, hogy aki az egyik beszédtípusban gyakrabban használja az adott diskurzusjelölőt, az a másikban is. Az *egyébként* és a *szerintem* esetében kevesebb beszédfeladatpár között találtunk szignifikáns korrelációt, ami arra enged következtetni, hogy ezek használata elsősorban a beszédfeladattól függ.



1. ábra. Az egyéni gyakorisági adatok alakulása (a beszédfeladatonkénti gyakoriságok átlaga)

Az előadásban a további kvantitatív és kvalitatív elemzések eredményeit is bemutatjuk, különös tekintettel a vizsgált diskurzusjelölők más diskurzusjelölőkkel alkotott társulásaira; a diskurzusjelölők jellemző pozícióira, illetve ezek esetleges beszédtípusfüggő eltéréseire.

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] Coulthard, M. & Johnson, A. (2007). *An Introduction to Forensic Linguistics. Language in Evidence.* London & New York: Routledge.

[2] Dér Cs. I. (2009). Mik is a diskurzusjelölők? In: Keszler B. & Tátrai Sz. (szerk.), *Diskurzus a grammatikában – grammatika a diskurzusban* (293–303). Budapest: Tinta Könyvkiadó.

[3] Dér, Cs. I. & Markó, A. (2010). A pilot study of Hungarian discourse markers. *Language and Speech, 53/2,* 135–80.

[4] Gold, E. & French, J. P. (2011). International practices in forensic speaker comparison. *International journal of speech, language and the law, 18 (2),* 293-307.

[5] Gósy M. (2012). Multifunkcionális beszélt nyelvi adatbázis – BEA. *Általános Nyelvészeti Tanulmányok, 24,* 329–349.

[6] Götz A. (2019). Diskurzusjelölők és kötőelemek gyakorisága írott és beszélt, mediált és nem mediált diskurzusokban. In Laczkó K. & Tátrai Sz. (szerk.), *Kontextualizáció és metapragmatikai tudatosság* (291–313). Budapest: ELTE Eötvös József Collegium.

[7] Kredens, K. (2000). *Forensic Linguistics and the Status of Linguistic Evidence in the Legal Setting.* PhD dissertation. University of Łodz.

[8] Mihajlik P., Gráczi T. E., Kohári A., Tarján B., Balog A., Mády K. (2022). A BEA továbbfejlesztése és alkalmazása kontrasztív gépi beszédfelismerési kísérletekre. *Általános nyelvészeti tanulmányok* 34. . 361−380.

[9] Schirm A. (2017). A diskurzusjelölők és a szövegtípusok viszonyáról. *Magyar Nyelv, 113,* 330–341.

[10] Schirm A. (2019). A diskurzusjelölő-használat életkori sajátosságai a nyelvi interjú szövegtípusában. *Beszédkutatás 2019,* 187–205. doi: 10.15775/Beszkut.2019.187-205

[11] Sclafani, J. (2017): Talking Donald Trump: A Sociolinguistic Study of Style, Metadiscourse, and Political Identity, Routledge

[12] Skarnitzl, R., & Vaňková, J. (2017). Fundamental frequency statistics for male speakers of common Czech. *Acta Universitatis Carolinae: Philologica, 2017/3,* 7–17.

[13] Sztahó, D., Beke, A., Szaszák, Gy. & Fejes, A. (2022). Forensic Authorship Classification by Paragraph Vectors of Speech Transcriptions. Berend Gábor, Gosztolya Gábor és Vincze Veronika (szerk.), *XVII. Magyar Számítógépes Nyelvészeti Konferencia* (275–288). Szeged: Szegedi Tudományegyetem Informatikai Intézet.

[14] Tatár Z. (2013). Beszélőprofil-alkotás lehetőségei a kriminalisztikai fonetikában. *Alkalmazott Nyelvtudomány, XIII/1–2,* 121–130.

Beszélőazonosítás 10 év különbséggel rögzített hangmintákon

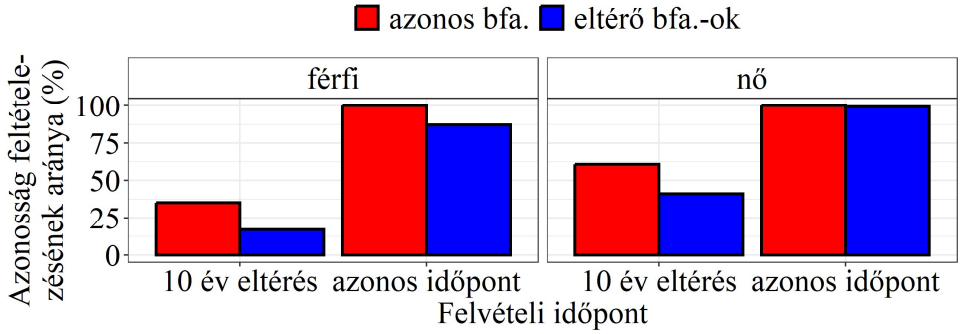
Fejes Attila[1], Huszár Anna[2], Krepsz Valéria[2,3], Gráczi Tekla Etelka[2]
[1]NBSZ Szakértői Intézet
[2]Nyelvtudományi Kutatóközpont
[3]Humboldt Universtität zu Berlin

A kriminalisztikai beszélőazonosítás során a vizsgálat anyagát képező beszédminták rögzítése között hosszabb-rövidebb idő telik el, ami – az ügy sajátosságaitól függően – éveket is jelenthet. A beszédjellemzők beszélőn belüli változása ezekben az esetekben befolyásolhatja az azonosítás eredményét. Korábbi kutatási eredmények nem minden esetben találtak adott időtáv alatt következetes eltérést a beszélők formánsértékeiben és alapfrekvenciájában, ugyanakkor a beszélők eltérő életkorúak voltak, így az azonos eltelt idő más-más életkorok között történt. Az alapfrekvencia 10 év elteltével történt vizsgálata során felmerült, hogy a fiatalabb (huszonéves) női beszélőkre jellemző, hogy tíz év után alacsonyabb értékek jellemzik beszédüket, míg az idősebb (harmincas) nőkre, illetve a férfiakra nem kaptak hasonló szisztematikusnak tűnő változást [4], a szegmentális jellemzők pedig nem mutattak szisztematikus eltérést [5]. Egy biometrikus beszélőazonosítással végzett kísérletben arra jutottak, hogy tíz év elteltével rögzített felvételek összevetése alacsonyabb hasonlósági pontszámokat, azaz Score-értékeket eredményez mind a női, mind a férfi beszélők esetében, mint az azonos felvételi időpontból származó beszédminták összevetése, de valamivel magasabbat, mint az eltérő beszélők esetében kapott eredmények [3]. A jelen vizsgálatban ezt a biometrikus elemzést bővítettük ki a vizsgálat időpontjában rendelkezésre álló 32 beszélőre, a Score-értékek kalibrációjával és a hangtechnikai szakértői véleményekben alkalmazott likelihood ratio (LR) számításával. Azt feltételeztük, hogy a tipikus eljárások során használt kalibrációval nyert Score- és LR-értékek azonos beszélőtől származó beszédminták összevetésére azonos beszédtípusból és azonos felvételi időből származó felvételek esetében a legmagasabbak, míg eltérő beszédtípusok esetében valamivel alacsonyabbak, míg a 10 év elteltével rögzített anyagokra a legalacsonyabbak, akár a beszélő azonosságának feltételezhetőségét megkérdőjelező érték alá csökkennek. Azt is feltételeztük, hogy mivel az akusztikai eltérések nem nagymértékűek a legtöbb beszélő esetében, a legtöbb beszélő azonosítható 10 év elteltével is, és hogy a legalacsonyabb értékeket az eltérő beszélőktől származó beszédminták összehasonlítására kapjuk.

A vizsgálatban 32 magyar anyanyelvű beszélő (19–45, majd 29–55 év; 15 nő, 17 férfi) spontán, olvasott és félspontán beszédfelvételét használtuk fel. A VOCALISE 2021 3.0.0.1746 verziószámú biometrikus beszélőazonosító rendszer [6] XVector-PLDA Default algoritmusával végeztük el a beszédminták összevetését. A biometrikus mérések első lépéseként raw Score eredményeket kaptunk. Ezen típusú pontszámok nagysága függ az azonosító rendszer és a biometrikus technológia típusától, valamint a hangminták karakterisztikájától, ezért különböző szoftverek és algoritmusok raw Score értékei nem hasonlíthatóak össze egymással. A Score értékek kalibrációját követően az adatokat hatványoztuk (mivel a VOCALISE szoftvernél az LR természetes logaritmusa a kalibrált Score), majd megkaptuk az LR értékeket, amelyek már összevethetőek egymással. A kalibrációhoz a ForVoice 120+ adatbázisának [2] 50 női és 50 férfi beszélő mintáját használtuk fel. A mintaadó személyek kiválasztása véletlenszerűen történt. A kalibrált Score-értékeket lineáris kevert modellek segítségével elemeztük [8, 1, 7]. A kapott értékeket mint függő változó, a beszélő, a beszédtípus, a felvételi időpont azonosságát mint fix hatás állítva be a modellben. Az LR-értékek alapján azonos és eltérő beszélőtől származónak kategorizáltuk a beszédmintákra kapott eredményeket ($LR \geq 1$ = azonos beszélő és $LR < 1$ = eltérő beszélő a hagyományos eljárások módszertana szerint). Ezeket a nyert

azonosságokat a valódi azonossággal összevetve a helyes és téves azonosság/eltérés csoportokra osztottuk, és a felvételi időpont azonossága, valamint a beszédtípus alapján elemeztük az arányukat.

A kalibrált Score- és LR-értékek jelentősen alacsonyabbak voltak 10 év eltérésével készült beszédminták összevetésekor, mint az azonos felvételi időpontban készült beszédmintáknál, ugyanakkor magasabbak, mint az eltérő beszélőktől származóak esetében. Emellett mind a nők, mind a férfiak esetében jelentősen csökkent az azonos beszélőtől származtathatóságra utaló LR-értékek aránya a 10 év elteltével rögzített beszédminták összevetésekor, mint az azonos időpontban rögzítettek esetében (1. ábra). Az összes eltérő beszélőtől származó beszédmintapár összevetés során a nők esetében 98,81%, a férfiak esetében 100%-ban kaptunk 1-nél alacsonyabb LR-értéket. A 10 év elteltével összehasonlított minták esetében a női beszélőknél közel a beszédmintapárok feléről (41,1% és 60,7%), míg a férfiak esetében 17,2% és 35,3%-ban feltételezhető az LR-értékek alapján, hogy azonos beszélőtől.



1. ábra: Azonos beszélőtől származó beszédminták LR-értékek alapján azonos beszélőtől feltételezhető származásának aránya (%) (bfa. = beszédfeladat)

Az LR-értékek alacsonyabb mivolta a 10 év elteltével rögzített anyagok esetében feltehetően abból következik, hogy ekkora időtávban a beszélőn belüli variabilitás nagyobb mértékű, mint a hagyományos eljárás során feltételezhető variabilitás annak ellenére, hogy a korábbi, akusztikai vizsgálatok alapján szisztematikus változás nem várható. Ennek magyarázata, hogy ekkora időtávban egészséges fiatal és középkorú beszélők esetében nem egyértelműen az életkor okozta következetesebb változások nagy mértékűek, hanem feltehetőleg az eltérő szociolingvisztikai tényezők változása mentén tapasztalható nagyobb mértékű eltérés az egyazon beszélőtől rögzített beszédminták között.

Köszönetnyilvánítás

A kutatás az FK128814-es pályázat keretében készült. A kalibrációhoz használt adatbázis az FK128615-ös pályázat keretében készült. Köszönjük Sztahó Dávidnak a projekt vezetőkutatójának, hogy a felvételeket a rendelkezésünkre bocsátotta.

Irodalom

[1] Bates, D., Maechler, M., Bolker, B. & Walker, W. (2015). ‘Fitting Linear Mixed-Effects Models Using lme4’. *Journal of Statistical Software,* 67(1), 1–48. old. DOI:10.18637/jss.v067.i01.

[2] Beke, A., Szaszák, Gy. & Sztahó, D. (2020). ‘FORvoice 120+: magyar nyelvű utánkövetéses adatbázis kriminalisztikai célú hangösszehasonlításra’. In Berend, G., Gosztolya, G., Vincze, V. (szerk.): XVI. Magyar Számítógépes Nyelvészeti Konferencia. MSZNY 2020. Szeged, Magyarország. Szegedi Tudományegyetem, Informatikai Intézet 95−101. old.

[3] Gráczi, T. E., Fejes A., Krepsz V., Huszár A. (2023, in press). ‘Speaker recognition over the course of 10 years and speech styles’. *Journal of Applied Linguistics.*

[4] Gráczi T. E., Huszár A., Krepsz V. & Markó A. (2022). ‘Az alapfrekvencia-jellemzők középtávú longitudinális változása fiatal és középkorú felnőttek beszédében’. In Mády K. & Markó A. (szerk.): *Általános nyelvészeti tanulmányok 34.: Fonetikai tanulmányok.* Budapest, Magyarország: Akadémiai Kiadó. 111−142. old.

[5] Gráczi, T. E., Krepsz, V., Huszár, A., Száraz, B., Deme, A., Juhász, K. & Markó, A. (kézirat). ‘Analysis of spectral features’ change for speaker comparison’.

[6] Kelly, F., Fröhlich, A., Dellwo, V., Forth, O., Kent, S & Alexander, A (2019). ‘Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)’. *Speech Communication 112.* 30−36. old.

[7] Kuznetsova, A., Brockhoff, P. B. & Christensen R. H. B. (2017). ‘lmerTest Package: Tests in Linear Mixed Effects Models.’ *Journal of Statistical Software, 82(13),* 1−26. old. DOI:10.18637/jss.v082.i13 <https://doi.org/10.18637/jss.v082.i13>.

[8] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

A megakadásjelenségek és a temporális paraméterek szerepe a borderline személyiségzavar felismerésében

Felletár Fanni[1,2], Gosztolya Gábor[3], Hoffmann Ildikó[4,5] és Babarczy Anna[1,2]
[1]Nyelvtudományi Kutatóközpont, Kísérleti Szemantika és Pragmatika Kutatócsoport
{felletar.fanni, babarczy.anna}@nytud.hu
[2]Budapesti Műszaki és Gazdaságtudományi Egyetem, Kognitív Tudományi Tanszék
[3]ELKH-SZTE Mesterséges Intelligencia Kutatócsoport
ggabor@inf.u-szeged.hu
[4]Nyelvtudományi Kutatóközpont, Nyelvi Folyamatok Agyi Reprezentációi Kutatócsoport
hoffmann.ildiko@nytud.hu
[5]Szegedi Tudományegyetem, Pszichiátriai Klinika

Bevezetés

A borderline személyiségzavarral (*borderline personality disorder*; BPD) élő személyeket az identitás, a társas kapcsolatok valamint az érzelem- és indulatszabályozás zavarának tartós mintázata jellemzi [1]. A tünetek számos lehetséges permutációja és a köztük létesülő lehetséges kölcsönhatások heterogén borderline populációt eredményeznek [8, 9, 11], a BPD diagnosztizálása ezért nehéz feladatot ró a klinikusokra. Mivel a beszédben akarva-akaratlanul kifejeződik a beszélő mentális állapota, a mentális zavarok jelei is megjelenhetnek a beszédben [7, 12, 13]. Azt feltételezzük, hogy a BPD és a kontroll csoport (*healthy controls*; HC) jól elkülöníthető a megakadásjelenségek és a temporális paraméterek alapján, az automatikus beszédfelismerés (*automatic speech recognition*; ASR) kényszerített illesztés módszerével. Az általunk bemutatásra kerülő objektív módszer hasznos kiegészítő eleme lehet a már meglévő, nem nyelvészeti szempontú tradicionális diagnosztikai módszereknek.

Módszer

Kényelmi és hólabda mintavétel eredményeként 27 BPD-vel diagnosztizált és 27 HC személy vett részt a kísérletben. A BPD-vel élők esetében a részvétel feltétele a BPD diagnózis megléte volt, a HC személyek esetében pedig bármilyen pszichiátriai vagy neurológiai diagnózis megléte kizáró oknak minősült. A két csoportot nem, életkor, és az elvégzett tanévek száma alapján illesztettük. A résztvevőket arra kértük, hogy meséljék el a tegnapi napjukat. A narratívákról készült hangfelvételekről gépi átiratokat készítettünk, majd kézzel annotáltuk a megakadásjelenségeket [2] alapján. A megakadásjelenségek gyakoriságát gépi módszerrel nyertük ki, az értékeket a szószám alapján arányosítottuk, majd a típusokat magasabb kategóriákba aggregáltuk [2] csoportosítása és [6] beszédprodukciós modelljének tervezési szintjei alapján. A néma szünetekből, a kitöltött szünetekből és a nyújtásokból az [5, 10]-ben mért temporális paramétereket számítottuk ki az ASR kényszerített illesztés módszerével [3, 4].

Eredmények

A hipotézistesztelést egy logisztikus regressziós modellben, "forward" módszerrel végeztük el, melyben az eredményváltozó a csoport volt. Az utolsó modell (R^2 = 0,426; $\chi^2(50) = 8{,}594$; p =

0,003) és a benne szereplő prediktorok – a *nyelvi enkódolási hibák* (grammatikai hibák, kontamináció) aránya (b = 2,330; z = 2,673; p = 0,008), a *kitöltött szünetek számának aránya* (b = 1,073; z = 2,692; p = 0,007) és a *néma szünetek gyakorisága* (b = –1,060; z = –2,649; p = 0,008) – is szignifikánsak voltak. A modell AUC értéke 0,834 lett, és a variancia közel 43%-át magyarázza. Ha 50%-os valószínűségnél húzzuk meg a kísérleti csoporthoz tartozás határát, a modell szenzitivitása 70,37%, vagyis a 27 borderline személyből 19 személyt sorol be megfelelően, specificitása pedig 66,67%, tehát a 27 kontroll személyből 18 személyt kategorizál helyesen, míg 8 (29,63%) borderline és 9 (33,33%) kontroll személyt diagnosztizál tévesen. A BPD csoportból 17 (62,96%) személy rendelkezett legalább egy komorbid zavarral, ez a változó azonban nem volt szignifikáns hatással az osztályozás kimenetelére (φ_c = 0,174; $\chi^2(1)$ = 0,819; p = 0,365). A nyelvi enkódolási hibák aránya, a kitöltött szünetek számának aránya és a néma szünetek gyakorisága tehát segíthet a BPD-vel rendelkező és nem rendelkező személyek elkülönítésében.

Irodalom

[1] American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders (5th ed., text rev.)*. Washington, DC: American Psychiatric Association.

[2] Gósy, M. (2002). ‘A megakadásjelenségek eredete a spontán beszéd tervezési folyamatában’. *Magyar Nyelvőr* 126.2, 192–204. old.

[3] Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., & Hoffmann, I. (2019). ‘Identifying Mild Cognitive Impairment and mild Alzheimer’s disease based on spontaneous speech using ASR and linguistic features’. *COMPUTER SPEECH AND LANGUAGE* 53, 181–197. old. DOI: https://doi.org/10.1016/j.csl.2018.07.007

[4] Gosztolya, G., Balogh, R., Imre, N., Egas-López, J. V., Hoffmann, I., Vincze, V., Tóth, L., Devanand, D. P., Pákáski, M., & Kálmán, J. (2021). ‘Cross-Lingual Detection of Mild Cognitive Impairment Based On Temporal Parameters of Spontaneous Speech’. *Computer, Speech & Language* 69.1, article no. 101215. DOI: https://doi.org/10.1016/j.csl.2021.101215

[5] Hoffmann, I., Tóth, L., Gosztolya, G., Szatlóczki, G., Vincze, V., Kárpáti, E., Pákáski, M., & Kálmán, J. (2017). Beszédfelismerés alapú eljárás az enyhe kognitív zavar automatikus felismerésére spontán beszéd alapján. In Z. Bánréti (Ed.), Kísérletes nyelvészet (pp. 385–405). Budapest: Akadémiai Kiadó.

[6] Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

[7] Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). ‘Automated assessment of psychiatric disorders using speech [A systematic review]’. *Laryngoscope Investigative Otolaryngology* 5.1, 96–116. old. DOI: https://doi.org/10.1002/lio2.354

[8] Millon, T., Grossman, S., Millon, C., Meagher, S., & Ramnath, R. (2004). *Personality Disorders in Modern Life (2nd ed.)*. Hoboken, NJ: Wiley.

[9] Smits, M. L., Feenstra, D.J., Bales, D. L., de Vos, J., Lucas, Z., Verheul, R., & Luyten, P. (2017). 'Subtypes of borderline personality disorder patients [A cluster-analytic approach]'. *Borderline Personal Disord Emot Dysregul* 4.16. DOI: https://doi.org/10.1186/s40479-017-0066-4

[10] Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., Zsura, F., Pákáski, M., & Kálmán, J. (2015). 'Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech using ASR'. *Proceedings of Interspeech*. Dresden, Germany. 2694–2698. old. DOI: http://dx.doi.org/10.21437/Interspeech.2015-568

[11] Unoka, Z. S., Richman, M. J., & Czégel, D. (2018, April 7). 'How many causal pathways must symptoms form before we call them a borderline? [A hierarchical network model of borderline personality disorder]'. DOI: https://doi.org/10.31234/osf.io/93cnh

[12] Wang, B., Wu, Y., Taylor, N., Lyons, T., Liakata, M., Nevado-Holgado, A. J., & Saunders, K. E. A. (2020). 'Learning to Detect Bipolar Disorder and Borderline Personality Disorder with Language and Speech in Non-Clinical Interviews'. *INTERSPEECH 2020*, Shanghai, China. 437–441. old. DOI: https://doi.org/10.48550/arXiv.2008.03408

[13] Wang, B., Wu, Y., Vaci, N., Liakata, M., Lyons, T., & Saunders, K. E. A. (2021). 'Modelling Paralinguistic Properties in Conversational Speech to Detect Bipolar Disorder and Borderline Personality Disorder'. *CASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7243-7247. old. DOI: https://doi.org/10.48550/arXiv.2102.09607

Relációs adatbázis fejlesztése beszélt nyelvi adatbázisok metaadatainak kezeléséhez LibreOffice Base rendszerben

Gocsál Ákos
Pécsi Tudományegyetem, Művészeti Kar, Zeneművészeti Intézet, Pécs
Nyelvtudományi Kutatóközpont, Budapest

A beszélt nyelvi adatbázisok fejlesztése Drude és munkatársai szerint három fő adattípus rögzítésén alapul. Az elsődleges adatok maguk a hang- és képfelvételek, a másodlagos adatok az annotációk, a harmadik adattípust pedig a metaadatok alkotják, amelyek a beszédesemény, illetve az elsődleges és másodlagos adatok időfüggetlen tulajdonságai. A metaadatok azért fontosak, mert segítségükkel lehet a konkrét kutatási céloknak megfelelő beszédfelvételeket kigyűjteni az adatbázisból, illetve csoportosítani azokat a kutató szándékainak megfelelően [4].

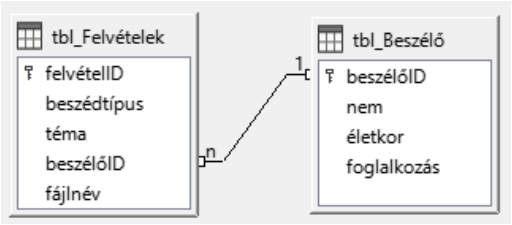
A metaadatok változatos típusúak lehetnek. Dobrushina és Sokur szláv nyelvek beszélt nyelvi adatbázisai kapcsán említi, hogy metaadatként tárolhatók a beszélő tulajdonságai (neme, életkora, iskolai végzettsége, anyanyelve), de a lejegyzést végző kutató neve, a hangfájl neve, a lejegyzés verziója, a kommunikációs helyzettel kapcsolatos információk, a kommunikációs csatorna, az esemény típusa, az esemény helye és ideje is [3]. Más kutatók televíziós műsorban, filmben elhangzó beszéd kutatásához készítettek adatbázist. Ezekben az esetekben metaadat volt a bemutatás időpontja, a gyártó, a film műfaja vagy a felvétel helyszíne (pl. [2, 6]).

A metaadatok tárolására és a rekordok szűrésére, lekérdezésére a legegyszerűbb megoldásnak egy Excel-tábla kínálkozik, azonban ennek hátrányai a bővítés során hamar megnyilvánulnak. A táblázat áttekinthetetlenül naggyá válhat, ugyanakkor sokszor nem kerülhető el egyes adatok ismételt beírása, például ha ugyanazon beszélőtől több bemondást is rögzít a kutató, és mindegyik bemondásnál szüksége van a beszélő demográfiai adataira. A problémára a relációs adatbázisok jelentenek megoldást. Ezek több kisebb méretű táblát használnak, amelyek között kapcsolatok alakíthatók ki, így kiküszöbölhető a redundancia. Ennek alapelve az, hogy az egyik táblában egyszer rögzített rekordra – például a beszélő adataira – a rekordhoz tartozó egyedi index, „kulcs” segítségével hivatkozunk az adatbázis más tábláiban, így az adatok az egyszeri rögzítés ellenére tetszőleges számban, tetszőleges helyen megjeleníthetők, használhatók. A korpusznyelvészeti kutatásokban (pl. [1]) gyakran alkalmaznak ilyen relációs adatbázisokat, azonban ezek általában jelentős informatikai tudást igénylő egyedi megoldások.

Az alábbiakban az ingyenesen és több platformra is elérhető LibreOffice rendszer Base programjában (https://hu.libreoffice.org/) készített adatbázis-struktúrákat mutatunk be, amelyek segítségével könnyen kiválaszthatók az adott kutatáshoz szükséges hangfájlok. A Base előnye, hogy használatával programozói tudás nélkül is hatékonyan működő relációs adatbázis építhető.

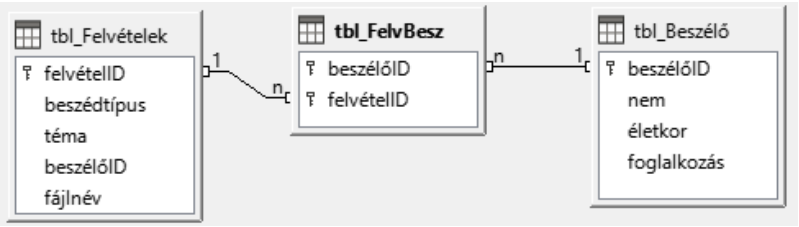
Az első, egyszerűbb példában azt feltételezzük, hogy több beszélőtől több beszédtípus is rendelkezésre áll, és ugyanazon beszélő több alkalommal is szerepelt. A jellemző kutatói igény ilyenkor az, hogy kigyűjtse az azonos beszédtípusokhoz tartozó felvételeket, majd azokat a beszélő neme, életkora szerint szétválogassa. Az adatokat két táblában rögzítjük. A tbl_Felvételek rekordjai a következők: felvételID, beszédtípus, téma, beszélőID, fájlnév. A felvételID a tbl_Felvételek elsődleges kulcsa, azaz az egyes felvételekhez tartozó egyedi

azonosító. A tbl_Beszélő rekordjai az alábbiak: beszélőID, nem, életkor, foglalkozás. Látható, hogy a beszélőID mindkét helyen szerepel, a tbl_Beszélő táblában elsődleges, a tbl_Felvételek táblában idegen kulcs, azaz ez a mező hozza létre a relációt a két tábla között (1. ábra).



1. ábra. A felvételek és a beszélők adattáblái közötti reláció

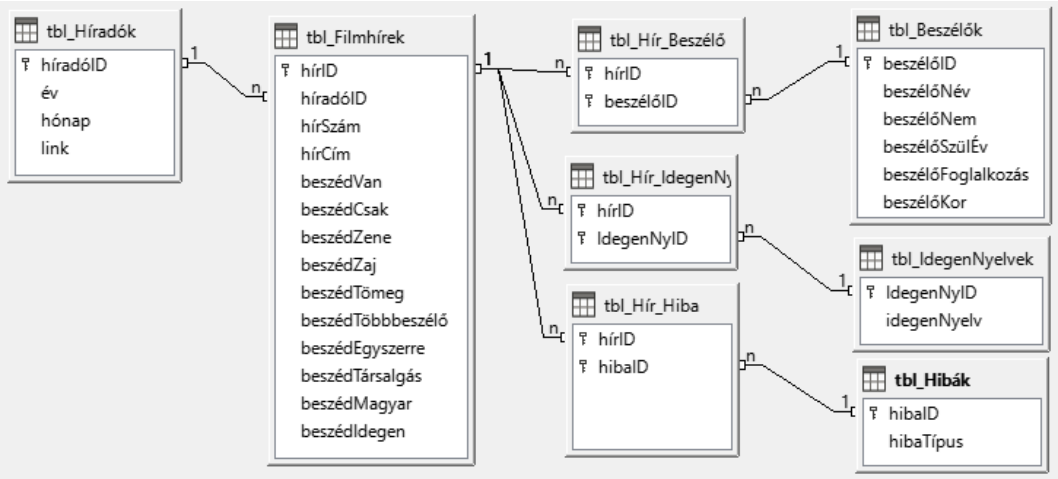
A reláció szükségtelenné teszi, hogy a beszélők adatait minden felvételhez beírjuk, ugyanakkor a lekérdezések során azok mindenhol megjelennek, ahol a beszélő megjelenítését beállítjuk. Ez a megoldás ugyanazon beszélőt több felvételhez tudja rendelni, de egy felvételt csak egy beszélővel kapcsol össze. Interjúk, társalgások esetében azonban szükség van arra, hogy egy felvételhez több beszélő is tartozhasson. Ennek megvalósításához egy összekapcsoló táblát alkalmazunk, amelynek két mezője a két másik táblázat elsődleges kulcsait tartalmazza, a rekordok pedig a két ID párosításai lesznek (azaz bármely beszélőhöz tetszőleges számban bármely felvétel, illetve bármely felvételhez tetszőleges számban bármely beszélő társítható).



2. ábra. A felvételek és a beszélők adattáblái összekapcsoló táblával

A 3. ábra egy fejlesztés alatt álló, a fentieknél összetettebb adatbázis felépítését mutatja 2023. januári állapotában, amely filmhíradók kutatását segíti [5]. Minden egyes híradó több filmhírből áll. Az egyes filmhírekben több beszélő, több nyelv, illetve többféle technikai hiba (szakadás, torzulás) is előfordulhat. Az ábrán látható, hogy a híradók, filmhírek, beszélők, nyelvek, hibák adatai külön táblákban helyezkednek el. A 2. ábrához hasonló módon összekapcsoló táblák teszik lehetővé a filmhírekkel való többszörös relációkat. A filmhíreket tartalmazó táblában számos boolean típusú, kétértékű változó szerepel, amelyek azt kódolják, hogy az adott filmhír tartalmaz-e a háttérzajt, tömeges beszédet, egyszerre beszéléseket, társalgást stb., így gyorsan kiválaszthatók e szempontok szerint is a kívánt hírek. Az adatbázisnak nem csak a könnyű lekérdezhetőség és a redundancia kiküszöbölése az előnye, hanem az is, hogy a kisebb táblák segítségével könnyen adhatók hozzá további beszélők, nyelvek, előre nem látott más jellemzők.

Dobrushina és Sokur szerint a metaadatok rögzítése bizonyos szempontból még az annotálásnál is fontosabb [3]. Bár a beszélt nyelvi adatok önmagukban is értékesek, a metaadatok hiánya jelentősen szűkíti a kutatási lehetőségeket. A szoftver alapfunkcióinak elsajátítását követően az itt bemutatott megoldások alapján bárki könnyen létrehozhat saját kutatói igényeinek megfelelő relációs adatbázist.



3. ábra. A filmHír relációs adatbázis felépítése

Irodalom

[1] Abbot, R., Ecker, B., Anand, P., & Walker, M. (2016). Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 4445–4452). Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1704 (2022. október 17.)

[2] Carrive, J., Beloued, A., Goetschel, P., Heiden, S., Laurent, A., Lisena, P., … Troncy, R. (2021). Transdisciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project. *DHQ: Digital Humanities Quarterly*, *15*(1). http://digitalhumanities.org/dhq/vol/15/1/000523/000523.html (2022. október 17.)

[3] Dobrushina, N., & Sokur, E. (2022). Spoken Corpora of Slavic Languages. *Russian Linguistics*, *46*(2), 77–93. https://doi.org/10.1007/s11185-022-09254-9 (2022. október 17.)

[4] Drude, S., Trilsbeek, P., Sloetjes, H., & Broeder, D. (2014). Best practices in the creation, archiving and dissemination of speech corpora at the language archive. In S. Ruhi, M. Haugh, T. Schmidt, & K. Wörner (eds.), *Best Practices for Spoken Corpora in Linguistic Research* (p. 183–207). Newcastle upon Tyne: Cambridge Scholars Publishing. https://www.mpi.nl/publications/item2051339/best-practices-creation-archiving-and-dissemination-speech-corpora (2022. október 17.)

[5] Gocsál, Á. (2022). Beszédtípusok és artikulációs tempóik a korai hangos filmhíradókban. In Csárdás L. & Bóna J. (szerk.), *Sokszínű beszédtudomány*. Budapest: Akadémiai Kiadó.

[6] Lleida, E., Ortega, A., Miguel, A., Bazán-Gil, V., Pérez, C., Gómez, M., & de Prada, A. (2020). *RTVE2020 Database Description*. http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf (2022. október 17.)

A kutatást a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal NKFIH-FK-128814 számú pályázata támogatta.

DYNAMO - reloaded

Reinhold Greisbach
Institute for Linguistics-Phonetics, University of Cologne, Germany

In 1976 Georg Heike, at that time head of the Institute for Phonetics at the University of Cologne, got the opportunity to acquire a mini-computer PDP 11/10 for phonetic research purposes. This machine was equipped with a special Tektronix screen as monitor, so that graphical data could be displayed. One of the first projects using this equipment was the development of a midsagittal 2D-computer model of the vocal tract. It was programmed in FORTRAN and controlled by phonetically defined parameters to produce midsagittal views of the articulators for each sound which was otherwise only available in sound atlases (e.g. [5, 6]).

The geometries of the sounds were generated via linear interpolation of the geometries of a few extremal articulatory configurations. This model was intended as an aid for the deaf [2] and later as an aid in the area of speech therapy. Several BA- and MA-theses revealed the useful therapeutical effect of the model in speech therapy.

In 1986 the midsagittal model was rewritten in the programming language C and was implemented on an ATARI microcomputer. The new model was called DYNAMO (DYNamic Articulation MOdel) and presented at the Hannover industrial fair in the spring of 1987 to the general public. Customers (mostly schools for the speech disabled or deaf students) had to own an ATARI computer to be able to use the model. The early ‘90s brought the dawning of the age of Windows. But the model was not ported into the Windows world and the development of DYNAMO ended with the retirement of Georg Heike in 1998.

Of course, DYNAMO was not the first articulatory model at that time and in the meantime a multitude of additional articulatory models were developed all over the world. [3] gives a commented overview (control parameters, purpose, etc.) of several of these models developed in the last 40 years. Today we find even 3D-models of the vocal tract. Models of today tend to be more realistic, are often based on MRI films of real speech and are used in speech research to investigate various topics. The modelling of the physiological control mechanisms is one goal of the research. Another one is the production of high-quality synthetic speech. Maximal realism of modelling the articulatory dynamics seems to be necessary for this goal.

At the same time, we still face the “classical” situation in the teaching of phonetics as 40 years ago. Articulatory Phonetics is taught primarily via phonetic symbols (e.g. the alphabet of the IPA). In the course of the teaching process one sometimes gets the impression, that many students believe that the symbols are identical with sounds and not mere abbreviations of specific articulatory configurations of the vocal tract and their corresponding acoustic output. They also seem to believe that diacritics are used to denote special (hence not so important) variants of the basic symbols but not that they were introduced for the sake of reducing the number of necessary symbols to describe sounds (e.g. voiceless nasals).

From the articulatory (geometrical) point of view there is a continuum between articulation manner and place of articulation of the consonants, and even between consonants and vowels.

The basic symbols denote fixed points in this continuum which are used frequently by various languages as their speech sounds. Focussing only on these symbols (i.e. on these reference points) and – what is more – restricting the focus of phonetic teaching to the sounds of the mother tongue of the students might lead to misdescriptions of speech errors by speech therapists as [1] found out.

Using geometrical displays of the articulations could enhance phonetic teaching and articulatory models are more beneficial in this respect than sound atlases because of their flexibility. But there are almost no articulatory models designed especially for the teaching of phonetics. A rare exception is [4]. However this articulatory model is conceptually designed primarily on the basis of the sounds of German and contains a division between consonants and vowels. For educational purposes an articulatory model should fulfil some requirements. It should

- allow a static view (not movies) of the sound geometries, so that the configurations may be learned and memorized by the students of phonetics,
- be a simple 2D-model (instead of 3D) containing only the outline of the vocal tract, so that the configurations can be redrawn on paper by the students by heart,
- be able to generate configurations which are not language specific and even articulations that may be found in disordered speech (for the education of speech therapists),
- contain invisible features, which are necessary for the production and classification of the sounds (e.g. in case of fricatives: location of the friction source in the vocal tract),
- display additional geometrical features beyond a mere midsagittal view (which can be found in some sound atlases) such as the synchronized frontal view and/or a display of the palatogram,
- be controlled by not too many parameters, which should correspond to or correlate with the traditional articulatory classification parameters of vowels and consonants.

The classical DYNAMO model fulfils several of these requirements. With the emergence of the HTML5 canvas graphics in all of the modern internet browsers (Firefox, Chrome, Safari etc.) it was decided to wake up DYNAMO again, now written in JavaScript. Based on the original geometry data of DYNAMO an additional synchronized 2D frontal view of the lips and the jaw was added. It is now used in beginner courses in Phonetics and Speech Therapy at the University of Cologne and the IB-school for Speech Therapy in Ulm.

Contrary to the old DYNAMO model, which only could display the geometry of the standard German sound, the new model is more flexible and can produce almost every possible articulation. The model is used in courses for Phonetic Transcription to demonstrate the articulatory configuration of the symbols of the IPA and mostly to the surprize of the students even beyond (e.g. simultaneous [p]+[t]+[k] (see figure 1), nasalized glottal stops, etc.).

The principles of the design and control of the model will be explained and a demonstration of the actual version (Version 2.0) will be given in the talk. This model is freely available as internet resource (https://www.logodynamo.de) and can be run with every modern browser with an internet connection.

Figure 1: Model display for simultaneous [p] + [t] + [k].

References

[1] Ball, M.J.; Müller, N.; Rutter, B.; Klopfenstein, M. (2009). ‚My client is using non-English sounds! A tutorial in advanced phonetic transcription. Part 1: Consonants". In: *Contemporary Issues in Communication Science and Disorders* 36, pp. 133-141.

[2] Heike, G. (1977). ‚Computer generation of a parameter controlled articulatory model as an aid for the deaf. In: *IPKöln-Berichte* 7, pp. 63-70.

[3] Kröger, B.J. (2022). 'Computer-implemented articulatory models for speech production: A review. In: *Front. Robot. AI* 9:796739. doi: 10.3389/frobt.2022.796739

[4] Kröger, B.J. & Kröger, B. (2022). SpeechTrainer: Visualization of speech movements [Computer program]. https://speechtrainer.eu/spetra/

[5] Wängler, H.-H. (1958). *Atlas Deutscher Sprachlaute*: Berlin, Akademie Verlag.

[6] Molnár, J. (1970). *A Magyar beszédhangok atlasza*: Budapest, Tankönyvkiadó.

Erdélyi magyarok akcentusa angolul beszélve: Egy laboratóriumi fonológiai vizsgálat első eredményei

Huszthy Bálint
ELTE Eötvös Loránd Tudományegyetem

A *hunglish* magyar anyanyelvűek internyelve angolul beszélve (vö. [1, 6, 8]). Ez a jellegzetes idegen akcentus számos fonetikai, fonológiai (és más nyelvi és nyelven kívüli) összetevőből jön létre (vö. [2, 4, 5]), melyek közül az egyik legfontosabb a magyar fonológia angollal való interferenciája. Az angolul tanuló magyarok ugyanis különböző transzferjelenségek útján átültetik anyanyelvük produktív fonológiai tulajdonságait a célnyelv kiejtésére [3, 8]. A Romániában (főleg Erdélyben) élő magyaroknak azonban jellemzően más összetételű az internyelve angolul beszélve, mint a magyarországiaknak, aminek két legfontosabb oka a román nyelv napi szintű használata és a beszélők erős dialektális háttere [3, 7].

Ez a vizsgálat 30 erdélyi magyar adatközlő angol kiejtését hasonlítja 15 magyarországi adatközlőéhez [3]. A beszélőknek (akik egyetemi hallgatók a kolozsvári BBTE-n és a budapesti PPKE-n) ugyanazt az angol újságcikket kellett kétszer felolvasniuk. A felgyűjtött hanganyag számos fonetikai és fonológiai vizsgálatra alkalmas, itt a legfontosabb általános következtetéseket vonom le, majd két magánhangzó transzferét részletesen is megvizsgálom.

A két adatközlőcsoport kiejtése sok tekintetben megegyezést mutat: így magán- és mássalhangzó-helyettesítések terén, fonotaktikai és laringális fonológiai folyamatokban, és prozódiai összetevőkben is. Mégis megfigyelhetők visszatérő, rendszerszerű különbségek. Az angol hátsó-alsó magánhangzókat [ʌ, ɒ, ɔ, ɑ] az erdélyi csoport (továbbiakban E) többnyire magyar [ɒ]-val helyettesíti, míg a magyarországi csoport (továbbiakban M) leginkább magyar [a]-val, pl. ang. *some come just once* [sʌm ˈkʰʌm ʤʌst ˈwʌns] > E [ˈsɒm ˈkɒm ʤɒst ˈwɒns], M [ˈsam ˈkam ˈʤast ˈvans]. Az angol svák az M-ben főleg egy [ø]-szerű internyelvi hanggal helyettesítődnek, ám az E megőrzi a svákat, hisz az [ə] fonéma a román hatására megtalálható a beszélők hangkészletében, pl.: ang. *summer* [ˈsʌmə(ɹ)] > E [ˈsɒm(ː)ər], M [ˈsam(ː)ør]; ang. *famous* [ˈfeɪməs] > E [ˈfejməs], M [ˈfeːmøs]; stb. Az E megőrzi az angol diftongusok javát (a román és a nyelvjárási diftongusok hatására), az M ellenben erősen monoftongizál, pl.: ang. *home* [həʊm] > E [howm], M [hoːm]; ang. *real* [ɹɪəɬ] > E [riəl], M [riːl]; stb.

A mássalhangzó-helyettesítések közül a legfontosabb különbség az angol [w]-t érinti, ezt ugyanis az E többnyire megőrzi, míg az M [v]-re cseréli, pl.: ang. *away* [əˈweɪ] > E [əˈwej], M [øˈvej] stb. Sok mássalhangzós jelenség azonban megegyezik a két csoport között, így például az interdentálisok [θ, ð] homorgán magyar hanggal való helyettesítése (pl. [t, d, s, z]), a zöngétlen zárhangok aspirációjának hiánya, az írásban kettőzött mássalhangzók gyakori hosszan ejtése, szillabikus zengőhangok és a sötét [ɬ] hiánya, valamint a zöngésségi hasonulás rendszeres alkalmazása, pl.: ang. *make the* [kθ̬] > [gd], *stop the* [pθ̬] > [bd], *voice that* [sθ̬] > H. [zd], *yet that* [tθ̬] > [dː] stb.

Prozodikus különbségek is akadnak a két csoport eredményei között, például a szóhangsúly helyzetét az M jelentősen többször téveszti el, mint az E (utóbbi számára ismét a román nyújthat segítséget a mozgó szóhangsúllyal), pl.: az ang. *professional, attraction, important* szavakban csak az M alkalmazott első szótagi hagsúlyt bizonyos százalékban (a pontos számadatok [3]-ban olvashatók), míg az E következetesen a második szótagot hangsúlyozta, az angolnak megfelelően. Ellenben az erdélyiek akcentusát is tudja befolyásolni a magyar prozódia,

amennyiben például az ang. *most famous* [məʊst ˈfeɪməs] kifejezést minden adatközlő az első szón hangsúlyozta: E [ˈmowst fejməs], M [ˈmoːst feːməs].

A felvett hanganyagból kiválogattam 5 háromszéki és 5 budapesti 21 éves lány kétszeri kiejtését, hogy az angol [ʌ] és a svá megfeleltetéseit részletesebben is megvizsgálhassam akusztikailag. 6 [ʌ]-t és 6 svát tartalmazó célszó esetében (összesen 60–60 előfordulás) mértem meg a hunglish-ban megjelenő magánhangzók első három formánsának Hertz-értékeit Praatban a magánhangzók közepébe kattintva, az eredményeket R-ben összesítettem (1. és 2. ábra).

Az 1. ábrán látható dobozdiagram az angol [ʌ] magánhangzó hunglish-beli megfeleléseit mutatja (célszavak: *summer*, *hundreds*, *some*, *come*, *once*, *others*). A dobozpárok közül az első doboz az M, a második az E középértékeit ábrázolja, az első három formánsra vonatkozóan. Az F1 alapján (M-átlag: 794 Hz, E-átlag: 667 Hz) az E-ben alacsonyabb nyelvállással valósul meg a magánhangzó, míg az F2 alapján hátsóbb nyelvállással (M-átlag: 1600 Hz, E-átlag: 1321 Hz), az F3 pedig inkább utal ajakkerekítésre az E-ben, mint az M-ben (M-átlag: 2658 Hz, E-átlag: 2480 Hz). Eszerint az angol [ʌ] az E-ben egy magyar veláris [ɒ]-hoz hasonló internyelvi hangként valósul meg, míg az M-ben az alsó-centrális kerekítetlen magyar [a]-ra hasonlít.

A 2. ábra az angol [ə] hunglish-ban mért középértékeit mutatja (célszavak: *summer*, *world*, *major*, *river*, *famous*, *away*). A magánhangzó az E-ben alapvetően ugyancsak sváként valósul meg (az átlagok: F1 529 Hz, F2 1573 Hz, F3 2561 Hz). Az M-ben azonban inkább magyar [ø]-re hasonlít a hunglish-ban létrejövő internyelvi hang (az átlagok: F1 470 Hz, F2 1622 Hz, F3 2339 Hz), amire az alacsonyabb F1 és F3, valamint a magasabb F2 értékek utalnak.

Összességében elmondható, hogy az erdélyi magyarok angol kiejtése autentikusabb a magyarországiakéhoz képest, ami elsősorban az internyelvük heterogeneitásának lehet köszönhető, hiszen minden adatközlő napi szinten használja a román nyelvet, valamint a magyar regionális köznyelv mellett mindannyian dialektust is beszélnek.

A kutatás a #142498-as számú OTKA-projekten belül valósult meg.

Irodalom

[1] Balogné Bérces K. & Szentgyörgyi Sz. (2006). *The Pronunciation of English*. Budapest: Bölcsész Konzorcium.

[2] Flege, J. E. (1987). ‘The production of “new” and “similar” phones in a foreign language: Evidence for the effect of equivalence classification.’ *Journal of Phonetics* 15, 47–65. old.

[3] Huszthy B. (2022). ‘Transylvanian Hunglish: Phonological properties of Hungarian accented English in Transylvania.’ *Hungarian Studies Yearbook* 4, 131–159. old.

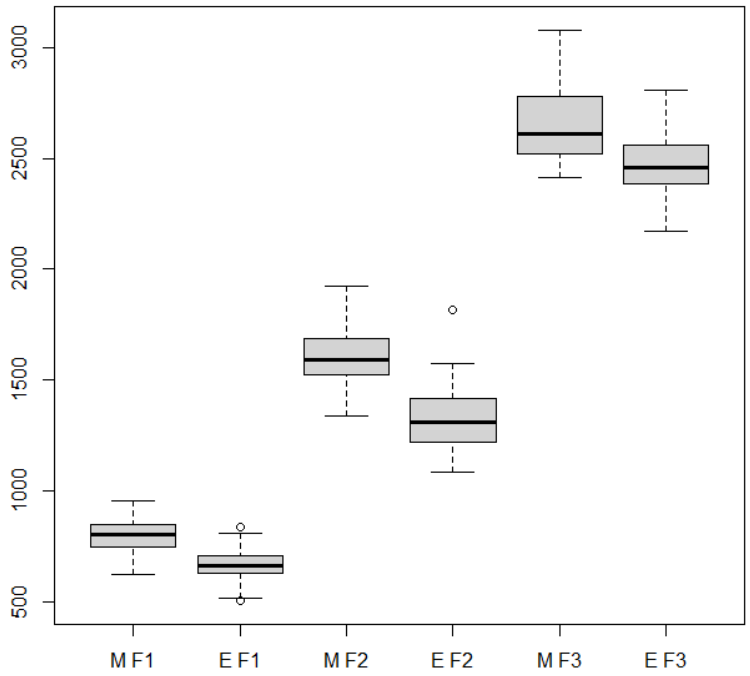
[4] Major, R. C. (2001). *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology*. Mahwah, NJ: Lawrence Erlbaum.

[5] Moyer, A. (2013). *Foreign Accent: The Phenomenon of Non-native Speech*. Cambridge: Cambridge University Press.

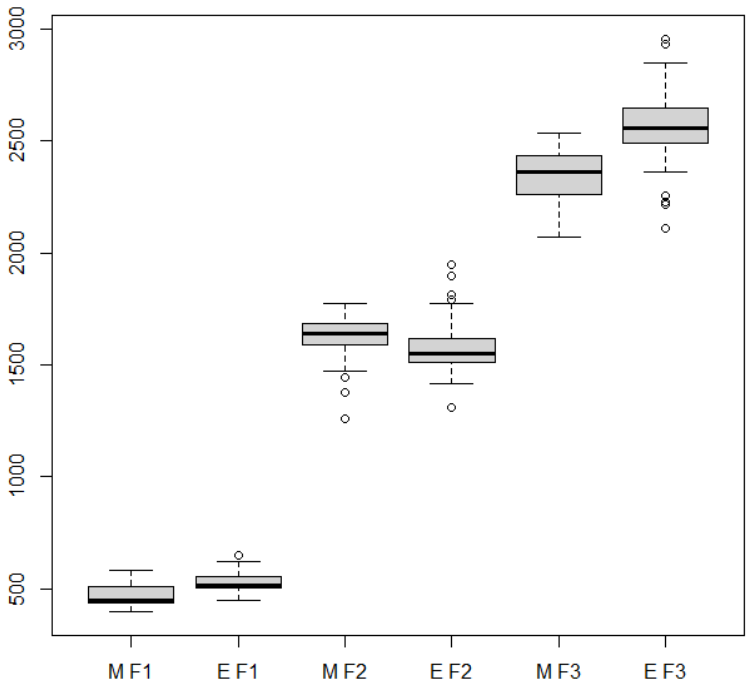
[6] Nádasdy Á. (2006). *Background to English Pronunciation*. Budapest: Nemzeti Tankönyvkiadó.

[7] Péntek J. & Benő A. (2020). *A magyar nyelv Romániában (Erdélyben)*. Kolozsvár, Budapest: Erdélyi Múzeum Egyesület, Gondolat Kiadó.

[8] Piukovics Á. (2021). *Phonological and non-phonological factors in non-native pronunciation acquisition*. PhD-disszertáció. Budapest: PPKE BTK.



1. ábra. Az angol [ʌ] magánhangzó magyarországi és erdélyi megfeleltetései



2. ábra. Az angol [ə] magánhangzó magyarországi és erdélyi megfeleltetései

Data Augmentation for Articulatory-to-Acoustic Mapping using Ultrasound Tongue Imaging

Ibrahim Ibrahimov[1], Gábor Gosztolya[2], and Tamás Gábor Csapó[3]

[1] Institute of Informatics, University of Szeged
[2] ELKH-SZTE Research Group on Artificial Intelligence
[3] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics

Introduction

The Silent Speech Interface (SSI) is a technology that has the potential to revolutionize the way individuals communicate. It enables individuals who are unable to produce audible speech to communicate effectively by predicting acoustic speech from their inaudible articulation. This is achieved through a process known as articulatory-to-acoustic mapping, which maps the movements of the mouth and tongue to the corresponding speech sounds. There are several methods of collecting the inputs needed to create an SSI system, including ultrasound tongue imaging (UTI), electromagnetic articulography (EMA), etc. Of these methods, UTI has become particularly popular in SSI research and development due to its relatively high spatial and temporal resolution, portability, and cost-effectiveness. One of the challenges faced in the development of SSI systems is the limited amount of training data available. To overcome this challenge, researchers often resort to a technique called data augmentation, which involves increasing the size of the training data set through various transformations and modifications. This helps to ensure that the system is better equipped to make accurate predictions even when faced with new and unseen data.

Methods

In articulatory data collection, one Azerbaijani male subject (the first author of this study) was recorded while reading Azerbaijani sentences aloud (154 sentences). In [3], more details about the recording set-up and articulatory data can be found.The total duration of the recordings was about 15 minutes, which was partitioned into training, validation and test sets in a 80-10-10 ratio.

In the baseline vocoder, similarly to the original WaveGlow paper [4], 80 bins were used for mel-spectrogram using librosa mel-filter defaults (i.e. each bin is normalized by the filter length and the scale is the same as in HTK) (Fig. 1a, 1d). The main purpose of data augmentation is to add new synthetic data which is created based on existing dataset to increase the number of examples which are given to the network as input. By using data augmentation techniques, we tried to improve the robustness of deep neural network, which targets to solve overfitting and generalization problem.
Consecutive time masking (CTM) – has been used to nullify some consecutive part of pixels in the ultrasound tongue image. This method aims to recreate new sample by masking a specific part of original data. In this study, we chose [0:100] range based on our initial experiments to be able to visualize changes and get the best results.
Intermittent time masking (ITM) – is the other way of masking pixels out on UTI which is focusing on different ranges rather than one consecutive part (Fig. 1b , 1e). This way we can achieve to our goal of creating synthetic data by making several parts of pixels zero (specifically $[\frac{1}{6} : \frac{1}{3}], [(\frac{1}{3} + \frac{1}{6}) : (2 \cdot \frac{1}{3})], [(2 \cdot \frac{1}{3} + \frac{1}{6}) : 1]$).

Sinusoidal noise injection (SNI) – has been used in this study to add predefined noise to our tongue images. Core idea behind this technique is to create sinusoidal wave by using properties of data itself and injecting created sinusoidal noise into each pixel line of UTI. We calculated the mean of each pixel line and by the given formula 1 this obtained sinusoidal wave for each pixel line.

$$I(t) = 0.02 \cdot A \cdot sin(2\pi \cdot 40t) \tag{1}$$

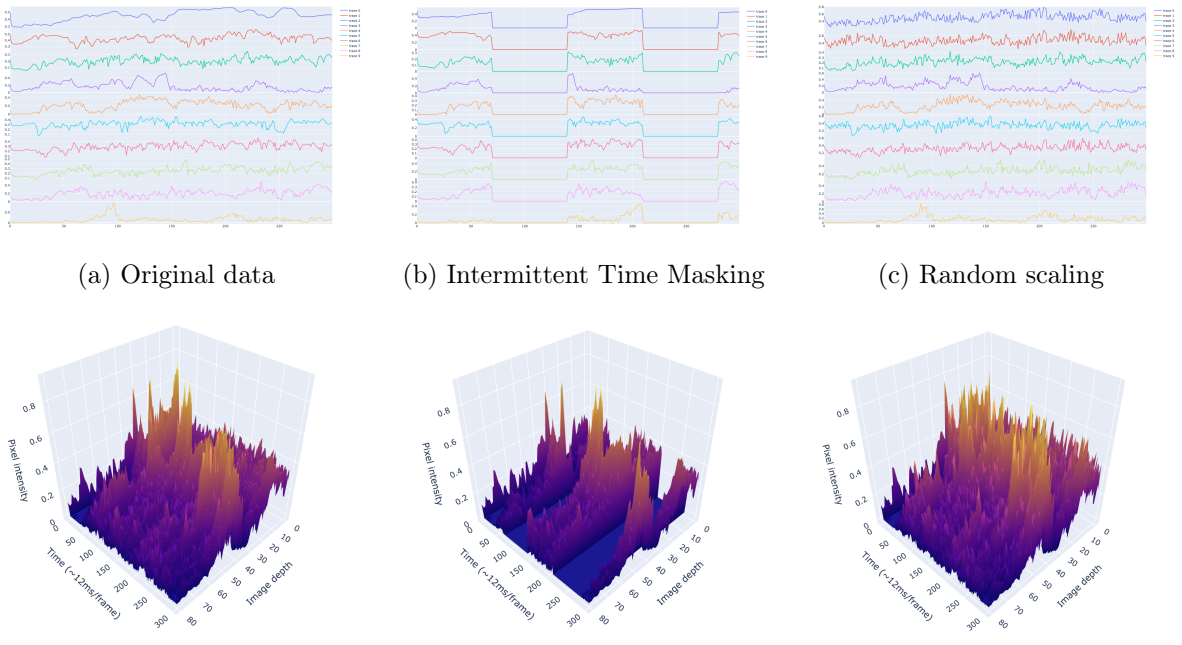
Random scaling (RS) – is used in different studies showed to be helpful to expand existing dataset by using random scalar on data. In this paper, we experimented different range to demonstrate the best results and among all [0.8-1.4] gave the least error rates. Here based on the scalar, pixels of our ultrasound images can get shrunk if the value of scalar is less than 1 or can be stretched in case the scalar is larger than 1. At the end, we obtain randomly scaled synthetic data which is helpful to train neural network better based on larger dataset that we provide (Fig. 1c, 1f).

Results and discussion

Based on the results shown in the bar chart (Fig. 2), it is clear that the data augmentation techniques we applied had a slight impact on the performance of our model. On our dataset, we used mean squared error as an evaluation metric to compare each augmentation method. The methods are compared using two errors that are tested in the training and validation datasets. According to both error rates, applying random scaling to the dataset could help us achieve minimizations on error rates (0.33 MSE and 0.47 V-MSE). We plan to apply these methods to a larger dataset or on other modalities (e.g., vocal tract MRI [2]), as well as observe these method combinations separately on a sample. Further experiments and studies are needed to verify these results.

Acknowledgements

We would like to thank Beiming Cao and his colleagues at the University of Texas at Austin, USA for the excellent research idea [1]. The research was partially funded by the NRDI Office (FK 142163). T.G. Cs.'s research was supported by the Bolyai Research Fellowship of the HAS, and by the ÚNKP-22-5-BME-316 New National Excellence Program of the Ministry for Culture and Innovation from the source of the NRDIF. G. G.'s research was supported by the NRDI Office of the Hungarian Ministry of Innovation and Technology (grant no. TKP2021-NVA-09), and within the framework of the Artificial Intelligence National Laboratory Program (RRF-2.3.1-21-2022-00004). The Titan X GPU used was donated by NVIDIA.



(a) Original data (b) Intermittent Time Masking (c) Random scaling

(d) Original 3D figure (e) ITM 3D figure (f) RS 3D figure

Figure 1: Examples of UTI before and after data augmentation methods.

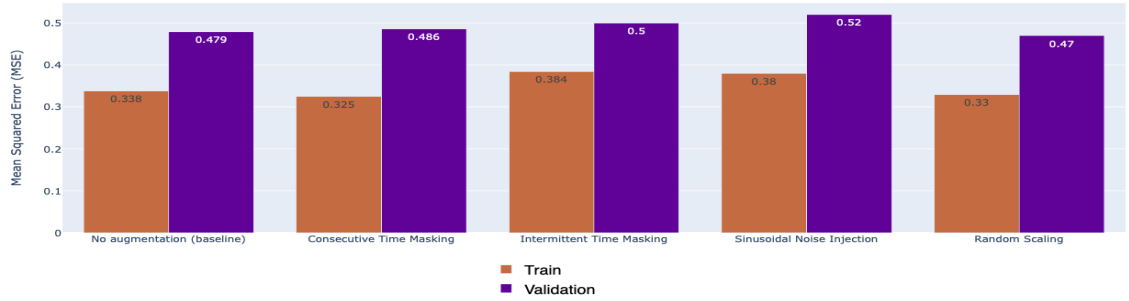


Figure 2: Mean Squared and Validation Mean Squared Errors of UTI example sample before and after data augmentation methods.

References

[1] Cao, B. et al. (2022). 'Data Augmentation for End-to-end Silent Speech Recognition for Laryngectomees'. In: *Proceedings of the Interspeech*, pp. 3653–3657.

[2] Csapó, T. G. (2020). 'Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal tract'. In: *Proc. Interspeech*. Shanghai, China, pp. 2722–2726.

[3] Csapó, T. G. et al. (2017). 'DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface'. In: *Proc. Interspeech 2017*, pp. 3672–3676.

[4] Prenger, R., R. Valle & B. Catanzaro (2019). 'Waveglow: A flow-based generative network for speech synthesis'. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3617–3621.

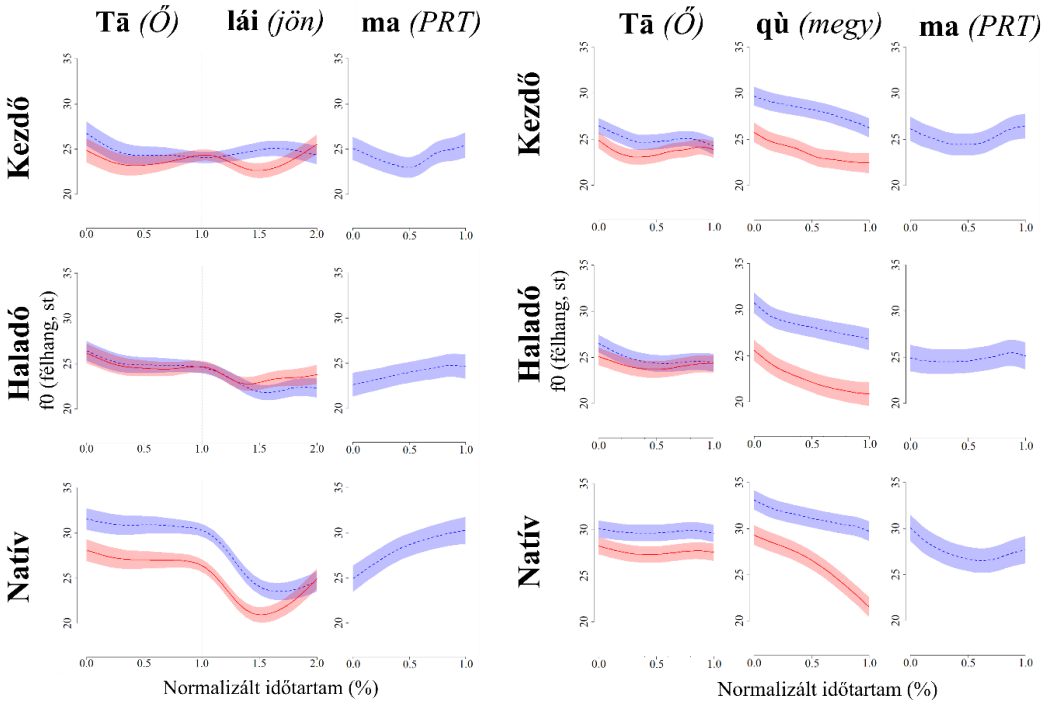
Az intonáció és a tónusok egymásra hatása kínaiul tanuló magyar anyanyelvűeknél

Juhász Kornélia, Bartos Huba
Eötvös Loránd Tudományegyetem, Nyelvtudományi Kutatóközpont

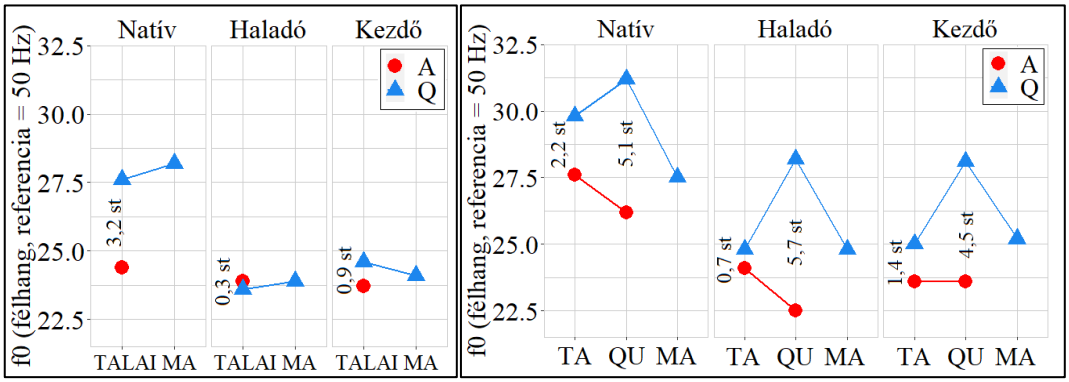
Az akusztikai kísérlet középpontjában a tónus- és intonációprodukció interakciója áll szintaktikailag jelölt mandarin kínai kérdésekben és kijelentésekben a kínaiul tanuló magyar anyanyelvűek produkciójában. A mandarin kínai tonális nyelv, ezért az alapfrekvencia (f_0) változása a szótagok szintjén jelentésmegkülönböztető szereppel bír, azonban az egyes tónusok megvalósulását (dallamkontúrját, f_0-regiszterét és f_0-terjedelmét) a közlés modalitása, illetve szegmentális és tónus-koartikulációs hatások is befolyásolják [2,4,5]. A mandarinban a kijelentő dallamot – a legtöbb nyelvhez hasonlóan – ereszkedő dallam jellemzi. Az eldöntendő kérdések teljes f_0-kontúrja a kijelentő dallamhoz képest magasabb frekvenciatartományban jelenik meg. A szintaktikailag jelölt kínai kérdések esetében a közlés végén elhelyezkedő tónustalan *ma* kérdő partikula a megelőző tónus dallamának (céljának) kiterjesztéseként valósul meg [5]. Ez azt jelenti, hogy a kínai emelkedő 2. tónust (LH) követő *ma* partikula relatíve magas (H); míg az ereszkedő 4. tónust (HL) követő partikula pedig az őt megelőző tónushoz viszonyítva alacsony (L) frekvenciatartományban valósul meg [5]. Ezzel szemben az atonális magyarban nem csak a tónusok hiánya, hanem az eltérő intonációs mintázatok is megnehezíthetik a kínai nyelv elsajátítását: a magyar kijelentések ereszkedő dallamával szemben a magyar eldöntendő kérdések a kínaitól eltérő dallammal valósulnak meg. A magyar eldöntendő kérdéseket emelkedő-eső mintázat (L*HL) jellemzi, azonban a kérdő karakterkontúr az utolsó hangsúlyos szótagon indul; továbbá az emelkedő-eső dallam a hangsúlyos szótagot követő szótagok számától függően realizálódik [6]. A kísérlet hipotéziseként azt feltételeztük, hogy a szupraszegmentális szintű anyanyelvi transzferből [1], és a kínai és magyar nyelv eltérő intonációs sémáiból (valamint a lexikai tónusok hiányából) fakadóan a magyar anyanyelvűeknek nehézséget okoz a tónusok és az intonáció összehangolása, ezáltal a kérdések és kijelentések elkülönítése. Továbbá közvetlen anyanyelvi transzfert is feltételezünk, azaz, hogy az anyanyelvi dallammenet felülírja a kínai tónusok produkcióját: a kísérletben vizsgált három szótagú szintaktikailag jelölt kérdő megnyilatkozásokban (ahol az első szótag az alany, a második az ige, illetve a harmadik szótag a *ma* partikula), az ige tekintendő az utolsó hangsúlyos elemnek a megnyilatkozásban, tehát az anyanyelv szempontjából e szótag esetében várjuk a kérdő karakterkontúr kezdetét. Ez azt jelenti, hogy a magyar két szótagú kérdő dallam mintájára tónusértéktől függetlenül a kínai kérdés második szótagján várjuk az L* megvalósulását, míg a harmadik szótag esetében az HL mintázatot (pl. *És **jön** ma*?) [6].

A kísérletben két, eltérő kínai nyelvi tapasztalattal rendelkező, azaz egy kínai alapszakos kezdő és egy kínai mesterszakos haladó magyar anyanyelvű csoportot, valamint egy kínai anyanyelvű kontrollcsoportot vizsgáltunk (csoportonként 5 főt, nőket). A kísérleti személyeknek a következő kérdő és kijelentő megnyilatkozásokat kellett szembeállítani egymással produkciójukban: 1. 他来吗？他来。Tā lái ma? Tā lái.[tʰá lǎi ma? tʰá lǎi.] *Ő jön PRT? Ő jön.* 2. 他去吗？他去。Tā qù ma? Tā qù. [tʰá tɕʰŷ ma? tʰá tɕʰŷ.] *Ő megy PRT? Ő megy.* Az f_0-t a szótagok vokalikus fázisaiban 5 ms-onként nyertük ki, majd ezeket az értékeket félhangokká (st) konvertáltuk az R-ben [3]. Az előálló f_0-kontúrokat általánosított additív kevert modellekkel (GAMM) hasonlítottuk össze. Az 1-es számmal jelölt *lai*-os kérdés-kijelentés párban a hangsort csak szonoránsok alkották (kivéve az első /tʰ/-t), így lehetőség nyílt a kérdő és kijelentő megnyilatkozások f_0-kontúrját egyetlen görbével modellezni az első két szótagban. A 2-es számmal jelölt, *qu*-t tartalmazó kérdés-kijelentés pár esetében a második szótag elején obstruens állt, így ebben az esetben az első szótagot a másodiktól elválasztva, izoláltan elemeztük. A harmadik szótagot, azaz a *ma* partikulát mindkét megnyilatkozásban önálló modellel elemeztük, hiszen a *ma* partikula nem rendelkezik kijelentő modalitású megvalósulással.

A kísérlet általános eredményei szerint a natív kontrollcsoport szignifikánsan megkülönböztette az összes szótagot, ami a kérdések kijelentésekhez viszonyítva magasabb frekvenciatartományban való megvalósítását jelentette (1. ábra, bal). A natív ejtésben a *lai*-t tartalmazó megnyilatkozások második szótagján késleltetett (L-)célmegközelítés figyelhető meg, ami a tonális célok közötti különbségből fakad: az első szótag H-ját a második szótag kezdetén egy L követné, és a közöttük felmerülő hangmagasság-különbséget egy tranzíciós fázis oldotta fel [7]. Továbbá a második szótag emelkedő íve a *ma* partikulára is kiterjedt, a szótag második felére érve el maximumát. A két kínaiul tanuló csoport más-más mintázatokat alkalmazott a különböző modalitások, valamint a tónus-koartikuláció megvalósítására: a haladók esetében a kérdő és kijelentő megvalósulások f_0-görbéi nagyrészt átfedtek, és a modalitás nem befolyásolta a tónusok megvalósulását, valamint a második-harmadik szótagon megvalósuló emelkedés is elmaradt. A kezdők a kérdő *lai*-os megnyilatkozásokban egy natívoktól teljesen ellentétes irányú f_0-görbét produkáltak, azaz látszólag az előrefelé ható koartikulációs hatás helyett inkább hátrafelé irányuló hatásként már az első szótagban bekövetkezett mind a csökkenő tranzíciós fázis a HL célok között, mind a második szótagra jellemző emelkedés, mely látszólagosan mutat némi hasonlóságot a magyar kérdő intonációs sémával. Azonban a magyar kérdő intonációval vonva párhuzamot fontos megemlíteni, hogy a kezdők kérdő mintázata olyan három szótagú magyar megnyilatkozásra hasonlít, amelynek az első szótagja a hangsúlyos, így az L*HL mintázat minden célja különböző szótagon valósul meg (pl. *Emlékszel?*, ***Ő** jön ma?*). Ez bizonyos szempontból ellentmondásos, hiszen a felolvasandó kérdések fókuszában az ige áll, így a karakterkontúr kezdetét is ezen a szótagon várnánk *(És **jön** ma?)*. Ez a jelenség feltételezhetően a felolvasási feladatnak tulajdonítható. A megnyilatkozások általánosabb struktúráját figyelembe véve, amit a GAM-modellek parametrikus eredményeiből nyertünk ki (2. ábra, bal), azt láthatjuk, hogy míg a natív beszélők esetében a becsült átlagos különbség a *lai*-os megnyilatkozás első két szótagja között 3,2 félhang, addig a nyelvtanulóknál ez kevesebb, mint 1 félhang. A *ma* partikula relatív pozícióját illetően a haladók jobban megközelítik a natív ejtést, mint a kezdők. A *qu*-t tartalmazó megnyilatkozások esetében a natív kínaiak a kérdő f_0-kontúrját szignifikánsan magasabb értékek jellemezték az első két szótagban, mint a kijelentő megvalósulást, ugyanakkor a kínaiul tanuló magyarok leginkább csak a második szótagban különítették el jelentősen a kérdő és kijelentő f_0-görbéket (1. ábra, jobb). Megjegyzendő továbbá a második szótag kijelentő módú megvalósulása: a natívok esetében ez a kontúr egy domború, relatíve nagy terjedelemmel rendelkező ereszkedő mintázatot mutat, szemben a kínaiul tanulókkal, akik homorú és kisebb terjedelmű f_0-görbét produkáltak. A szótagszinten kinyert becsült átlagértékekből kiindulva azt mondhatjuk, hogy az első szótagban a kezdők jobban megközelítették a natív kontrasztot a kérdő és kijelentő modalitású megnyilatkozások között (natív különbség: 2,2 st, haladó: 0,7 st, kezdő: 1,4 st) (2. ábra, jobb). Ezzel ellentétben a második szótagban a haladók nem csak elérték, hanem felülmúlták a kérdő és kijelentő modalitású megvalósulások közötti különbségtételt (natív különbség: 5,1 st, haladó: 5,7 st, kezdő: 4,5 st) (2. ábra, jobb). A kérdő megnyilatkozások általánosabb struktúráját tekintve azonban a két nyelvtanuló csoport viszonylag hasonló, a natívtól mintázattól eltérő „egyenlő szárú háromszöges” szerkezetet produkált, azaz míg az első és harmadik szótag átlagértékei közel megegyeztek, a második szótag ezeknél magasabban realizálódott. Ezzel szemben a natív kínai produkciót inkább aszimmetria jellemezte, azaz a harmadik szótag az első szótagnál alacsonyabb becsült átlagértékkel valósult meg. Összegezve a kínaiul tanuló magyar anyanyelvűek tónusszekvencia függvényében eltérő hatékonysággal különböztették meg a kínai kérdő és kijelentő megnyilatkozásokat, illetve nem csak az eltérő modalitások okoztak problémát, hiszen a kínaiul tanulók ejtése a kijelentésekben is eltért a natív produkciótól.



1. ábra: A három beszélői csoport f_0-görbéi a három szótagra vetítve (ahol a piros folytonos vonal a kijelentő, a kék szaggatott vonal a kérdő f_0-görbét jelöli)



2. ábra: A három beszélői csoport görbénkénti átlagértékei a GAMM parametrikus eredményei alapján (balra a *lai*-os; jobbra a *qu*-s megnyilatkozások), ahol a kék háromszögek a kérdő, a piros körök pedig a kijelentő megnyilatkozások szótagok/modellek szintjén becsült f_0-átlagait mutatják be; illetve a szótagok pontjai között megjelenített számértékek a becsült különbségeket jelölik félhangokban (st)

Irodalom

[1] Jun, S. A. & Oh, M. (2000). ’Acquisition of Second Language Intonation’. *Acoustical Society of America Journal*. 107. 73–76.

[2] Lee O. J. (2005). *The prosody of questions in Beijing Mandarin*. PhD Disszertáció. Ohio, The Ohio State University.

[3] R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

[4] Shen X. N. (1989). ’Interplay of the four citation tones and intonation in Mandarin Chinese’. *Journal of Chinese Linguistics*, 17.1, 61–74.

[5] Shen, X. N. S. (1990). *The prosody of Mandarin Chinese*. California, University of California Press.

[6] Varga L. (2002). ’The intonation of monosyllabic Hungarian yes-no questions’. *Acta Linguistica Hungarica*, 49.3–4, 307–320.

[7] Xu, Y., & Wang, Q. E. (2001). ’Pitch targets and their realization: Evidence from Mandarin Chinese’. *Speech Communication*, 33.4, 319–337.

A hiátustöltőként és fonémarealizációként megjelenő [j]-variánsok elemzése a nyelvállás akusztikai vetületének szempontjából

Juhász Kornélia[1,2] és Deme Andrea[1]
[1] Eötvös Loránd Tudományegyetem,
[2] Nyelvtudományi Kutatóközpont

A jelen kutatásunkban azt a kérdést vizsgáljuk meg a [j] példáján, hogy okoz-e eltérést a beszédhangok megvalósításában az a tény, hogy a beszédhang a hangsor feltételezett fonológiai reprezentációjának részeként, egy fonéma realizációjaként valósul-e meg (/ijaː/), vagy pedig a koartikuláció sajátos jelenségének következtében hangátmenetként (/iaː/). A kísérletben dinamikus elemzéssel hasonlítjuk össze a hiátustöltőként és a /j/ fonéma realizációjaként megjelenő [j]-t a képzés során létrejövő szűkület akusztikai vetületének szempontjából. A kísérlet hipotéziseinek alapját [8] és [9] fonológiai elemzése szolgáltatja. Ezek szerint a mögöttesen jelenlévő /j/ likvidaként osztályozandó [+msh, +szon], tehát mássalhangzós tulajdonságokkal rendelkezik (obstruentizálódik (pl. *kapj* [kɔpç]), és asszimilálódik (pl. *moss*)), míg a hiátustöltő [j] siklóhang [–msh, +szon] [9: 6.], az /i/ jegyterjedésének eredménye a szomszédos onszet pozícióra [8: 285], amivel együtt a hiátustöltő megvalósulása is gyengébb és átmenetibb jellegű , mint a mögöttes /j/ realizációja [8: 91]. A kétféle [j] között továbbá azért is tehetünk különbséget, mert a fonetikaoktatásban is ismétlődő tapasztalat az, hogy az írás-olvasás tanulása során a helyesírás ki- illetve átalakítja a kiejtett hangalakhoz tartozó hangsorsémát a magyar anyanyelvű nyelvhasználók fejében azáltal, hogy bizonyos hangokhoz absztrakt sémát rendel – pl. ***kijár*** –, míg másokhoz – amilyen például a glottális zárhang, pl. *ha*[ʔ]*akkor* vagy a hiátustöltő [j], pl. ***hiá****ny* – nem, és ez utóbbiakban ezért gátolttá válik ezeknek a hangoknak a szegmentálása a beszélők számára. A fentebbi fonológiai elemzések mentén [8, 9], azaz a szerint, hogy a fonemikus /j/ realizációja „mássalhangzósabb", mint a hiátustöltő [j] a fonetikai megvalósításban az várható, hogy míg a /j/ képzése magasabb nyelvállással, azaz jelentősebb toldalékcső-szűkülettel valósul meg [4, 5], addig a hiátustöltő [j] esetében pusztán egy vokalikus hangátmenet jelenik meg a /j/-nél nyíltabb, felső nyelvállású, elöl képzett /i/ és a legalsó nyelvállású centrális /aː/ között úgy, hogy abban nem találunk egy önálló, a hiátustöltő [j]-hez tartozó akusztikai/artikulációs célt (jelentősebb szűkületet a toldalékcsőben) [vö. 2].

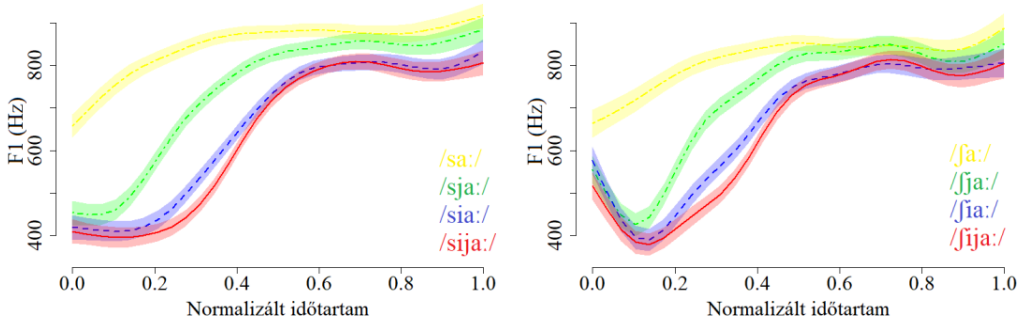
Korábbi kísérletek azt mutatták, hogy a fonémikusan /j/-t tartalmazó (pl. /ijaː/) hangsorok megvalósulása, amelyben a fenti okokból három artikulációs-akusztikai célt feltételezhetünk hosszabb, mint a hiátusos (pl. /iaː/) hangsoroké, ahol csak kettőt [3, 6]. [6] azt is megállapította, hogy az /ijaː/ hangsorban az első vokalikus cél az /iaː/ első céljánál akusztikailag palatalizáltabban (magasabb F_2-értékkel), illetve ezzel összefüggésben meredekebb tranzícióval valósult meg. A jelen kísérletben az előzőekben elemzett hangsorokban a nyelvemelkedés fokának az akusztikai vetületét vizsgáljuk. Várakozásaink szerint a fonemikus /j/ esetében a mássalhangzókra jellemző jelentősebb szűkület miatt magasabb nyelvállást, azaz alacsonyabb F_1 értéket tapasztalunk, mint a „magánhangzósabb" hiátustöltővel, azaz koartikulációs tranzícióval megvalósuló hangsorokban. Az eddigi eredményekhez képest a jelen kísérlet azért tekinthető újszerűnek, mert a /j/ approximánst a korábbiaktól eltérően nem a hangsorból kiemelve, statikusan elemzi, hanem a hangsorokat teljes egészében vizsgálja, ezáltal globálisabb képet kapva a környező beszédhangokban bekövetkező akusztikai változásokról is.

A kísérletben 14 magyar anyanyelvű felnőtt nő ejtését elemeztük. A vizsgált /iaː/ és /ijaː/ hangsorokat /s/- és /ʃ/-környezetben vettük fel nyolc ismétléssel, izolált ejtésben (*sziá szij, siá, sijá*, továbbá

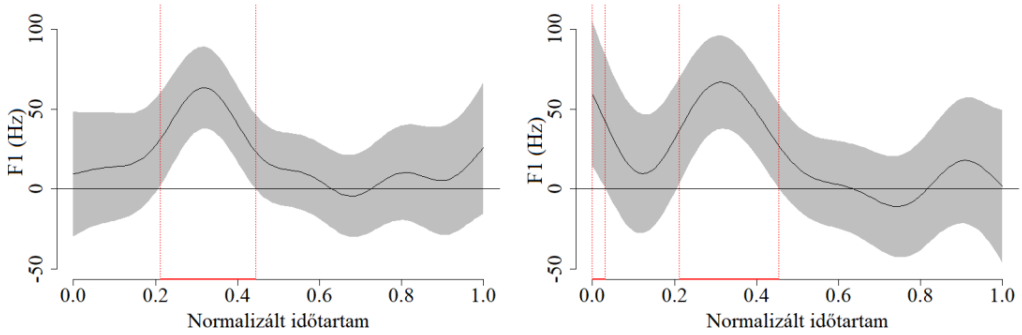
referenciaként rögzítettük a *szá sá* és *szjá, sjá* hangsorokat is. A nyelvemelkedés, akusztikai vetületét az ezzel összefüggésben álló F_1 frekvenciaértékével, valamint az intenzitás dB-ben mért értékével vizsgáltuk, ugyanis a szakirodalom szerint minél nagyobb a szűkület keresztmetszete (azaz minél alacsonyabb a nyelvállás) annál magasabb az F_1 értéke, és annál magasabb intenzitásérték mérhető [4, 5, 10]. A hangsorok vokalikus fázisában az F_1-et 5 ms-onként nyertük ki a Praatban [1], ahol az /ia:/ időtartama átlagosan 308 ms, az /ija:/ tartama pedig átlagosan 327 ms volt. Az F_1-kontúrokat a mássalhangzó-kontextusok szerint bontva, külön-külön elemeztük általánosított additív kevert modellekkel (GAMM). A két modellben a vokalikus szakasz teljes tartamához normalizált időtartam függvényében vizsgáltuk az F_1 változását, illetve a modellt tovább bővítettük a *hangsor* (/ia:/ vs. /ija:/) faktorral, valamint az F_1-kontúrokra görbénként random simítást illesztettünk. Ugyan a beszédhangokat nem szegmentáltuk a hangsorokban, de az intenzitást csak abban az intervallumban vizsgáltuk, ahol az /ia:/ és az /ija:/ F_1-görbéje eltérést mutatott: erről feltételezzük ugyanis, hogy az eltérő [j]-variánsokhoz tartoztak. Itt megvalósulásonként átlagértéket nyertünk ki és ezeket lineáris kevert modellekkel hasonlítottuk össze, melyekben a *hangsor* hatását vizsgáltuk (random hatás: beszélő).

A GAMM-elemzés eredményei szerint az /ia:/ és az /ija:/ hangsorok F_1-görbéi szignifikánsan eltértek egymástól ($p < 0{,}001$) mindkét mássalhangzó-kontextusban (1. ábra) ($R^2_{/s/\text{-kontextus}} = 92{,}3\%$; $R^2_{/ʃ/\text{-kontextus}} = 86{,}4\%$). A görbéken jól látható, hogy a normalizált időtartam elején az /i/ célja alacsony F_1-értékkel valósult meg (mindkét kontextusban), amit egy magasabb frekvenciaértékek felé irányuló tranzíció követett, majd a hangsor utolsó /a:/ eleme magas F_1-értékekkel realizálódott. Az /ia:/ és az /ija:/ görbéit összevetve a normalizált időtartam azonos fázisaiban találtunk szignifikáns eltérést (a /sia:/ és /sija:/ esetében 21% és 44,5% között, a /ʃia:/ és /ʃija:/esetében pedig 21% és 45% között), így ezt a fázist tekintettük a két eltérő [j]-variáns megvalósulásának (2. ábra). Ezekben a fázisokban minden esetben az /ija:/ görbéje valósult meg homorúbban, azaz az adott fázisban az /ija:/ mindig az /ia:/-nál alacsonyabb F_1-értékkel rendelkezett. Az itt mért intenzitásértékek az /ia:/ és /ija:/ hangsorok között ugyancsak szignifikáns különbséget mutattak: az /ija:/ hangsorban az intenzitás alacsonyabb volt, mint az /ia:/ esetében (/ia:/: 64,5±4,1 dB; /ija:/: 63,8±4,5 dB) ($p < 0{,}01$), bár itt a legjobb illesztést biztosító kevert modell szignifikáns eredménye elsősorban nem a független változó, hanem a random hatásból fakadt ($R^2_m < 1\%$; $R^2_c = 66\%$). Továbbá szignifikáns eltérést találtunk a két görbe /ʃ/-kontextusú megvalósulásai között a normalizált időtartam első 4%-ában is, ami nem volt jellemző a /s/-kontextusban. Ennek feltételezhetően az az oka, hogy a vizsgált hangsorokban a /ʃ/-ben nem volt megfigyelhető a /s/-ben tapasztalt palatalizálódás [7], és ennek a hatása jelent meg a vokalikus szakasz elején. Láthattuk, hogy az /ija:/ /i/-je az /ia:/ /i/-jéhez képest tendenciózusan alacsonyabb F_1-gyel realizálódott, és a /s/-kontextusú görbék vízszintes fázissal indultak, míg a /ʃ/-kontextusúak egy jelentősebb F_1-csökkenéssel. Ezek alapján feltételezhetjük, hogy az /ija:/ /i/-je a vokalikus szakasz legelejére (azaz a /ʃ/ és /i/, valamint a /s/ és /i/ tranzíciójának elejére) az F_1-et csökkentő koartikulációs hatást fejtett ki, és ennek volt köszönhető a vokalikus fázis kezdetén (a normalizált időtartam első 4%-ában) talált különbség.

Az eredmények alapján arra következtethetünk, hogy akusztikai szempontból a /j/ fonéma megvalósulásaként a [j] a várt módon magasabb nyelvemelkedéssel és szűkebb toldalékcsővel, azaz „mássalhangzósabban” valósult meg, mint a hiátustöltő [j]. Ezek az eredmények bővítik az ismereteinket arról, hogy a beszéd vélt fonológiai (mögöttes) reprezentációja miként hat a beszédhangok megvalósítására, illetve hogyan függ össze azzal.



1. ábra. Az /ia:/ és az /ija:/ hangsorok, valamint referenciaképpen az /a:/, /ja:/ hangsorok F_1-görbéi /s/- (balra) és /ʃ/-kontextusban (jobbra)



2. ábra. Az /ia:/ és az /ija:/ hangsor F_1-görbéi közötti eltérés mássalhangzó-kontextusonként (balra: /s/, jobbra: /ʃ/), ahol mindig az /ia:/ görbéje a referencia

Irodalom

[1] Boersma, P., & D. Weenink (2020). Praat: Doing phonetics by computer [Computer program]. 6.1.15-ös verzió http://www.praat.org.

[2] Davidson, L., & Erker, D. (2014). Hiatus resolution in American English: The case against glide insertion. Language, 90/2, 482–514. old.

[3] Gósy M. (2014). 'A palatális közelítőhang kétféle funkcióban'. *Beszédkutatás* 22: 17–40. old.

[4] Hunt, E. (2003). *Acoustic Characterization of the Glides /j/ and /w/ in American English*. PhD Disszertáció. Princeton: Princeton University.

[5] Jaggers, Z. S. 2018. 'Evidence and characterization of a glide-vowel distinction in American English'. *Laboratory Phonology* 9: 1–27. old.

[6] Juhász K., & Deme A. (2022). 'Palatális approximánsok a kínaiban és a magyarban – a kínai alveolopalatális /ɕ/ szibilánst követő vokális szakasz produkciója kínaiul tanuló magyar anyanyelvűeknél'. *Általános Nyelvészeti Tanulmányok* XXXIV. 287–332. old.

[7] Juhász K., & Deme A. (2022). 'Mandarin kínai és magyar szibilánsok palatalizációja kínaiul tanuló magyar anyanyelvűek ejtésében'. Szerk. Mády, K., & Markó, A. *Alkalmazott Nyelvtudomány*, XXII. 1, 64–89. old.

[8] Siptár P., & Törkenczy M. (2000). *The phonology of Hungarian.* Oxford: Oxford University Press.

[9] Siptár P. (2013). 'Palatálisok'. *„…hogy legyen a víznek lefolyása…" : Köszöntő kötet Szilágyi N. Sándor tiszteletére*. Szerk. Benő, A., Fazakas, E., & Kádár, E. Kolozsvár: Erdélyi Múzeum Egyesület 433–448. old.

[10] Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Multilingualism - our friend in the school

Annamária Kacsur
University of Pannonia, Hungary

According to a recent handbook on pedagogical translanguaging, bringing in different languages to a classroom may have a positive impact, as they are capable of reinforcing each other. Moreover, learners' pre-existing linguistic skills and competences may be beneficial. Therefore, accessing two or more languages during the same lesson does not mean less exposure to the target language [1]. However, this is influenced by two factors: the attitude of the learners and the teacher's ability to control the situation [2]. The present survey seeks to answer the question of how pupils in Transcarpathia, the most western part of Ukraine on the Hungarian border, really see the simultaneous use of several languages as an advantage, and what possible difficulties they face.

The questionnaire, compiled at the beginning of the research, was completed online with the help of 50 pupils from four different secondary schools with Hungarian as the language of instruction in Transcarpathia. Respondents, aged between 15 and 17, were starting 10th and 11th grade, when the research was conducted. Some had Hungarian as L1 and others had Ukrainian as L1. They all acquired a second language (either Hungarian or Ukrainian) before entering institutional education, while English often became part of their lives as early as kindergarten.

The questionnaire had nine open ended questions. After questions on their language history, they were asked to comment on difficulties they face in class, the strategies they use to solve them, their use of languages with their friends, and finally their views on the advantages of bilingualism.

It was revealed that one third of the respondents had language difficulties in the school. They identified one main reason as the source of their problems. These pupils previously used to study in schools with Ukrainian language of instruction. Although they were fluent in two or three languages during everyday conversations, they had problems with the subject specific vocabulary used during the lessons, which they had learned formerly in another language. However, in response to the following question, these pupils named the ways in which they try to overcome the obstacles. Among others, bilingual note-taking and detailed explanations from the teacher were mentioned.

Nevertheless, during conversations with their friends at school or interpreting and preparing homework together, multilingualism was not a problem. As learners do not feel bound by the rules of the classroom, translanguaging can be given greater scope.

On the other hand, it is also an important question whether a learner discovers when he/she can take advantage of translanguaging. Based on their answers, the respondents can be divided into three groups: those who have been taught by their parents to use this phenomenon effectively, those who use their own intuition to navigate between languages and those who are not aware of other people's language usage when speaking.

In conclusion, it can be stated that pupils really benefit from using several languages at the same time. However, there is still room for improvement in learners' ability to assess in which situations translanguaging can be used.

It is intended to further investigate the topic paying particular attention to the role of translanguaging in the English classroom

References

[1] Cenoz, J., & Gorter, D. (2021). *Pedagogical translanguaging*. Cambridge: Cambridge University Press.

[2] Tódor, E.-M. (2021). ’Language use during Romanian classes in bilingual settings’. In:*Acta Universitatis Sapientiae - Philologica, 13*(2). Ed. by Tonk M. Kolozsvár: Scientia Publishing House. pp. 1-20.

Jaw position as an articulatory correlate of rhythm in spoken Polish. Study design.

Hanna Kasperek
Adam Mickiewicz University in Poznań, Poland

Research on speech rhythm in Polish has been focused mostly on its acoustic determinants [8, 2, 10, 11, 12, 7, 14, 15]. The articulatory aspects of prosody, including rhythm, have been the subject of various studies which indicate the effect of syllable stress on the jaw lowering in English (American variant) and Japanese [4, 3, 9]. Studies of jaw displacement patterns have also been conducted for Spanish [6], Mandarin Chinese [5], and French [13]. In each of the aforementioned languages, distinctive and unique patterns of lowering the mandible were noted, reflecting the metrical structure of the sentences of these languages.

Taking into account the above, as well as the fact that Polish has a phonotactic structure different from the aforementioned languages (Polish allows phonotactically complex words with CCCCVC structure, e.g., *pstryk*, *pstrąg*) and a different phonological inventory, it might be expected that unique articulatory correlates of speech prosody in Polish will emerge in the course of the study. It should also be noted that Polish is considered a language with an ambiguous rhythmic structure [11, 16, 14], so the present research will contribute to expanding knowledge in this area, and it might shed new light on the issues related to the rhythm of speech in Polish. Unlike most of the traditional phonetic research on the prosody of Polish that has been concerned primarily with acoustic variables, the present study focuses on the articulatory features.

As a research hypothesis, it is assumed that there is a correlation of jaw position with the metrical structure of Polish speech. Prominent syllables will be associated with a stronger lowering of the mandible, relative to the occlusal plane, and the graph visualising the jaw movements will coincide with word stress and, to a greater extent, sentence stress. See Fig. 1 for a visualisation of possible displacement (lowering) of mandibula.

The project combines articulatory research methodology with phonetic-acoustic research. Quantitative analysis based on data from acoustic and articulatory measurements will be used. The course of the research tasks can be divided into four main stages; in the first stage, stimuli will be prepared, which involves the selection of linguistic material based on a literature review. The stimuli ought to be chosen to control for vowel height as it affects jaw opening [see e.g., [13]]. In the second research stage, articulographic measurements will be carried out using the only Carstens AG501 electromagnetic articulograph (EMA) available in Poland at the Maria Przybysz-Piwko Applied Phonetics Laboratory located at the Institute of Applied Polish Studies at the University of Warsaw. EMA enables recording, storing, presenting, and evaluating the movements of the articulators (including the tongue, lips, mandible) in real time, i.e. while speaking, and present them in 3D space. Experimental research with human subjects will be carried out after obtaining written consents from the subjects to participate in the study and to use the results in scientific research. An application for the approval for the EMA studies from the Rector's Committee on Ethics in Research with Human Participation at the University of Warsaw has already been prepared. The collected data will be anonymized as a first step and only in this form will it be used subsequently.

It is planned to collect data from 3 speakers. To obtain information on articulatory

movements, their positioning, shape, and speed during utterance production, EMA sensors will be placed and attached to selected points, including the speaker's mandible, e.g. onto lower central incisor, and reference points, e.g., the bridge of the nose. Simultaneously with the articulographic examination, the subjects' utterances will be recorded for acoustic analysis. The phonetic-acoustic parameters included in the study are pitch, F1 and F2 formants, and speech rate and its intensity. Measurements will be carried out using the Praat software [1]. Speech segmentation for speech rate measurement will be carried out using the ANNPRO software plug-in [10]. Then, rPVI (raw Pairwise Variability Index), nPVI (normalized Pairwise Variability Index) and the difference between F1 and F2 will be calculated. The obtained data will be then used to statistically investigate the correlation between the acoustic correlates of rhythm in Polish (dependent variables) and the position of the mandible in relation to the maxillary occlusal plane during speech (independent variable).

It is expected that articulatory and acoustic analysis of spoken Polish will reveal the unique articulatory features of Polish prosody. The results will also enable comparisons to be made. This phase will enable the testing of the research hypotheses. Conducting and publishing the results of the following study will provide new insights into such correlations in Polish speech, as well as contribute to scientific dialogue by providing materials for comparative studies.

Rhythmic structure influences the interpretation and understanding of utterances, and for this reason the study of rhythm has the potential to enrich linguistics in terms of knowledge of prosody, among other things, and can also be used, for example, in glottodidactics or speech synthesis enhancement.

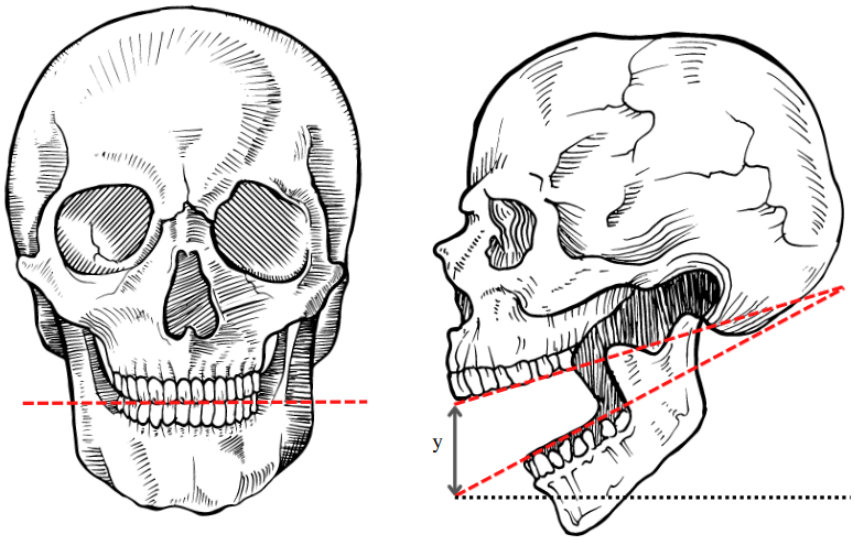


Figure 1: Possible mandibula displacement in vertical plane (y) during articulation. Red line represents the maxillary occlusal plane.

This contribution presents the background and description of planned research financed by IDUB 054/13/SNJL/0002 grant to be realised at The Institute of Applied Linguistics (ILS) at Adam Mickiewicz University in Poznań and The Institute of Applied Polish Studies (IPS) in the Faculties of Polish Studies of the Warsaw University.

References

[1] Boersma, P. & D. Weenink (2014). 'Praat: Doing phonetics by computer [computer program]'. Version 6.2.23. In: URL: http://www.praat.org.

[2] Demenko, G. (2003). 'Phrase time structure modelling for speech synthesis purposes'. In: *Proc. ICPhS*. Vol. 15, pp. 2441–2444.

[3] Erickson, D. (2004). 'On phrasal organization and jaw opening'. In: *Proceedings of From Sound to Sense, June 13, MIT*, p. 24.

[4] Erickson, D. (1998). 'Effects of contrastive emphasis on jaw opening'. In: *Phonetica* 55.3, pp. 147–169.

[5] Erickson, D. & S. Kawahara (2016). 'Articulatory correlates of metrical structure: Studying jaw displacement patterns'. In: *Linguistics Vanguard* 2.1.

[6] Erickson, D. et al. (2015). 'Spanish articulatory rhythm'. In: *Acoustical Society of Japan Fall Meeting*, pp. 319–322.

[7] Gibbon, D., K. Klessa & J. Bachan (2014). 'Duration and speed of speech events: A selection of methods'. In: *Lingua Posnaniensis* 56.1, pp. 59–83.

[8] Jassem, W., M. Krzyśko & P. Stolarski (1981). 'Regresyjny model izochronizmu zestrojowego w sygnale mowy'. In: *Prace IPPT IFTR REPORTS*.

[9] Kawahara, S., D. Erickson & A. Suemitsu (2015). 'Edge Prominence and Declination in Japanese Jaw Displacement Patterns: A View from the C/D Model'. In: *Journal of the Phonetic Society of Japan* 19.2, pp. 33–43.

[10] Klessa, K. et al. (2022). 'ANNPRO: A Desktop Module for Automatic Segmentation and Transcription'. In: *Language and Technology Conference*. Springer, pp. 65–77.

[11] Malisz, Z. (2013). 'Speech rhythm variability in Polish and English: A study of interaction between rhythmic levels'. PhD thesis. URL: https://pub.uni-bielefeld.de/record/2604114#fileDetails.

[12] Malisz, Z., M. Żygis & B. Pompino-Marschall (2013). 'Rhythmic structure effects on glottalisation: A study of different speech styles in Polish and German'. In: *Laboratory Phonology* 4.1, pp. 119–158.

[13] Smith, C. L., D. Erickson & C. Savariaux (2019). 'Articulatory and acoustic correlates of prominence in French: Comparing L1 and L2 speakers'. In: *Journal of Phonetics* 77, p. 100938.

[14] Wagner, A. (2017). *Rytm w mowie i języku w ujęciu wielowymiarowym*. Dom Wydawniczy "Elipsa".

[15] Wagner, A., K. Klessa & J. Bachan (2016). 'Polish Rhythmic Database—New Resources for Speech Timing and Rhythm Analysis'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4678–4683.

[16] Wagner, P. (2008). 'The rhythm of language and speech: Constraining factors, models, metrics and applications'. Habilitation dissertation. URL: https://pub.uni-bielefeld.de/record/1916845#fileDetails.

Nyelvjárási *ö*-zés akusztikai vizsgálata

Kocsis Zsuzsanna
kocsis.zsuzsanna@nytud.hu
Nyelvtudományi Kutatóközpont,
Történeti Nyelvészeti és Uralisztikai Intézet, Lexikológiai Intézet

A magyar nyelvjárások meghatározásának, rendszerezésének, csoportosításának egyik legfontosabb szempontja mindig is a magánhangzó-rendszerek leírása volt (vö. [1, 2, 3, 5, 4]). A különböző hangtani sajátosságok közül az *ö*-zés jelensége kiemelkedő szerepet kapott az egyes régiók és csoportok elkülönítésében ([3, 7]). A hagyományos dialektológiai kutatások az *ö*-zést elsősorban két szempontból vizsgálták: egyfelől arra koncentráltak, hogy az *ö*-zés mely szótagi pozícióban realizálódik; másfelől az *ö* [ø] magánhangzó köznyelvi *e* [ɛ], illetve *ë* [e] hanggal szembeni gyakoriságát nézték (l. [3, 4, 7]).

A Magyar nyelvjárások akusztikai és dialektometriai vizsgálata című projekben a magyar nyelvjárások maganhangzó-rendszereit vizsgáljuk (NKFIH FK-138396, http://nyelvfoldrajz.nlp.nytud.hu). A magyar nyelvjárások, nyelvjárási változások vizsgálatát objektivitásra törekedve, akusztikai adatokon keresztül valósítjuk meg (vö. [8, 9, 6, 11, 12, 10]). Jelen előadás ennek a vállalkozásnak egy alprojektjét mutatja be, melynek során az *ö*-zést nem hagyományosan, vagyis a gyakoriság felől, hanem az akusztikum oldaláról közelítjük meg. Feltételezzük ugyanis, hogy vannak olyan nyelvjárások, amelyeket nemcsak az *ö* hang gyakorisága, hanem annak minősége alapján is le lehet írni. Sőt, az akusztikai vizsgálatok az *ö*-zés sajátosságainak, illetve a magyar nyelvjárások osztályozásának, csoportosításának újragondolását, finomítását tehetik lehetővé. Az a megállapítás, hogy a hang különböző minőségekben létezik, nem újkeletű a dialektológiai kutatásokban. Több nyelvatlaszunk (pl. MNYA. [13], S–ZA. [14]) különböző finomságban, de jelöli az *ö* és az *ë* közti átmeneti hangrealizációkat (*ȍ*, *ȩ̈*). Imre Samu az *ö*-zés bemutatása során maga is említést tesz ezen hangokról, ugyanakkor ezeket az *ö*-ző területeken köznyelvi hatásként értelmezi, „rendszerszerűtlen, nem megfogható” jelenségnek véli ([3]: 202–203). A nyelvatlaszok alapján azonban feltételezhetjük, hogy ezeknek az „átmeneteknek” ([3]: 202) a felbukkanása korántsem olyan rendszertelen, mint Imre megállapítja, sőt, a dél-dunántúli kutatópontokon nagyobb gyakorisággal fordulnak elő, kvantitatív elemzéssel kimutatható térbeliségük van.

Jelen kutatás ezen előfeltevést kívánja részletesen megvizsgálni a szubjektivitás minimalizálása érdekében akusztikai mérésekkel. Vizsgálatunk anyaga a Magyar nyelvjárások atlaszának gyűjtőmunkálatai során készített beszélgetések hangfelvételei (1960–1964). Az elemezni kívánt hangfelvételeket az MNyA. és S–ZA. adatai alapján jelöltük ki: olyan településeket kerestünk melyeknél a középső nyelvállású palatális hangok labializáltságában átmenetiség található. Az atlaszok eredményei alapján kiválasztottuk azokat a dél-dunántúli kutatópontokat, ahol a vizsgálni kívánt jelenség a leggyakoribb volt a MNyA. 1055 térképlapjának adatai alapján. A kutatás jelenleg még kezdeti fázisában van, a szóban forgó kutatópontok közül eddig négyet találtunk, ahonnan a vizsgálathoz megfelelő terjedelemben és elfogadható minőségben maradt fönn hangfelvétel a MNyA. ellenőrző gyűjtéséből: Bödeháza, Döbrököz, Mesztegnyő, illetve Péterhida hanganyagával dolgoztunk, jelen előadás ezen kutatópontok eddigi vizsgálatát mutatja be. A felvételekből először kiválasztottunk egy-egy minimum 5 perc terjedelmű interjút, amelyet

a Bihalbocs szoftverrel lejegyeztünk (http://www.bihalbocs.hu), időzítési markerekkel biztosítva a kapcsolatot az eredeti hanganyaggal. Ezekben első és második szótagban, viszonylag semlegesnek tekinthető mássalhangzó-környezetben mértük a magánhangzók első és második formánsainak értékét a Praat szoftverrel, különös tekintettel az *ö*-zés szempontjából meghatározó helyen realizálódó hangokra. Eddigi eredményeink abba az irányba mutatnak, hogy a) a mért eredmények részben igazolják az MNyA. adatai alapján kirajzolódó jelenséget; b) ugyanakkor elmondható, hogy az *ö* és a zárt *ë* közti átmenetek – a zárt *ë* adott nyelvjárásra jellemző minőségével összefüggésben – az *ö* tipikus ejtésénél nyíltabbak, ami a MNyA. lejegyzéseiből nem következik.

Úgy gondoljuk, hogy az akusztikai vizsgálatok objektívebb képet adhatnak a magánhangzó-minőségek összefüggéseiről, mint a korábbi, hallás után készült, s nem ritkán a kutatói előfeltevések által befolyásolt szubjektív lejegyzésekre támaszkodó kutatások, s ezzel egyúttal megalapozzák a későbbiekben elvégzendő valósidejű vizsgálatokat.

Irodalom

[1] Balassa, J. (1891). *A magyar nyelvjárások osztályozása és jellemzése*. Budapest: Magyar Tudományos Akadémia.

[2] Horger, A. (1934). *A magyar nyelvjárások*. Budapest: Kókai Lajos kiadása.

[3] Imre, S. (1971). *A mai magyar nyelvjárások rendszere*. Budapest: Akadémiai Kiadó.

[4] Juhász, D. (2001). ‘A nyelvjárási régiók’. *Magyar dialektológia*. Szerk. Kiss, J. Budapest: Osiris Kiadó.

[5] Kálmán, B. (1975). *Nyelvjárásaink*. 3. kiadás. Budapest: Tankönyvkiadó.

[6] Kocsis, Zs. (2013). ‘Nyelvjárási e hangok vizsgálata különös tekintettel a labializáció mértékére.’ *Elmélet és empíria a szociolingvisztikában: válogatás a 17. Élőnyelvi Konferencia előadásaiból*. Szerk. Kontra, M., Németh, M., Sinkovics B. Budapest: Gondolat Kiadó.

[7] Kontra, M., Németh, M., Sinkovics B. szerk. (2016). *Szeged nyelve a 21. század elején*. Budapest: Gondolat Kiadó.

[8] Labov, W., Ash, S., Boberg, Ch. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin/New York: Mouton de Gruyter.

[9] Leinonen, T. (2010). *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects*. Groningen Dissertations in Linguistics 83.

[10] Mády Katalain (2015. A fogadodban sok a vendég: hosszú magánhangzók egy mezőségi nyelvjárásban. In: É. Kiss Katalin – Hegedűs Attila – Pintér Lilla eds. *Nyelvelmélet és dialektológia 3*. PPKE BTK Elméleti Nyelvészeti Tanszék – Magyar Nyelvészeti Tanszék, Piliscsaba. 196–210.

[11] Vargha, F. S. (2007). ‘Nyelvi változók A magyar nyelvjárások atlasza hangfelvételeiben.’ *V. Dialektológiai szimpozion*. Szerk. Guttmann, M., Molnár, Z. Szombathely: Berzsenyi Dániel Főiskola, 279–288.

[12] Vargha, F. S. (2013). ‘A hangzó adat szerepe a magyar dialektológiában’. Szerk. Szoták, Sz., Vargha, F. S. *Változó nyelv, nyelvváltozatok, területiség: a VII. Nemzetközi Hungarológiai Kongresszus szekcióelőadásai.* Kolozsvár: Egyetemi Műhely Kiadó, Bolyai Társaság, 194–204.

Rövidítések feloldása

[13] MNYA. = *A magyar nyelvjárások atlasza* 1–6. (1968–1977) Szerk. Deme, L., Imre, S. Budapest: Akadémiai Kiadó.

[14] S–ZA. = Király, L. (2005). *Somogy–zalai nyelvatlasz*. Budapest: Magyar Nyelvtudományi Társaság.

A frázishatár zöngeminőségének vizsgálata felnőttekhez és gyerekekhez szóló beszédben

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

A beszédben a frázishatárokat gyakran jelölik frázisvégi nyúlással, alacsony határjelző tónussal, szünettel vagy glottalizációval [1, 2]. A különböző frázishatárok jelölése segíti a befogadó számára a beszéd egységekre tagolását és így a beszédmegértést is, ezért nemcsak a felnőttek közötti kommunikációban, hanem a nyelvelsajátítás szempontjából is kulcsfontosságú lehet.

A gyerekekhez szóló beszédben a frázishatárokon történő nyúlást, szünetezést több nyelvben is vizsgálták, köztük a magyar beszédben is [3]. Felolvasott mondatokban azt találták, hogy a frázis határ előtt a szótagok magánhangzói nyúlnak, és gyakoribbak a szünetek, és ez a két jelenség a felnőttekhez és a gyerekekhez szóló beszédben hasonló volt. A frázishatár zöngeminőséggel történő jelöltségének vizsgálata ugyanakkor eddig elmaradt, miközben ismert, hogy a glottalizáció gyakran megjelenik intonációs frázisok (IP) végén [2]. A magyar beszédben az intonációs frázis mellett feltételezik, hogy az úgynevezett akcentuális frázis (AP) is megjelenik prozódiai egységként [4], amely hangsúlyos szótaggal kezdődik, és a követő hangsúlytalan szótagokat foglalja magában a következő hangsúlyos szótagig vagy az IP végéig. Az akcentuális frázis határon megvalósuló zöngeminőséget eddig még nem vizsgálták a magyar beszédben.

A zöngeminőség akusztikai mérésére alapvetően 3 mérőszámtípust szoktak alkalmazni, amelyek a percepciós tesztek alapján a különböző zöngeminőségekkel jól korrelálnak: H1*−H2*, HNR és CPP [5]. A glottalizációhoz tipikusan alacsony H1*−H2* érték társul, míg a modális zöngére magasabb, leheletes zöngére még magasabb érték jellemző. Habár a prototipikusnak tekintett glottalizációhoz kevésbé erős periodikusságot, így alacsonyabb HNR-t és CPP-t társítanak, egy másik glottalizációtípus esetén magas HNR értékek is megjelenhetnek [6].

Jelen pilotkutatás célja egyrészt az, hogy felolvasott mondatokon megvizsgálja a különböző típusú frázishatárokon a zönge minőségét. Másrészt arra a kérdésre is keressük a választ, hogy a különböző frázishatárok glottalizációval történő jelöltsége eltér-e a gyerekekhez és a felnőttekhez szóló beszédben.

A vizsgálathoz 20 édesanya felvételeit használtuk fel. A beszélők egy felnőtt kísérletvezetőnek és saját gyereküknek is elmesélték ugyanazt a történetet részben saját szavaikkal, részben a narratívába beépített mondatokkal. A kísérletet a baba különböző életkoraiban megismételték, amelyek közül a gyerekek 4, 8 és 18 hónapos korában rögzített, a laborban fejmikrofonnal készült felvételeit használtuk fel. A vizsgálathoz a felolvasott mondatokból hármat választottunk ki, amelyekben szerepelt a *látlak* szó különböző prozódiai határt megelőzően: frázis belsejében szóhatárként (W), akcentuális (AP) és intonációs frázishatáron (IP). A három különböző felolvasott mondat a következő volt: *Látlak ám a málnásban.* (W); *Látlak odalent a tóparton.* (AP); *Látlak, ott vagy a barlangban.* (IP). A határt megelőző *-lak* szótag magánhangzójának egészén végeztünk akusztikai méréseket (pl. H1*−H2*, CPP, HNR) a VoiceSauce programmal. Az R-ben végzett statisztikai elemzésekhez lineáris kevert modelleket alkalmaztunk a különböző mérőszámokra a *lmerTest* csomag segítségével, amelyekben a regisztert, a gyerek életkorát és a

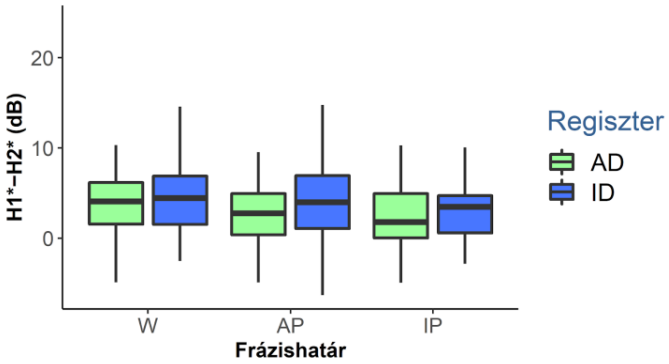
frázishatár típusát tekintettük fix hatásként, és a beszélőket random hatásként, és szükség esetén az *emmeans* csomagban a Tukey-féle post hoc tesztet használtuk.

A kevert modell eredményei azt mutatták, hogy a H1*−H2* értékei alacsonyabbak voltak IP határon, mint szóhatáron (β = 1,59, SE = 0,45, t = 3,54, p = 0,001; marginális R^2: 0,04, kondicionális R^2: 0,23), azaz a célszó utolsó magánhangzóját feltehetően glottalizáltabban ejtették. Az AP határ magánhangzóinak értékei ugyanakkor nem mutattak eltérést a másik két prozódiai határ magánhangzóitól. Ugyanezen mérőszám a dajkanyelvben magasabbnak mutatkozott a felnőttekhez szóló beszédhez képest (β = 0,75, SE = 0,36, t = 2,07, p < 0,04), ami arra utalhat, hogy a hangszalagoknak nagyobb a nyitottsági hányadosa ebben a regiszterben. A modell nem jelzett interakciót a regiszter és a prozódiai egység típusa között. Továbbá az életkor nem mutatott hatást a H1*−H2* értékeire sem önmagában, sem interakcióban.

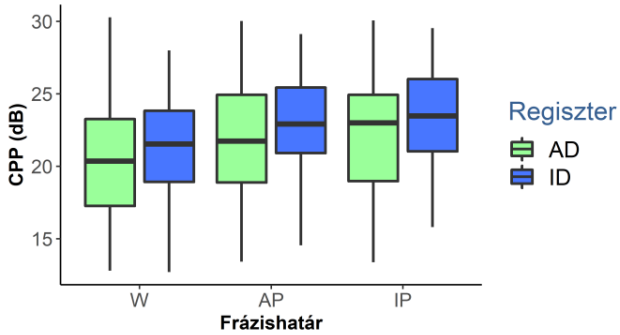
A CPP-értékek magasabbak voltak IP és AP határon lévő magánhangzók esetében, mint szóhatáron, azaz kevésbé zörejes zöngeképzés volt jellemző ezeken a határokon (AP W-hez képest: β = −1,52, SE = 0,46, t = −3,33, p = 0,003; IP W-hez képest β = −1,90, SE = 0,46, t = −4,12, p < 0,001, marginális R^2: 0,05, kondicionális R^2: 0,28). A HNR mérőszám, amely a CPP-hez hasonlóan a zönge periodikusságát számszerűsíti, szintén az IP és AP határon lévő magánhangzók esetében mutatkozott magasabbnak a szóhatáron előforduló magánhangzókon mért értékekhez képest (AP W-hez képest: β = −4,26, SE = 1,17, t = −3,65, p < 0,001; IP W-hez képest β = −5,61, SE = 1,19, t = −4,73, p < 0,001, marginális R^2: 0,08, kondicionális R^2: 0,45). Továbbá a CPP és HNR értékek jellemzően magasabbak voltak, azaz periodikusabb zöngeképzés volt jellemző a dajkanyelvben, mint a felnőttekhez szóló beszédben (CPP: β = 0,94, SE = 0,37, t = 2,52, p = 0,01; HNR: β = 4,14, SE = 1,50, t = 2,77, p = 0,01). Az életkor nem mutatott eltérést a különböző mérőszámok esetében.

A különböző típusú frázishatárok tehát eltérő zöngeminőséget mutattak. Az intonációs frázishatáron előforduló magánhangzók spektrális lejtése valamivel alacsonyabb volt, mint a szóhatáron lévő magánhangzóké. Ez arra utalhat, hogy az intonációs frázishatárt a beszélők hajlamosabbak glottalizáltabban ejteni [hasonlóan 1, 2]. A másik két vizsgált mérőszám (HNR, CPP) értékei ugyanakkor arra utalnak, hogy az intonációs frázishatáron lévő magánhangzókat periodikusabb zönge jellemezte, mint a frázis belsejében lévő magánhangzókat. Az akcentuális frázishatáron lévő magánhangzók hol inkább az intonációs frázishatáron lévő (HNR, CPP), hol a frázis belseji magánhangzók zöngeminőségéhez (H1*−H2*) hasonlítottak, tehát nem különültek el egyértelműen az IP és W kategóriáktól. A frázishatártípusok mellett a regiszter is hatással volt a zönge minőségére. A dajkanyelvben tipikusan periodikusabb zöngeminőség volt jellemző a felnőttekhez szóló beszédhez képest, és a spektrális lejtés magasabb volt, ami arra utalhat, hogy a zönge képzése a modálisabb képzés felé tolódik. A későbbiekben percepción alapuló kategorizáció (leheletes, modális, glottalizált zönge) adhatja meg a mérőszámok által mutatott regiszterbeli és prozódiai határon lévő eltérések tényleges jelentőségét.

A vizsgálat *A korai nyelvfejlődés neuro-kognitív előjelzői* című projekt (NKFI-115385), a Prozódiai szerkezet és mondattípusok vizsgálata nagy beszédadatbázisokon mély tanulási támogatással című projekt (NKFI-135038) és *A dajkanyelv longitudinális vizsgálata multimodális módszerekkel* című projekt (NKFI-134775) keretében készült.



1. ábra. H1*−H2* mérőszám regiszterenként (AD: felnőttekhez szóló beszéd, ID: gyerekekhez szóló beszéd) különböző prozódiai határok esetében



2. ábra. CPP mérőszám regiszterenként (AD: felnőttekhez szóló beszéd, ID: gyerekekhez szóló beszéd) különböző prozódiai határok esetében

Irodalom

[1] Krepsz, V., Huszár, A., & Horváth, V. (2022). 'Acoustic realization of prosodic cues in boundary functions in Hungarian'. In: *42nd Annual Meeting of the Department of Linguistics*. Thessaloniki: Aristotle University of Thessaloniki, pp. 1–3. https://www.lit.auth.gr/amgl42/abstracts/poster_Krepsz%20Huszar%20Horvath.pdf

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

[3] Szalontai, Á., Mády, K., Deme, A., & Kohári, A. (2018). 'Prosodic boundaries in Hungarian infant-directed speech'. In: *Proceedings Experimental and Theoretical Advances in Prosody (ETAP) 4*, pp. 1–2.

[4] Beňuš, Š., Reichel, U. D. & Mády, K. (2014). 'Modelling accentual phrase intonation in Slovak and Hungarian'. In: *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*. Ed. by Veselovská, L. & Janebová, M. Olomouc: Palacký University, pp. 677–689.

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

[6] Keating, P. A., Garellek, M., & Kreiman, J. (2015). 'Acoustic properties of different kinds of creaky voice'. In: *Proceedings of ICPhS 2015*, pp. 2–7.

Az átszótagolás jelenléte és megítélése a magyar nyelvben attitűdvizsgálat alapján

Kovács Dorottya
Eötvös Loránd Tudományegyetem, Budapest

A beszédfolyamat során, vagyis ha különböző szavak összekapcsolásának köszönhetően két szótag szomszédossá válik, úgynevezett *átszótagolás* jöhet létre ([9] idézi [2]). E jelenség alatt (általában) azt értjük, hogy két szomszédos szó közül az elsőnek az utolsó mássalhangzója a következő, magánhangzóval kezdődő szó elejére kerül át ([7]). Az átszótagolás alapvető jellemzője például a spanyol és a francia nyelvnek ([1, 10]), de megjelenik az olaszban, a hollandban, az egyiptomi arabban, és bizonyos források szerint az angolban is ([10]). A magyar nyelvre vonatkozóan [5] (idézi [8]) arról ír, hogy az átszótagolás a szótő utolsó szótagjának zárlata és a csatolt, magánhangzóval kezdődő toldalék első hangja között jöhet létre. Erre egy példa az „ablakok" szó „ab.la.kok"-ként való szótagolása. Ezzel szemben viszont [2] szerint a szótagzárlat utolsó mássalhangzója akkor helyeződik át a következő szótag kezdetébe, ha az adott zárlat két mássalhangzóból, a rákövetkező szótagkezdet pedig egyetlen mássalhangzóból áll. Ekkor azonban a magyar nyelv esetében fonotaktikai akadályokba ütköznénk, hiszen a nyelv számára idegen mássalhangzó-torlódások jönnének létre (például: „mos.d meg"). Jelen kutatásban arra keressük a választ, hogy a magyar nyelvben valóban előfordul-e az átszótagolás, illetve mennyire tűnik elfogadhatónak a magyar fül számára. Ehhez olyan attitűdvizsgálatot végeztünk, melyben az adatközlők egy online kérdőív formájában értékeltek meghallgatott mondatokat.

A kérdőívhez használt felvételeken egy 21 éves, magyar anyanyelvű nő felolvasása hallható, aki nem kapott semmilyen utasítást arra vonatkozóan, hogyan olvasson vagy artikuláljon. A kérdőívben 15 mondat kapott helyet, melyből négy olyan mondatpár, ami az átszótagolástól vagy annak hiányától függően kétértelmű (pl. „A**z á**rban / A **zá**rban benne van a kulcs?"). Ezen felül egy mondatpár csak abban különbözik egy korábbitól, hogy az esetleges átszótagolást tartalmazó szókapcsolat a mondat végén helyezkedik el („A kulcs benne van a **zá**rban?" és „A kulcs benne van a**z á**rban?"), ezzel megadva a lehetőségét annak, hogy a lehetséges átszótagolás pozíciójának hatását is vizsgálni tudjuk. Annak érdekében, hogy a lehető legkevesebb tényező befolyásolja a kitöltőket (így például a felolvasás milyensége), minden mondatpár egyik tagja manipulált hangfelvétel. A szerkesztés úgy történt, hogy a vizsgált pozícióban átszótagolásra alkalmas kontextust nem tartalmazó felvétel (például: „A **zá**rban benne van a kulcs?") jelentette az alapot, melyre az esetlegesen átszótagolt hang- vagy szókapcsolatot a másik felvételből vágtuk át (például: „**Az árban** [benne van a kulcs?]" + „[A zárban] **benne van a kulcs?**"). Az előzőek mellett helyet kapott a kérdőívben két, átszótagolásra alkalmas szókapcsolatot tartalmazó mondat is (nem szerkesztett és nem kétértelmű), melyek csak az esetleges átszótagolás megítélésének vizsgálatát hivatottak segíteni. Egy mondat oly módon manipulált, hogy a mondatba illő, átszótagolásra nem alkalmas szókapcsolat („e **gyé**r") helyett csak egy másik mondatban értelmes átszótagolható kifejezés lett beillesztve („e**gy é**r"), azonban végül ez a mondat nem képezte részét az elemzéseknek, mert a kitöltők számára nehéznek bizonyult az értelmezése. Végezetül pedig a fennmaradó két mondat ismétlődés, a kitöltés következetességének tesztelése érdekében.

A kérdőívben a beágyazott 15 hangfelvétel úgy követi egymást, hogy hasonló mondatok illetve a mondatpárok ne kerüljenek egymás mögé. Minden esetben arra kértük a kitöltőket, hogy az adott

felvétel meghallgatása után írják le, mit hallottak, illetve ítéljék meg, az adott mondat mennyire természetes, magyartalan, mindennapi, elfogadható, furcsa, helytelen és tipikus. Az értékelés öt paraméter alapján történt (teljes mértékben, inkább igen, inkább nem, egyáltalán nem, nem tudom). Kérdőívünket 200 fő töltötte ki, többségben nők (n=152, 76%), illetve a legtöbb adatközlő a 19-25 éves korosztályt képviseli (n=116, 58%), ezt követi a 26-29 (n=28, 14%), és a 30-39 (n=23, 11,5%) évesek csoportja. Mind a származási hely, mind pedig az aktuális lakhely tekintetében (sorrendben) Budapest, Pest megye és Heves megye bizonyult a leggyakoribbnak. A kitöltők 89%-a (n=178) megkezdte, vagy már befejezte alap- illetve mesterképzését felsőoktatási intézményben, viszont jelentős részük (n=159, 79,5%) nem részesült korábban nyelvészeti képzésben, vagy ennek tényét nem tudja eldönteni (n=9, 4,5%).

A kapott adatok alapján azt találtuk, hogy a kitöltők inkább az átszótagolt „a zárban” (átlagosan 62,17%) illetve „a zűrben” (49,75%) kifejezést hallották akkor is, amikor valójában „az árban” (33,83%) és „az űrben” (40%) állt a felolvasott mondatban. Ezzel szemben azonban az „akart enni / akar tenni” és a „miért akarod / miért takarod” mondatpárokban az „akart enni” (79,5%) és a „miért akarod” (84%) bizonyult gyakoribb válasznak. Mindez alapján az a tendencia látszik megmutatkozni, mely szerint az átszótagolás nagyobb arányban valósul meg névelőt tartalmazó kifejezésben, mint két fogalomszó esetében. Az „az árban / a zárban” mondatpár esetében kétféle, mondatkezdő és mondatvégi pozícióban is vizsgáltuk a kérdéses kontextust, és az adott válaszok alapján a kitöltők nagyobb arányban vélték az átszótagolt „a zárban”-t hallani a mondat végén (96,5%), mint a mondat elején (45%). A mondatok tartalmának megadása mellett a kitöltőket arra is kértük, hogy lássák el különböző jellemzőkkel a hallottakat. Az eredmények alapján úgy tűnik, hogy nem a felvételek manipulált vagy eredeti mivolta befolyásolta a mondatokhoz hozzákapcsolt pozitív jelzők arányát, hiszen csak egy mondatpár esetében tapasztaltunk jelentősebb eltérést (eredeti: 80,25%, manipulált: 61,15%). Érdemes kiemelni, hogy csak az „az űr / a zűr” mondatpár esetében 50% alatti a pozitív tulajdonságok aránya, ami talán a kevésbé gyakori szavaknak köszönhető. A nem kétértelmű, de átszótagolásra alkalmas kontextusokat tartalmazó mondatokhoz a kitöltők 78,63%-os („Egy alma van a fán”) illetve 91,63%-os arányban („Nem tudom, mit akart inni.”) társítottak pozitív jelzőket.

Mindebből arra következtethetünk, hogy a magyar nyelvben valóban megjelenhet az átszótagolás, azonban az eredmények alapján elsősorban névelőt (is) tartalmazó kifejezésekben, hiszen az ilyen kétértelmű mondatok közül többször is az átszótagolt kifejezést vélték hallani a kitöltők. Az átszótagolás jelenlétét látszik alátámasztani a kérdőívbe beillesztett mondatok akusztikai elemzése is. Azt a tendenciát találtuk ugyanis, hogy hasonlóan egy francia nyelvre vonatkozó kutatáshoz ([3]), a felolvasó rövidebben ejtette az átszótagolható (és ezek szerint átszótagolt) mássalhangzót, mint más környezetben, például szó elején. Ez a mintázat azonban csak az „az” névelő utolsó hangjára vonatkozik. Emellett érdemes azt is megemlíteni, hogy a minimál vagy kontraszív párok alkalmazása több kritikát is kapott már a szemantikai kontextus figyelembe vételének hiánya miatt ([6]), hiszen a párok egyike hátrányba kerülhet amiatt, hogy kevésbé gyakori ([4]), vagy azért, mert az adott környezetben (mondatban) szokatlanul hat, és mindez a kapott eredményeinket is befolyásolhatta.

A kutatás a Kulturális és Innovációs Minisztérium ÚNKP-22-3 kódszámú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

Irodalom

[1] Civit Contra, R. (2018). ‘Del Fonema a la Frase: Resilabificación en la Clase de Español Elemental’. *Kwansei Gakuin University Humanities Review* 23, 205-223 old.

[2] Dressler, W. U. & Siptár P. (1998). ‘A magyar nyelv természetes fonológiája felé’. *Általános Nyelvészeti Tanulmányok XIX.* Szerk. Péter, M. Budapest: Akadémiai Kiadó, 35-39. old.

[3] Fougeron, C., O. Bagou, A. Content, M. Stefanuto & U. Frauenfelder. (2003). ‘Looking for acoustic cues of resyllabification in French’. *15th International Congress of Phonetic Sciences*. Szerk. Solé, M. J., D. Recasens & J. Romero. Barcelona, 2257-2260. old.

[4] Lahoz-Bengoechea, J. M. & M. Jiménez-Bravo. (2020). ‘Spoken word boundary detection in ambiguous resyllabification contexts in Spanish’. Kézirat. *Laboratory Phonology*.

[5] Levin, J. (1985). *A metrical theory of syllabicity*. Doktori disszertáció. Cambridge: Massachusetts Institute of Technology.

[6] Mattys, Sven L., L. White & J. F. Melhorn. (2005). ‘Integration of multiple speech segmentation cues: A hierarchical framework’. *Journal of Experimental Psychology: General* 134.4, 477–500. old. DOI: https://doi.org/10.1037/0096-3445.134.4.477

[7] Steriade, D. (1988). ‘Greek accent: A case for preserving structure’. *Linguistic Inquiry* 19.2, 271-314. old.

[8] Törkenczy, M. & P. Siptár. (2000). ‘Magánhangzó ~ semmi váltakozások a magyarban’. *Nyelvtudományi Közlemények* 97, 64-130. old.

[9] Vennemann, T. (1988). *Preference Laws for Syllable Structure and the Explanation of Sound Change*. Berlin: Mouton de Gruyter.

[10] Vogel, I. (1991). ‘Prosodic phonology: second language acquisition data as evidence in theoretical phonology’. *Cross Currents in Second Language Acquisition and Linguistic Theory*. Szerk. Huebner, T. & C. A. Ferguson. Amszterdam-Philadelphia: John Benjamins Publishing Company, 47–65. old.

The Budapest Games Corpus

Katalin Mády[1], Anna Kohári[1], Uwe D. Reichel[1], Ádam Szalontai[1], Péter Mihajlik[1,2]

[1]Hungarian Research Centre for Linguistics
[2]Budapest University of Technology and Economics

Introduction. Large speech corpora are inevitable for the application of classical machine learning and novel deep learning techniques, along with quantitative analyses of phonetic and linguistic variation. Two well-known Hungarian examples of such corpora are the BEA Hungarian Spoken Language Database and the HuComTech corpus [2, 5, 4]. Next to these, two smaller speech corpora have emerged recently: the Akaka Maptask Corpus [6] and the Budapest Games Corpus. Both contain dialogues in which the partners cooperate to achieve a common goal. The task-oriented design makes it possible to enhance spontaneous interactions and pragmatically varied utterances including emotion expressions such as surprise, disbelief, joy, irony etc. Another advantage is the fixed context that helps annotators and data analysts to control for speaker intention without knowing the participants' personal background. The design was created based on the Columbia Games Corpus [1] in order to establish contexts with varying types of foci (borad, narrow, contrastive). In order to motivate a lively dialogue, the description task was turned into a competition between the groups.

Design. The paradigm was adapted from the Object Game subset of the Columbia Games Corpus [1, 3]. The dialogue partners were seated in front of a laptop in a nonechoic chamber being divided by a board in order to prevent eye contact between them. Both participants saw a dashboard on their screen with 5–7 objects, one of which was placed among the other objects on the first speaker's screen, while it appeared on the bottom of the second speakers screen.

One of the main goals for creating the corpus was to study prominence realisations in spontenous utterances with various information structures (e.g. broad, narrow and contrastive focus) with the names of objects as reocurring words. All target words contained sonorous consonants (sonorants or voiced obstruents) in their stressed syllables to allow for reliable f0 measurements in the potentially prominent syllables which are always word-initial in Hungarian. The final set of 35 objects was chosen according to a pre-test in order to make sure participants associated the expected words with the images presented, including targets being 1 to 4 syllables long. Some objects were intended to trigger contrast marking, e.g. *narancssárga nyíl, lila nyíl* 'orange/purple arrow' (see Fig. 1).

The game was divided into a practice unit and 3 target blocks for each pair of speakers. At the beginning of each block, the Describer was asked to name all visible objects in order to make sure players used the target expressions during the block rather than mistaking a *rénszarvas* 'reindeer' for a *jávorszarvas* 'elk'. If participants agreed on an object name different from the target, the experimenter pointed them to the expected name.

In Block 1, Speaker 1 took the role of the Describer, who was asked to describe the position

of the blinking object to the partner in relation to the other objects to Speaker 2, the Follower. After the Follower was confident they placed the target object into the correct position, they pressed the button FINISH. The proportion of the overlap of the two objects on the two screens was calculated in pixels, allowing for a maximum of 100 points (100% overlap) in each turn, with the score saved in a separate file. Participants switched roles after the first block of 4 turns with new objects they had not seen previously. The third block contained 6 further games with shifting roles, the set of objects being identical for the 12 pairs of speakers. The role of the Describer and the Follower was equally distributed between Speaker 1 and Speaker 2 throughout the altogether 14 games. The first two blocks with 4 games each were varied between the two sessions a speaker participated in.

12 speakers forming four groups participated in the recordings (5 females, 7 males). Within each group, the three participants were relatives or close friends, and they belonged to the same age group (∼20, 30, 40 and 60 years). All speakers were invited twice to play the game with one or the other member of their group. The team with the highest score received double remuneration in order to motivate participants to make their descriptions as precise as possible during the game, in turn resulting in a more realistic and involved language use.

Each speaker was equipped with an omnidirectional head-mounted microphone (DPA 4066-F), and they were recorded on two separate channels with a sampling rate of 44.1 kHz, 16bit. Recordings took place in the Speech Acoustics Laboratory of the Research Institute for Linguistics, Hungarian Academy of Sciences, in 2015. The database contains 8 hours and 38 minutes of dialogues.

Annotation. Annotations based on standard Hungarian orthography were aided by automatic speech recognition described in [8] on the phrase level, followed by manual corrections in Praat's TextGrid format on three separate levels for Speaker 1 (SPK1), Speaker 2 (SPK2) and noise (NOI), including external noise and instructions by the experimenter who gave instructions via a loudspeaker located in the recording booth from the monitoring room.

Realisations diverging from the lexicon, i.e. mispronounced, foreign words or proper names with irregular spelling were given both according to the phoneme sequence and the written form of the word, the latter in square brackets (i.e. *kvescsön [question]*). Marks for non-lexical signals were added manually in <>, such as laughter, hesitation and grunt (see [7] for a clustering procedure), along with the punctuation marks *.,?!* that our ASR system did not output.

Wav files, manual phrase and automatic word and sound segmentations are available under `phon.nytud.hu/voxbox/corpora/databases.html` and are being extended continuosly. Access to the corpus is free, but restricted to research purposes.

Acknowledgements. The creation of the corpus was funded by the National Research and Development Fund (NKFIH), grants PD 1010150, K 135038 and 143075. The idea for creating a Hungarian version of the Columbia Games Corpus by Štefan Beňuš is highly appreciated, along with Julia Hirschberg and Agus Gravano for providing and assisting us with their Java software.

Figure 1: Players' screens, left: the Describer's screen showing a blinking image of raspberries, right: the Follower's screen displaying the raspberry on the bottom panel along with other objects. Instruction left: *Describe the position of the blinking object*, right: *Place the object on the screen.*

References

[1] Online. http://www1.cs.columbia.edu/~agus/games-corpus/.

[2] Gósy, M. (2008). 'Magyar spontánbeszéd-adatbázis – BEA'. In: *Beszédkutatás*, pp. 194–207.

[3] Gravano, A. et al. (2007). 'On the role of context and prosody in the interpretation of 'okay''. In: *Proc. 45th Annual Meeting of Association of Computational Linguistics*. Prague, pp. 800–807.

[4] Hunyadi, L. et al. (2012). 'Az ember–gép kommunikáció elméleti-technológiai modellje és nyelvtechnológiai vonatkozásai'. In: *Nyelvtechnológiai kutatások*. Ed. by Prószéky, G. & T. Váradi. Akadémiai Kiadó, pp. 265–309.

[5] Mihajlik, P. et al. (2022). 'BEA-Base: A benchmark for ASR of spontaneous Hungarian'. In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 1970–1977. URL: `https://aclanthology.org/2022.lrec-1.211`.

[6] Molnár, C. (2023). 'The Akaka Maptask Corpus'. In: *Beszédkutatás – Speech Research Conference*.

[7] Reichel, U. D., A. Kohári & K. Mády (2023). 'Acoustics and prediction of non-lexical speech in the Budapest Games Corpus'. In: *Beszédkutatás – Speech Research Conference*.

[8] Varga, Á. et al. (2015). 'Automatic close captioning for live Hungarian television broadcast speech: A fast and resource-efficient approach'. In: *International Conference on Speech and Computer*. Springer, Cham, pp. 105–112.

Creaky Voice via Speaker Adaptation within End-to-End Text to Speech Synthesis

Ali Raheem Mandeel, Mohammed Salah Al-Radhi, and Tamás Gábor Csapó
Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
{alimandeel, malradhi, csapot}@tmit.bme.hu

Text-to-Speech Synthesis (TTS) aims to create human-like speech signal from input text. Speech synthesis has been based recently on end-to-end approaches. These approaches can develop high-quality speech (e.g., in terms of naturalness, expressiveness, or similarity to the original speaker). These models fall under low-quality speech when the dataset used to train them is insufficient. One key solution to unraveling these issues is utilizing speaker adaptation (also known as speaker customizing or cloning). It adapts a TTS model trained on a large dataset to a target speaker's smaller dataset.

Creaky voice (also called irregular voice, glottalization, or vocal fry) often occurs in ordinary speech communication due to the irregular vibration of the human vocal folds [1]. It appears more frequently in the speech of young female American English speakers [2]. The presence of creaky voices may alter the expression's perceived or emotional content. Creaky voice modeling was proven to enhance the naturalness of synthetic speech [3]. Specifically, synthetic speech with creaky voices is more natural and similar to the natural speaker than conventional synthetic speech.

Based on the displayed creakiness directions, we hypothesize that vocal fry does not appear as a vocal disorder but as a vocal experience to be considered in synthesizers. A few research studies attempted to develop TTS models to synthesize speech with creaky voice, such as Hidden Markov model (HMM) based TTS has been expanded with creaky voice modeling [3]. However, with modern end-to-end TTS systems, creaky voice synthesis with a limited target speaker dataset has not been investigated before.

In this paper, we modeled the end-to-end pretrained FastSpeech 2 model (trained on a large single-speaker female English dataset (LJSpeech)) with a creaky voice using a limited target speakers' dataset of only 100 sentences from the Hi-Fi multi-speaker dataset. We used the pretrained HiFi-GAN neural vocoder which is fast, effective, and high-Fidelity Speech Synthesis. Moreover, we applied Montreal Forced Aligner (MFA) trained with English US ARPA to align the phonemes to the texts. Additionally, we utilized NVidia Titan X GPU for training FastSpeech2 on the target datasets. The adaptation datasets were divided into a training set (95%) and the rest was kept as a validation set.

We adapted FastSpeech 2 to four target speakers (two females and two males). We used three dataset types: 1) frequent creaky voices, 2) randomly chosen sentences, and 3) sentences with few creaky voices. We selected the adaptation sentences based on a creakiness percentage (see Table 1) that was measured automatically [4]. Ten sentences were created in synthesis for each target speaker and data selection scenario (a total of 120 sentences). We conducted both objective and subjective evaluations to estimate how closely the synthesized speech matched the ground truth (references).

The creakiness percentage, Jitter, Shimmer, mean F0, and the Harmonic to Noise ratio (HNR) were measured in the synthesized sentences and compared to the original utterances. Fig. 1 illustrates speech samples for the sentence "And conduct him to the school." from the first

female target speaker, showing the creakiness percentage in the original and the synthesized speech. The objective evaluation showed that modeling a creaky voice in the end-to-end model using speaker adaptation with a dataset of 100 sentences is successful when a significant amount of creaky voice is in the adaptation data.

We used an online MUSHRA- like test to determine which altered version is closest to the target speaker's speech. Each utterance was assessed by 16 non-native individuals (all in a silent place, nine males and seven females, average age of 36). The subjective evaluation revealed that the synthesized creaky voices received lower ratings in terms of similarity and naturalness to the reference utterances compared to synthesized speech on the regular dataset (sentences with few creaky voices).

In conclusion, the objective metrics have demonstrated synthetic speech has noticeable irregular voice patterns when the adaption dataset contains a considerable proportion of creaky voices. According to the subjective listening test, the output speech from the fry-adapted model received a lower evaluation rate in terms of similarity to the original target speaker than the output speech from the regular-adapted model. This inconsistency may result from the synthesized sentences on the adaptation dataset of frequent creaky voices having an excessive degree of creaky voice (more than is necessary). Another explanation could be that the listeners were non-native English speakers, hence less sensitive to the peculiar English voice patterns. Our findings help develop expressive, natural, and customized speech synthesis. Finally, more investigation of the acoustic features of the produced irregular voice samples, subjective evaluations with native listeners, and comparisons with signal processing-based creaky voice synthesis algorithms are also planned for future work.

Table 1: Creakiness percentage on the three adaptation datasets.

Speaker	frequent creaky voices dataset	Randomly chosen dataset	few creaky voices dataset
F1	25.51%	9.23%	0%
M1	5.39%	0.38%	0%
F2	21.38%	0.57%	0%
M2	13.50%	0.49%	0%

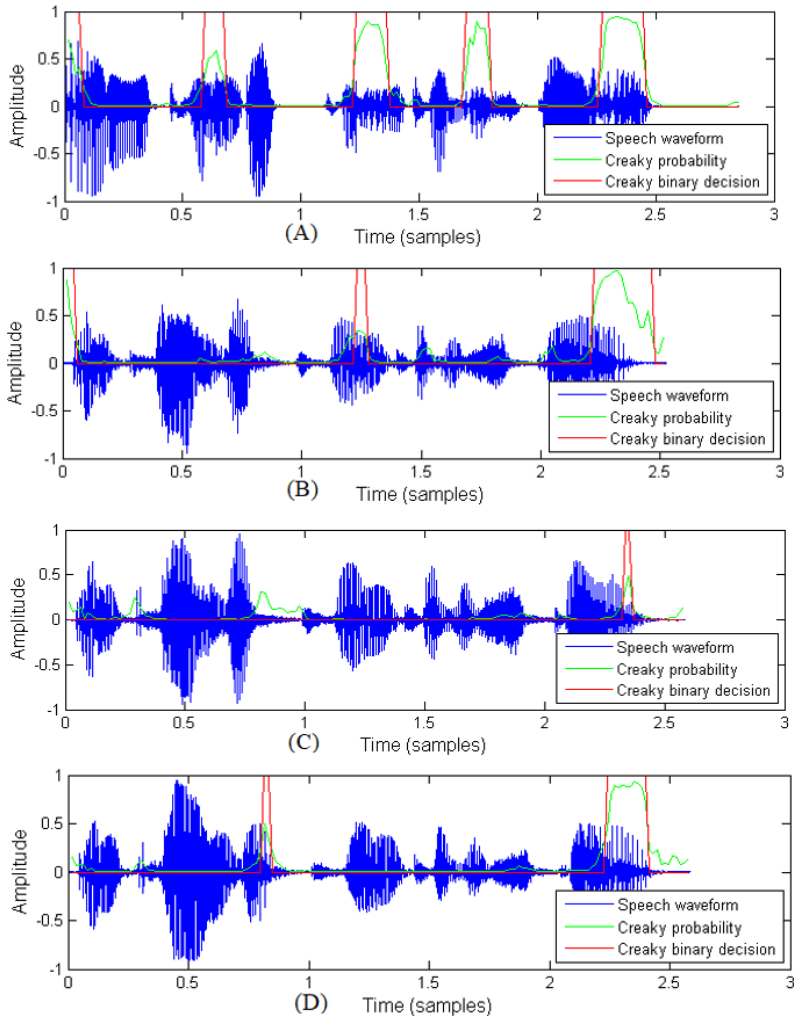


Figure 1: Speech samples from a female speaker: a) natural, b) synthesized using the creaky dataset (dataset 1), c) synthesized using few creaky voices dataset (dataset 3), d) synthesized dataset using the random dataset (dataset 2).

References

[1] J. Laver (1980). 'The Phonetic Description of Voice Quality'. Cambridge Univ. Press.

[2] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin (2012). 'Habitual Use of Vocal Fry in Young Adult Female Speakers'. Journal of Voice, vol. 26, no. 3, pp. e111–e116.

[3] N. P. Narendra and K. Sreenivasa Rao (2017). 'Generation of creaky voice for improving the quality of HMM-based speech synthesis'. Computer Speech & Language; vol. 42, pp. 38-58.

[4] J. Kane, T. Drugman, and C. Gobl (2013). 'Improved automatic detection of creak'. Comp. Speech & Lang., vol. 27, no. 4, pp. 1028-1047.

The Akaka Maptask Corpus

Cecília Sarolta Molnár[1], Katalin Mády[1], Péter Mihajlik[1,2], Beáta Gyuris[1,3]

[1]Hungarian Research Centre for Linguistics
[2]Budapest University of Technology and Economics
[3]Eötvös Loránd University, Budapest

Goal. The aim of this presentation is to introduce a new Hungarian spoken language corpus, the Akaka Maptask Corpus (AMC), which is available for research purposes for free. Access is provided via the website https://phon.nytud.hu/voxbox/corpora/databases.html.

Motivation. Spoken language corpora currently available for Hungarian contain mainly interview questions produced by the linguistic fieldworker or experimenter whose aim is to set up a theme and make the interviewee talk. The motivation for creating the AMC was to have a corpus that enables the study of various form types used for asking canonical and non-canonical questions (cf. [2]) in Hungarian, in contexts where the values for well-known parameters of bias can easily be identified. These parameters include the contents of the common ground ([8]), including the type of evidence available in the context, the private knowledge of the participants, and their aims and preferences. Cf. [9, 3].)

Realization. The dialogues in the AMC were recorded while pairs of speakers were performing a collaborative game. The game is a type of map task [1] designed by the first author. According to the story, archeologists have found several stone tablets in a cave system under the Caucasus Mountains, containing inscriptions that belong to an undeciphered writing system. They need to be investigated by scholars after being transported to a particular town in the area as soon as possible. One of the participants plays the role of a researcher situated in the cave system, and receives a map of the caves (Figure 1), while the other one plays the role of a researcher situated outside the caves, and receives a map of the surface (Figure 2). It is the "researcher" outside the caves who knows where the destination is, so she has to guide the one in the caves to the appropriate exit. The task involves setting up correspondences between the two maps, localising the position of the participant "in the caves", and finding an optimal path through the cave system to the destination. The game finishes when the "researcher" in the caves has reached the destination.

The corpus currently consists of 23 dialogues of about 5 hours, recorded with separate cardioid head-mounted microphones and on separate channels for each speaker in a silent environment (sample rate: 44.1 kHz). During the recording, the participants could not see each other or each other's maps. Recordings were made in 2019, with two different groups of teenagers and young adults: 10 dialogues were produced by speakers of Hungarian living in Romania (mean age: 19.1; 13 women and 7 men), and 13 dialogues by speakers from Budapest (mean age: 21.8; 11 women and 15 men).

Transcription and annotation. Sound files were transcribed with the help of an automatic speech recognizer software into TextGrid format (ASR; [5]). These transcriptions were manually corrected in two rounds by trained annotators and researchers following the BEA transcription and annotation guidelines [6].

Advantages. An advantage of the AMC is that it contains numerous interactions with

short turns, representing a large variety of initiative and reactive moves, and canonical or noncanonical question acts. Since the "universe of discourse" is restricted, it can easily be determined what constitutes the shared vs. individual knowledge of the participants, what their preferences are, and what visual and auditory types of evidence are available for them. This makes it possible to identify relevant parameters for the analysis of bias. In addition, the speech recorded in the AMC is truly spontaneous: participants concentrate on the goal of the game and not on the linguistic form of their production.

The AMC currently contains 1137 utterances marked with a question mark (i.e. that the annotators judged as "questions" in the given context), representing constituent interrogatives, positive and negative polar interrogatives, interrogatives with the *-e* question particle [3], as well as declaratives with a multiple rise-fall intonation contour ("declarative questions"; [4]), tag question constructions with *ugye* and other tags [7] in addition to a high number of fragments.

Disadvantages. Since the recordings were not made in laboratory conditions, they contain some environmental noise. However, this makes the corpus ideal for being used for testing the efficiency of ASR tools.

Acknowledgements. The creation of the corpus was funded by the National Research and Development Fund (NKFIH), grants K 115922 and K 135038.

References

[1] Anderson, A. H. et al. (1991). 'The HCRC Map Task Corpus'. In: *Language and Speech* 34.4, pp. 351–366. DOI: https://doi.org/10.1177/002383099103400404.

[2] Farkas, D. F. (2022). 'Non-intrusive questions as a special type of non-canonical questions'. In: *Journal of Semantics* 39.2, pp. 295–337.

[3] Gyuris, B. (2017). 'New perspectives on bias in polar questions: A study of Hungarian *-e*'. In: *International Review of Pragmatics* 9, pp. 1–50.

[4] — (2019). 'Thoughts on the semantics and pragmatics of rising declaratives in English and rise-fall declaratives in Hungarian'. In: *K + K = 120. Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*. Ed. by Gyuris, B., K. Mády & G. Recski. MTA Nyelvtudományi Intézet, pp. 247–280.

[5] Mihajlik, P. et al. (2022). 'BEA-Base: A Benchmark for ASR of Spontaneous Hungarian'. In: *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*. Ed. by Calzolari, N. Marseille: European Language Resources Association (ELRA), pp. 1970–1977.

[6] Mihajlik, P. et al. (2023). 'A BEA továbbfejlesztése és alkalmazása kontrasztív gépi beszédfelismerési kísérletekre'. In: *Általános Nyelvészeti Tanulmányok XXXIV. Fonetikai tanulmányok*. Ed. by Mády, K. & A. Markó. Budapest: Akadémiai Kiadó, pp. 361–380.

[7] Molnar, C. (2019). 'Speciális kérdések? – Az *ugye* partikulát tartalmazó megnyilatkozások formája és használata'. Doctoral Dissertation. Eötvös Loránd University, Budapest.

[8] Stalnaker, R. (2002). 'Common Ground'. In: *Linguistics and Philosophy* 25.5/6, pp. 701–721. DOI: 10.1023/a:1020867916902.

[9] Sudo, Y. (2013). 'Biased polar questions in English and Japanese'. In: *Beyond expressives. Explorations in Conventional Non-truth-conditional Meaning*. Ed. by Gutzmann, D. & H.-M. Gärtner. Leiden: Brill, pp. 275–295.

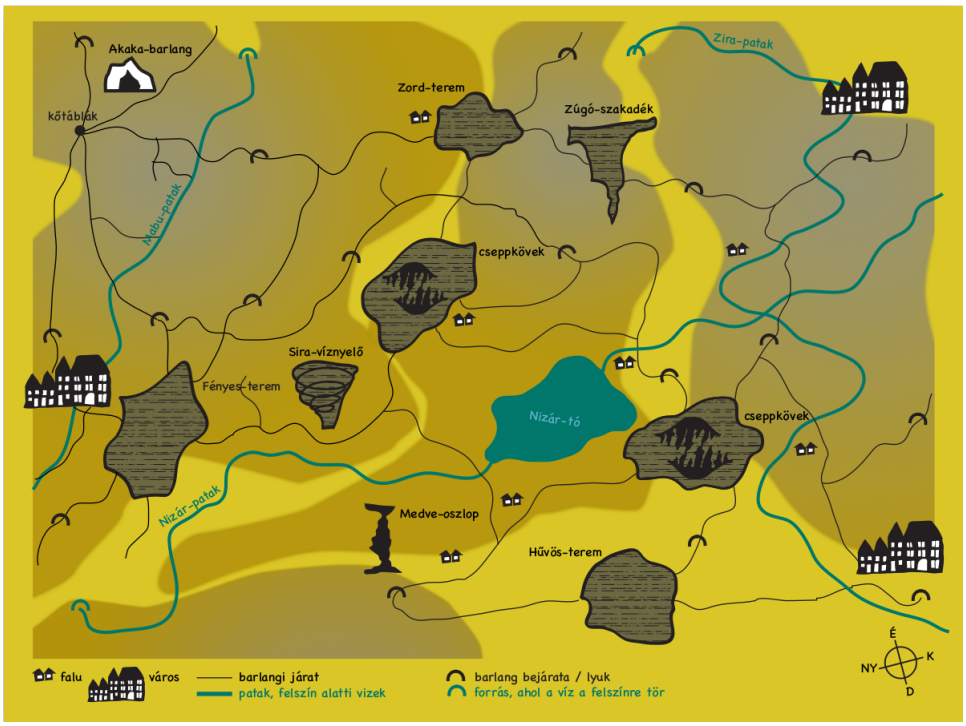


Figure 1: Map of the cave system. (Design: eszterwolf.com.)

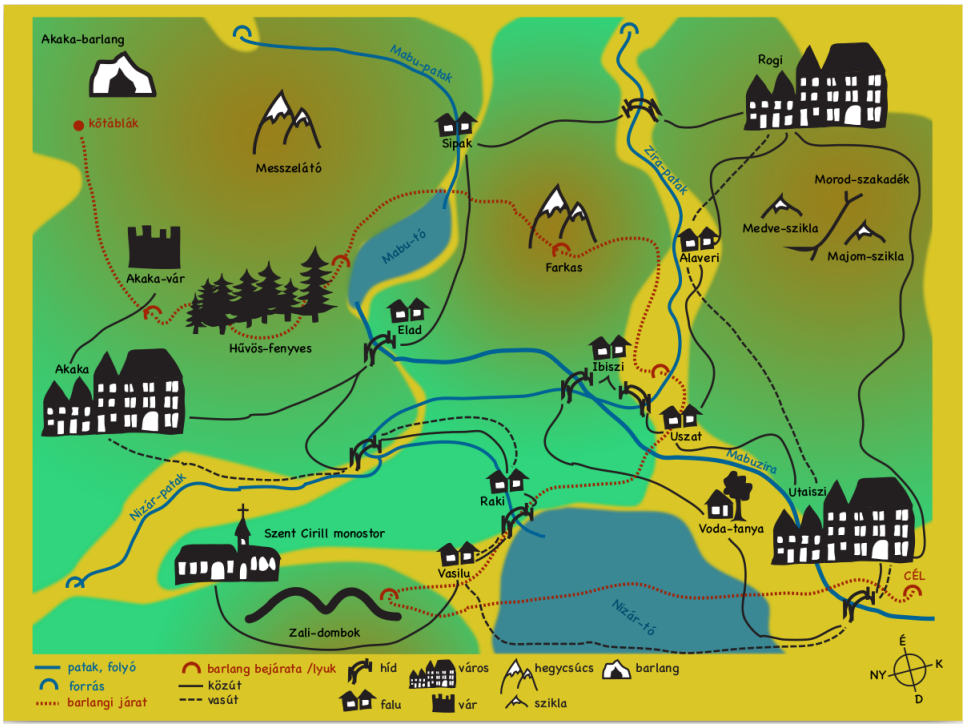


Figure 2: Map of the surface. (Design: eszterwolf.com.)

The role and functional motivation of High target tones in Tundra Nenets

Nikolett Mus, Katalin Mády

Hungarian Research Centre for Linguistics, Budapest

Aims and claims In this paper we examine High target tones in Tundra Nenets – a Uralic language that belongs to the Samoyedic branch of the language family – and attempt to explain their function. High tones can occur in three positions within words in Tundra Nenets: (i) in **initial** position; (ii) in **medial** position, e.g. in the second syllable of a 3-syllabic word; and (iii) in **final** position. Patterns in (i) and (ii) have already been observed and described in the language, while the one in (iii) has not been reported so far [3, 2]. In general, High target tones can refer to several sources of functional motivation, among others word-level prominence (lexical stress) or phrase-final boundary marker (boundary tone). Since empirical studies on Tundra Nenets prosody are rare, we try to link the observed High tones to potential prosodic and syntactic functions using an exploratory approach.

Background Previous observations suggest that primary word-level stress in Tundra Nenets usually falls on the initial syllable, and so the language exhibits bisyllabic trochaic (heavy + light) feet aligned to the left. According to [3, 2], the trochaic structure leads to secondary stress on the odd syllables (3rd, 5th, etc.). Given that secondary stress is not associated with tones, they will not be considered further in this paper. In a few loanwords in which a *schwa* appears in the initial syllable, primary lexical stress is shifted to the next, i.e. second syllable. According to descriptions provided by [3, 2], the final syllable is always excluded from the domain of stress assignment.

Data were recorded with a 31-year old male native speaker of Tundra Nenets. He speaks the Yamal dialect of Tundra Nenets that is the only dialectal variety of the language which has been systematically described so far. Given the limitations of accessible data for our study, we regard our findings as preliminary observations.

Recordings took place in a quiet room with a digital portable recording device (Zoom H4n) and a cardioid head-mounted microphone (Beyerdynamic TG H54c) with a sample rate of 44.1 kHz. The material contained read sentences and spontaneous speech. High tones were manually labelled in Praat and discussed among the two authors in detail. The examples presented below are taken from a 93 seconds long route description monologue.

Observed patterns with non-final High tones As was said before, the second syllable of a word can be stressed in non-native words in which there is a *schwa* in the first syllable [3, 2]. Our material contains some counterexamples to this observation.

In the word *xarad-mʔ* 'house-ACC', for instance, the first syllable does not contain a *schwa*, but it is the second syllable that carries a High tone signalizing stress (see Fig. 1a for the intonation pattern). Similarly, the word *tarkxajuta* 'by the fork (of the river)' – that is a native word – shows a similar pattern to non-native words like *păraśinə̆* 'tarpaulin' having the stress signalised by a High tone on the second syllable (see Fig. 1b). As will be shown in

the next section, the presence of a High tone in this position cannot be explained by phrase structure alone.

Observed patterns with final High tones Some occurrences of word-final High tones were followed by a pause. These pauses do not carry other signals of hesitations, thus, they are most likely to be interpreted as phrase-final boundary markers. This pattern is observed in various syntactic positions. The first case is the syntactic boundary of coordinate clauses such as (1) (Fig. 2a).

(1) *tańćan-da tăńa, ńo-nda śi tăńa.*
stairs-3sg exist.3sg door-gen.3sg hole exist.3sg
'It has stairs, it has a door.'

Second, a High boundary tone (optionally also followed by a silent pause) appears after non-finite embedded clauses, e.g. (2) (Figure 2b).

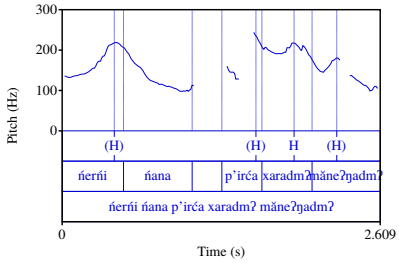
(2) *tarka-ʔ pija-m măda-ʔma-xadańi ŋańiʔ ŋoxolir-ć jăxa-ʔ ńabí tarka-mʔ măda-w.*
fork-gen end-acc pass-an-abl.1sg again swim-cvb river-gen other fork-acc pass-sg.1sg
'After I crossed the end of the fork, I swam across the other branch of the river again.'

Note that non-finite embedded clauses in Tundra Nenets usually precede the finite verb of the main clause.

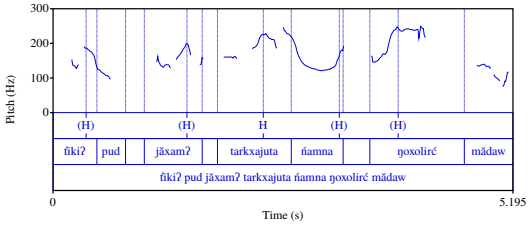
Additionally, topicalized phrases that appear at the left periphery of the clause can also be followed by a final High boundary tone (and an optional silent pause; see *jăxa-mʔ* 'river-ACC' in Fig. 3a). Furthermore, we assume from the context that the phrase *pulʔ ńimńa* 'over the bridge' with a final High tone in Fig. 3b acts as a contrastive focus. Thus, a final High tone may also mark a focussed phrase.

The examples presented here illustrated clear instances of the function of a final High boundary tone. There are further observed patterns in which the function of the High tone is less clear. For instance, the High tone may occur in a syllable containing a glottal stop. As suggested in the literature [1], the glottal stop may result in a High boundary tone in some languages, however, this phenomenon has not been documented in Tundra Nenets so far.

Acknowledgements The support of the research project "Theoretical and experimental approaches to dialectal variation and contact-induced change: a case study of Tundra Nenets" by the National Research and Development Fund (NKFIH FK 129235) is gratefully acknowledged.

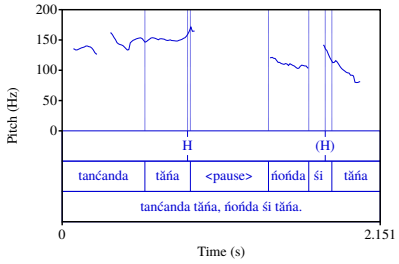


(a) 'I saw a high house in front of me.'

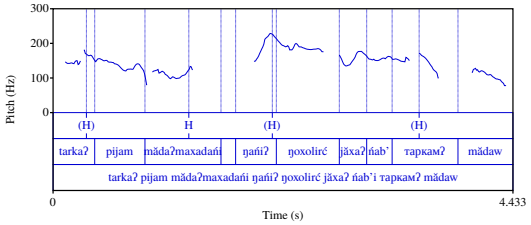


(b) 'After that, I swam across the river by its fork.'

Figure 1: Non-final High tone

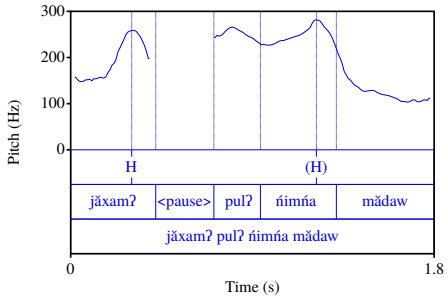


(a) 'It has stairs, it has a door.'

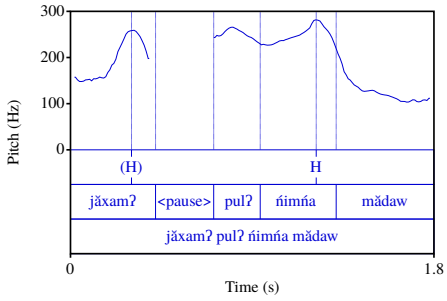


(b) 'After I crossed the end of the fork, I swam across the other branch of the river again.'

Figure 2: Final High tone in coordination and subordination



(a) 'I crossed [$_{\text{TopP}}$the river] over the bridge.'



(b) 'I crossed the river [$_{\text{FocP}}$over the bridge].'

Figure 3: Final High tone marking TopP and FocP

References

[1] Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

[2] Nikolaeva, I. (2014). *A grammar of Tundra Nenets*. Berlin/Boston: Mouton de Gruyter.

[3] Salminen, T. (1998). 'Nenets'. In: *The Uralic languages*. Ed. by Abondolo, D. London: Routledge, pp. 516–547.

Mi a fonológiai komplexitás?

Rebrus Péter[1,2] és Szigetvári Péter[2]
[1]Nyelvtudományi Kutatóközpont
[2]Eötvös Loránd Tudományegyetem

A ***fonotaktikai*** mintázatok, azaz a nyelvben létező vagy elfogadható hangszekvenciák leírása és magyarázata a fonológia egyik központi témája. A fonotaktika a nyelvi tudás része, központi szerepet játszik a szavak grammatikai megítélésében, a kölcsönszavak fonetikai realizációjában, a fonológiai folyamatok kiváltásában, illetve blokkolásában, a szintaktikai és morfológiai elhatárolásban és kategorizációban, ezek mellett a nyelvi rétegekre, illetve a lektusokra vonatkozó információkat is tartalmaz. A nyelvekben megfigyelhető grammatikus szekvenciák egyrészt nem véletlenszerűek, hanem nyelvspecifikus vonások mellett ***univerzális*** összefüggések vonatkoznak rájuk (pl. [4]). Másrészt a fonotaktikai minták határai gyakran nem élesek (a grammatikai ítéletek fokozatosak), és a minták ***graduális***ak (a grammatikus szekvenciákat sokszor nem lehet megadni általánosan, a hangok nagyobb osztályaira alapozva).

A nyelvekben lehetséges mássalhangzó-kapcsolatok (CC) fonotaktikailag komplex rendszert alkotnak; ezek leggyakoribb altípusa a zárhangra (explozíva, affrikáta) végződő kapcsolatok (CT). A CT-kapcsolatok a bennük szereplő első magánhangzó típusa alapján szigorú ***jelöltségi hierarchiát*** mutatnak, amely ***implikációs univerzálé***ban nyilvánul meg: a nyelvekben a jelöltebb típusú kapcsolatok léte maga után vonja a jelöletlenebbeket. A CT-típusonak ezt az univerzális skáláját mutatja (1) – a skála nem tartalmazza a glide-kezdetű és gemináta-kapcsolatokat, mivel azok viselkedése eltérő.

A fenti jelöltségi skála ***nyelvtípusok***at definiál, az alapján, hogy az adott nyelvtípusban mi a legjelöltebb elérhető CC-típus. A 6 lehetséges nyelvtípus a CT-ket egyáltalán nem tartalmazótól (pl. hawaii) a közbenső típusokon át – pl. japán (csak NT), szidamó (NT és RT), olasz (NT, RT, ST), latin (NT, RT, ST, PT) – az összes CT-típust megengedő nyelvtípusig (pl. hindi, német) terjed.

Az egyes CT-típusokon belül előforduló mássalhangzó-kapcsolatokra a nyelvekben több egyéb megszorítás vonatkozhat (homorganitási kényszer vagy tilalom, a szillabikusok kontaktjának tilalma, a koronálisok preferálása stb). Ennélfogva gyakran egy adott típusban a kapcsolatok előfordulása nem teljes: az összes lehetséges konkrét kapcsolat nem feltétlenül létezik. Újabb vizsgálatok azt mutatják, hogy az előforduló és az összes elméletileg lehetséges kapcsolat számának az egy típuson belüli ***aránya*** a fenti (1) skálán megadott jelöltség növekedtével ***monoton csökken***. Azaz minél jelöltebb egy CT-típus, annál kisebb arányban vannak benne előforduló különböző mássalhangzó-kapcsolatok: az arány a legjelöletlenebb típusnál tapasztalt 100%-tól (minden kapcsolat létezik) a már nem elérhető típus(ok) 0%-áig terjedhet úgy, hogy a közbenső típusokban az arány akármennyi lehet, betartva a monotonitást.

A konkrét nyelvekre vonatkoztatva ez az (1) skálához tartozó diszkrét tipológia graduális kiterjesztése: a 6 nyelvtípust az adott ***nyelv fonotaktikai profil***ja váltja fel, amelyet a típusokhoz tartozó lecsengő arányok sorozata mutat. A (2) ábra a magyar ilyen profiljait mutatja, ahol az arányokat a CT-k egyre jelöltebb jobboldali környezeteiben (magánhangzós, szóvégi, mássalhangzós) külön számoltuk. Látható, hogy a lecsengő profilok a magyarban a legjelöletlenebb típusoktól (NT-ben prevokalikusan és szóvégen, illetve RT-ben prevokalikusan az arány közel 100%) a legjelöltebb típusokig terjednek (MT-ben mindhárom típus, PT-ben a nem

intervokalikus előfordulás rendkívül csekély: az arány közel 0%).

Az (1) skála értelmezése több szempontból problematikus: egyrészt az nem vezethető le egyéb skálákból, így a szonoritási hierarchiából [1, 6, 8] sem, másrészt a szótagalapú elméletekben a szótagra vonatkozó jólformáltsági szabályokon túl feltételezni kell szótagfűző szabályokat (*syllable contact laws*, [5]) sokszor ad hoc módon. A reprezentacionalista elméletek a jelöltséget egyfajta ***komplexitás***ként értelmezik: minél jelöltebb a CT, annál komplexebb az ábrázolása. Példa erre az értelmezésre a fonológiai pozíciók engedélyezése (pl. [7], illetve a kormányzás-fonológián belül [2]). A legtöbb reprezentációt használó elmélet hátulütője viszont az, hogy az általános elvek nem képesek megragadni a fonotaktika inherensen graduális jellegét. Ezt demonstrálja a (3) ábra, amely a magyarban lehetséges CT-kapcsolatokat mutatja be (minél sötétebb a háttér, az adott CT annál több környezetben jelenik meg, a fehér hátterűek monomorfemikusan nem fordulnak elő). Látható, hogy egyfajta jelöltségi különbség nemcsak a C_1 által meghatározott típusok között van (sorok), hanem a C_2 képzési helyei és módjai között is (oszlopok), ezenkívül a zöngésségnek is szerepe van, ami így többdimenziós graduális mintákat eredményez.

A komplexitás feltételezése azonban nem feltétlenül haszontalan, és a fonetikailag megalapozott fonológiai elméletekkel (pl. [3]) is kompatibilis, azaz artikulációs és/vagy percepciós fonetikai szempontból is értelmezhető. [7] a ***kontraszt***ok fonológiai ***perceptibilitásá***nak univerzális különbségével magyarázza a fonotaktikai jelenségeket: a szonoránsok környezetében a legjobb, az obstruensek környezetében a legrosszabb a felismerhetőség. Ez a skála bizonyos elemeit magyarázza, így azt, hogy miért jelöletlenebbek bizonyos szonoránskezdetű kapcsolatok (NT, RT) az obstruenskezdetűeknél (ST, PT): a C_2 (a zárhang) azonosíthatóságát az előbbi környezet jobban elősegíti. Hasonlóan a C_1 azonosíthatóságát a homorganitás könnyíti meg: az NT típus jobban felismerhető, mint az MT, mivel az előbbiben az egész kapcsolatban a képzési hely állandó, míg az utóbbiban nem – ezt a nyelvekben gyakori aktív hasonulási folyamatok is mutatják, pl. mt > nt, nk > ŋk (ugyanez áll a másik kötelezően heterorgán típusra (PT) is: pt, kt > tt). Ennélfogva a felismerhetőségi nehézség a komplexitással együtt nő, amennyiben a komplexitást a szegmentum-kontrasztok megadásához szükséges ***információ mennyiségével*** értelmezzük (ez homorganitás esetén kevesebb). Ennek a megközelítésnek artikulációs fonetikai, és reprezentációs fonológiai kapcsolódásai is lehetnek. Az előadásban a fonológia–fonetika interakcióját a fenti problémák szempontjából vizsgáljuk.

Hivatkozások

[1] Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In Kingston – Beckman (szerk.) *Papers in Laboratory Phonology I: Between the grammar and the physics of speech.* 283–333.

[2] Harris, J. (1997). Licensing inheritance: An integrated theory of neutralisation. *Phonology* 14: 315–370.

[3] Hayes, B., R. Kirchner &. Steriade szerk. (2004). *Phonetically based phonology.* C.U.P.

[4] Jakobson, R., G. Fant & M. Halle (1952). *Preliminaries to speech analysis.* Cambridge, MIT

[5] Murray, R. W., T. Vennemann (1983). Sound change and syllable structure in Germanic phonology. *Language* 59: 514–528.

[6] Sievers, E. (1876). *Grundzüge der Lautphysiologie zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* Leipzig: Beitkopf und Härtel.

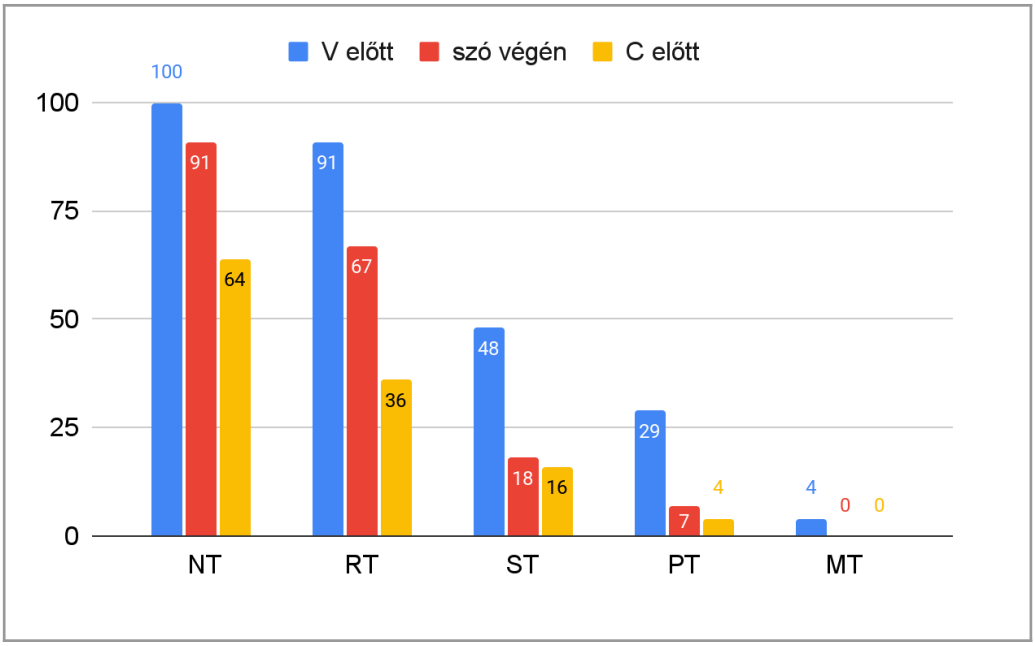
[7] Steriade, D. (1999). Alternatives to syllable-based accounts of consonantal phonotactics. In *Proc. of the 1998 Linguistics and Phonetics Conference: Item Order in Language and Speech.* 205–245.

[8] Zec, D. (1995). Sonority constraints on syllable structure. *Phonology* 12: 85–129.

(1) CT-típusok univerzális jelöltségi skálája

rövidítés:	**NT**	<	**RT**	<	**ST**	<	**PT**	<	**MT**
C_1 jellemzője:	homorgán nazális		likvida		réshang		nem homorgán zárhang		nem homorgán nazális
példák:	nt mp ŋk ɲc		rt rk lt lp		st sp ʃk ft		pt kt pk tk		mt ɲp nk ŋt

(2) A magyarban monomorfemikusan előforduló posztvokalikus CT-k aránya az összeshez képest típusonként különböző jobboldali környezetekben (%)



(3) A magyar mássalhangzó+zárhang-kapcsolatok (a gemináták kivételével)

T=	**Alv.**		**Vel.**		**Lab.**		**Pal.**		**Alv.**	**Pal.**		_V	_#	_C	össz	%		
	zárhang								**affrikáta**							_V	_#	_C
NT	nt	nd	ŋk	ŋg	mp	mb	ɲc	ɲɟ	nʦ	nʧ	nʤ	**11**	**10**	**7**	11	**100**	**91**	**64**
RT	rt	rd	rk	rg	rp	rb	rc	rɟ	rʦ	rʧ	rʤ	**30**	**22**	**12**	33	**91**	**67**	**36**
	lt	ld	lk	lg	lp	lb	lc	lɟ	lʦ	lʧ	lʤ							
	jt	jd	jk	jg	jp	jb	jc	jɟ	jʦ	jʧ	jʤ							
ST	st	zd	sk	zg	sp	zb	sc	zɟ	sʦ	sʧ	zʤ	**21**	**8**	**7**	44	**48**	**18**	**16**
	ʃt	ʒd	ʃk	ʒg	ʃp	ʒb	ʃc	ʒɟ	ʃʦ	ʃʧ	ʒʤ							
	ft	vd	fk	vg	fp	vb	fc	vɟ	fʦ	fʧ	vʤ							
	xt	xd	xk	xg	xp	xb	xc	xɟ	xʦ	xʧ	xʧ							
PT	–		tk	dg	tp	db	tc	dɟ	tʦ	tʧ	dʤ	**16**	**4**	**2**	55	**29**	**7**	**4**
	kt	gd	–		kp	gb	kc	gɟ	kʦ	kʧ	gʤ							
	pt	bd	pk	bg	–		pc	bɟ	pʦ	pʧ	bʤ							
	ct	ɟd	ck	ɟg	cp	ɟb	–		cʦ	cʧ	ɟʤ							
	ʦt	ʣd	ʦk	ʣg	ʦp	ʣb	ʦc	ʣɟ	–	ʦʧ	ʣʤ							
	ʧt	ʤd	ʧk	ʤg	ʧp	ʤb	ʧc	ʤɟ	ʧʦ	–								
MT	–		nk	ng	np	nb	nc	nɟ	–	nʧ	nʤ	**1**	**0**	**0**	24	**4**	**0**	**0**
	mt	md	mk	mg	–		mc	mɟ	mʦ	mʧ	mʤ							
	ɲt	ɲd	ɲk	ɲg	ɲp	ɲb	–		ɲʦ	–								
_V	19		19		15		9		7	10		**79**				**47**		
_#	17		10		6		5		3	3			**44**		167		**26**	
_C	12		8		6		0		2	0				**28**				**17**

Acoustics and prediction of non-lexical speech in the Budapest Games Corpus

Uwe D. Reichel, Anna Kohári, Katalin Mády
Hungarian Research Centre for Linguistics

Introduction This study is about the acoustic analysis and prediction of conversational grunts and filled pauses with the purpose of automatic annotation of the Budapest Games Corpus. Acoustic analyses show, that both types of non-lexical speech can be well distinguished by voice quality, articulatory, and prosodic parameters. In non speech classification experiments we obtain encouraging UARs of .96 and .98 for feature-based and end-to-end models, respectively, which qualifies them for the intended corpus annotation use.

Conversational grunts and filled pauses both are non-lexical and acoustically simple vocal expressions that frequently occur in human dialogs. Conversational grunts (*cg*) are intentionally used for a variety of pragmatic functions like backchannel, agreement, disagreement, uncertainty, and surprise. Filled pauses (*fp*) in turn are expressions of hesitation of the speaker. Tools for automatic corpus labeling like automatic speech recognition or speech-text alignment, the latter being applied in annotating and segmenting the Budapest Games Corpus (BGC, [4]) are generally not trained to distinguish between variants of non-lexical speech. Thus with the final purpose of automatic non-lexical speech labeling, this pilot study aims (1) to detect acoustic differences between *cg* and *fp* in an exploratory data analysis of the BGC, and (2) to run machine learning experiments for non-lexical speech classification.

Acoustics For the acoustic analyses we extracted a large amount of acoustic features from the 1697 *cg* and *fp* segments (327 *cg*, 1370 *fp*) of the manually corrected word-level segmentation of the BGC including 36 dialogues by 12 speakers. The underlying feature sets are openSMILE Compare_2016 [2] and eGeMAPS [3], as well as CoPaSul features [5]. The openSMILE features comprise functionals and summary statistics from robust low-level descriptors for spectral, intonation, and voice quality features. These sets were complemented by CoPaSul features that represent prosodic aspects of the utterance.

In this exploratory data analysis we tested all features for significant differences between *cg* and *fp* by Mann-Whitney tests. A type-1 error correction was not feasible given the huge amount of features. Instead, we further filtered the features by their effect size in terms of Cohen's d, and report only the number of features which differ by at least medium effect size level ($d \geq .5$). For eGeMAPS we found 49 out of 88 features to differ significantly. For ComParE_2016 we found 2721 out of 6373 features. Among the CoPaSul subset 18 out of 61 features turned out to differ significantly. Amongst others, *fp* compared to *cg* is characterized by **voice quality** (lower jitter, shimmer, and harmonics-to-noise ratio), **articulation** (lower mean cepstral frequency coefficients 1–3), and **prosody** (lower f0 and energy variation, and a horizontal – as opposed to rising – f0 midline; see Figure 1).

Modeling We trained several feature-based and end-to-end models to predict *cg* vs. *fp* in a classification task. For this purpose we split the data into a speaker-disjunct training, development and test set by the ratio 80-10-10. The split was stratified for our target variable.

As feature-based models based on eGeMAPS we trained a K-Nearest Neighbor classifier (KNN), a Classification and Regression tree (CART), a Support Vector Machine (SVM), and an Extreme Gradient Boosting classifier (XGB). The hyperparameters of these models were optimized in a 5-fold cross-validation on the merged training and development set by means of Bayesian optimization with a constant random seed. The optimized models were then evaluated on the held-out test set.

For end-to-end non-lexical speech classification we used Transformers. As encoder we took the pretrained multilingual *facebook/wav2vec2-large-xlsr-53* base model with 24 layers [1]. To this base model we added a simple classification head consisting of one feed-forward hidden layer with a *tanh* activation function, and a final output projection layer to 2 classes. Input to this head is the mean pooling over the hidden states of the last encoder layer to which we applied a dropout of 0.1. We finetuned the model on the downstream task in 10 epochs with a learning rate of $2e-4$, the ADAM optimizer with a weighted Cross Entropy Loss, and a batch size of 8. Following the approach of [7], only the encoder weights but not the weights of the initial convolutional feature extractor layers were finetuned. We kept the best model in terms of Unweighted Average Recall (UAR) on the development set and evaluated it on the held-out test set.

We additionally trained an ablated transformer variant with the same learning configuration but removing the final 12 encoder layers. Given the relatively small amount of training data, this serves to find out whether a reduction of the number of model parameters to be fine-tuned helps in improving performance.

Table 1 gives the results of all models in terms of UAR and accuracy. For the best performing model architecture (the ablated Transformer TRANS-12), we repeated finetuning and validation with 5 different random seeds to ensure that the good results were not only accidentally obtained. The best feature-based model was the SVM with a UAR of .96. This model was still outperformed by the best end-to-end model TRANS-12, that yielded a mean UAR of .98.

Discussion and conclusion The exploratory data analysis indicates, that *cg* and *fp* can be acoustically well distinguished from each other in terms of voice quality, articulatory, and prosodic features.

Furthermore, the good classification results with an UAR up to .98 for non-lexical expressions suggest to use the models for automatic corpus annotations. Interestingly, the ablated end-to-end transformer variant TRANS-12 outperformed the big variant TRANS-24. We will further examine by means of probing experiments (see an overview in [6]) whether this performance gain is only due to increased robustness by reducing the amount of model parameters or also due to a better representation of non-lexical aspects of speech in non-final encoder layers.

Acknowledgements This study was funded by the National Research, Development and Innovation Office, grants *K 135038* and *K 143075*.

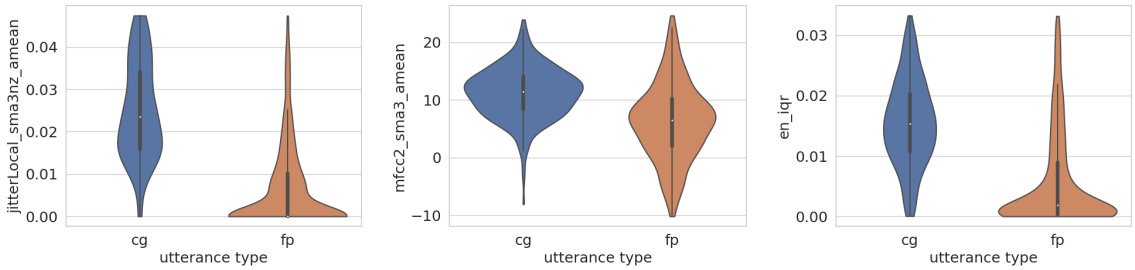


Figure 1: **Acoustic features:** Compared to conversational grunts (cg) filled pauses (*fp*) are amongst others characterized by lower jitter (left), a lower MFCC2 (mid), and lower energy variation (right).

Model	**UAR**	**Accuracy**
CART	.87	.94
KNN	.82	.92
SVM	.96	.96
XGB	.90	.96
TRANS-12	**.98 (0.007)**	**.99 (0.004)**
TRANS-24	.93	.98

Table 1: **Non-speech classification performances** on the test set in terms of UAR and Accuracy. For the best performing model TRANS-12 we give the mean and standard deviation obtained from training and evaluation with 5 different random seeds.

References

[1] Conneau, A. et al. (2020). 'Unsupervised cross-lingual representation learning for speech recognition'. In: *arXiv preprint arXiv:2006.13979*.

[2] Eyben, F. (2015). *Real-time speech and music classification by large audio feature space extraction*. Springer.

[3] Eyben, F. et al. (2015). 'The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing'. In: *IEEE transactions on affective computing* 7.2, pp. 190–202.

[4] Mády, K. et al. (2023). 'The Budapest Games Corpus'. In: *Proc. Beszédkutatás – Speech Research Conference*. Budapest.

[5] Reichel, U. D. (2016). *CoPaSul Manual – Contour-based parametric and superpositional intonation stylization*. https://arxiv.org/abs/1612.04765. RIL, MTA. Budapest, Hungary.

[6] Triantafyllopoulos, A. et al. (2022). 'Probing Speech Emotion Recognition Transformers for Linguistic Knowledge'. In: *Proc. Interspeech*, pp. 146–150.

[7] Wang, Y., A. Boumadane & A. Heba (2021). 'A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding'. In: *arXiv preprint arXiv:2111.02735*.

Improving the expressiveness of TTS synthesis with non-autoregressive neural vocoding

Mohammed Salah Al-Radhi[1], Tamás Gábor Csapó[1], and Géza Németh[1]
[1] Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
{malradhi,csapot,nemeth}@tmit.bme.hu

Speech prosody plays a vital role in communication as it conveys not only the meaning of words but also the emotional and social context of the message. Stress, intonation, and rhythm in spoken words, phrases, and sentences all contribute to the prosodic features of speech [1] [2]. Text-To-Speech (TTS) synthesis systems have been developed to enable machines to generate speech from text. With the advent of new machine learning techniques such as Deep Neural Networks (DNN), these systems have seen significant improvements in recent years. DNN-based acoustic models have been developed to represent the mapping function between the linguistic feature sequences and the acoustic feature sequences. End-to-end neural network-based approaches have also made considerable progress in recent years, leading to an increase in the quality of synthesized speech.

However, neural TTS models are not without their weaknesses. Autoregressive architecture, such as that used in Tacotron2 [3], has been found to be slow and less robust. This is because every frame needs to iterate through all previous time steps to make predictions, which can be computationally expensive and time-consuming. In human-computer interactions, emotional expressions and controllable speech synthesis are critical factors in generating natural responses. Despite the progress made in TTS synthesis systems, the expressiveness of synthesized speech remains an area that requires further improvement [4].

To achieve our goal of creating a highly expressive TTS system, we propose utilizing a non-autoregressive Mel-spectrogram prediction model, such as FastSpeech2 [5], which has demonstrated improved speed and robustness compared to traditional autoregressive models. This approach allows for a more efficient and accurate generation of speech by predicting the acoustic features in parallel rather than sequentially (see Figure 1). Additionally, to further enhance the emotional expressiveness of the synthesized speech, we explore the F0 characteristics of various emotional expressions, such as anger, fear, and happiness [6]. By analyzing the F0 range, correlation, and contour of these expressions, we aim to extract the necessary features to generate emotionally expressive speech that can convey a range of different emotions. Our study reports initial results on the efficacy of our proposed approach for expressive and controllable speech synthesis, demonstrating promising results in generating natural and expressive speech with a high degree of emotional expressiveness.

The validation process provides a critical evaluation of the proposed model's performance, which is essential to ensure the model's effectiveness in generating high-quality synthesized speech. The four different loss functions were implemented (as presented in Table 1) to capture various aspects of the speech signal, such as prosody and acoustic features, and to ensure the proposed model's robustness in handling different speech expressions. The validation loss scores show that the proposed model has achieved significant improvement in performance, indicating the

effectiveness of the model in generating synthesized speech that closely resembles the ground-truth speech. This is particularly important for speech communication applications, where natural and expressive speech is critical for enhancing the user experience.

Furthermore, the visualization of the Mel-spectrograms generated by the proposed model and the ground-truth models provides an additional level of evaluation (see Figure 2). By comparing the two Mel-spectrograms, we can observe that the proposed model generates Mel-spectrograms with more detailed frequency information, which is an indication of the model's ability to capture and reproduce the nuances of the speech signal. The similarity between the Mel-spectrograms generated by the proposed model and the ground-truth Mel-spectrogram further validates the proposed model's effectiveness in generating high-quality synthesized speech. Overall, the results demonstrate the proposed model's capability in producing natural and expressive speech that can be applied in various human-computer interaction applications.

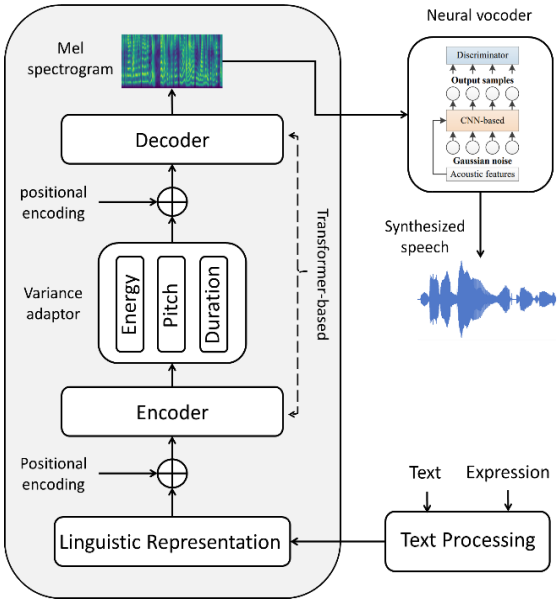


Figure 1: Suggested diagram of expressive speech synthesis.

Table 1: Results of training and validation using different loss functions.

Metrics	Training	Validation
Mel Loss	0.13	0.39
Duration Loss	0.002	0.11
F0 Loss	0.006	0.42
Energy Loss	0.005	1.07
Number of Parameters	27M	

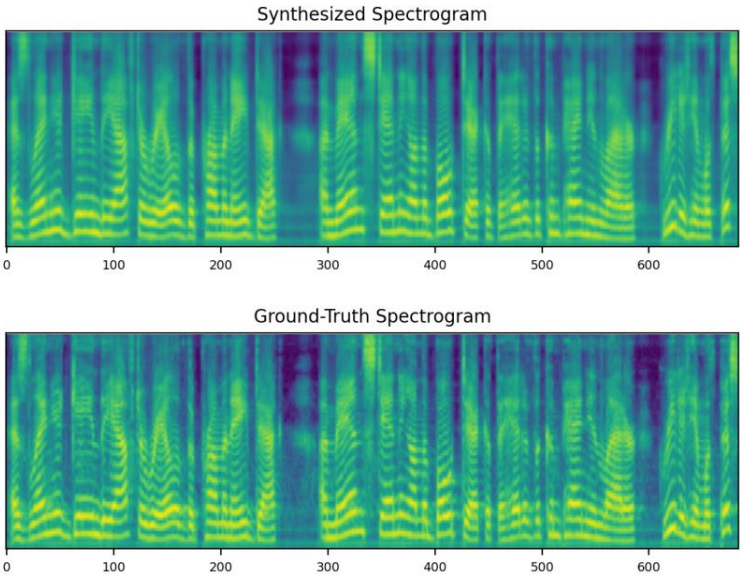


Figure 2: Visualizations of the ground-truth and synthesized Mel-spectrograms.

References

[1] Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, R. J., Weiss, R., Clark, R., Saurous, A. (2018). 'Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron'. In: Proc. *ICML*, Stockholm.

[2] Xiao, Y., He, L., Ming, H., Soong, F.K. (2020). 'Improving Prosody with Linguistic and Bert Derived Features in Multi-Speaker Based Mandarin Chinese Neural TTS'. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6704-6708.

[3] Shen, J., Pang, R., Weiss, J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan R., et al. (2018). 'Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions'. in Proceedings of *ICASSP*, pp. 4779-4783.

[4] Schiller, D., Mertes, S., Rijn, P., Andre, E. (2021). 'Analysis by Synthesis: Using an expressive TTS Model as Feature Extractor for Paralinguistic Speech Classification'. in Proc. of *Interspeech*, Brno, Czechia.

[5] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y. (2021). 'FastSpeech2: Fast and High-Quality End-to-End Text to Speech'. In: *The International Conference on Learning Representations (ICLR)*.

[6] Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S. (2008). 'IEMOCAP: interactive emotional dyadic motion capture database'. Lang Resources & Evaluation, vol. 42, pp. 335-359.

Few-Shot Multi-Language Text-to-Speech Synthesis with State-of-the-Art Neural Networks

Layan Sawalha[1] and Mohammad Salah Al-Radhi[1]
[1] Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics

layan.sawalha@edu.bme.hu, malradhi@tmit.bme.hu

Abstract

This paper investigates the artificial creation of human voices by generating speech waveforms from textual input, known as Text-To-Speech (TTS) synthesis. The motivation is to create a synthetic, understandable speech that is as close to human speech as feasible in multiple languages with limited data. To achieve optimal Text-to-Speech results, we often need a large dataset which causes a barrier when implementing it in different languages. Nowadays, there are almost seven thousand spoken languages with insufficient datasets, which constrains the applicability of TTS. To eliminate these barriers, we introduce Few-Shot Speaker Adaptive Text-to-Speech Synthesis.

In this framework, FastSpeech2 [1] is implemented as a non-autoregressive text-to-speech model; it focuses on extracting pitch, energy, and duration from speech waveform and uses them in training and interference. It was implemented to overcome common TTS issues and provide high-quality speech synthesis faster while avoiding controllability and robustness issues. But in this study, we implement it using the English language and in a different one (ex., Arabic). Arabic Speech Corpus [2] contains a large dataset that consists of 3.7 hours, we implemented only 1.5 hours which is less than half of the original dataset while maintaining high-quality results. In the future, we want to create a system where a user can train and generate speech using a minimal dataset, which will apply to more languages. As depicted in Figure 1, the FastSpeech2 model operates by converting text to speech through several stages. The process begins with the conversion of phoneme embedding using an encoder, which produces a phoneme hidden sequence. The variance adaptor then adds information on pitch, energy, and duration to the phoneme hidden sequence. Subsequently, the Mel-spectrogram decoder converts the adapted hidden sequence into a Mel-spectrogram sequence in a parallel manner. This pipeline of operation allows the FastSpeech2 model to effectively generate high-quality speech output by incorporating the relevant prosodic information. Additionally, the parallel architecture of the Mel-spectrogram decoder is essential in achieving real-time performance, making it a suitable choice for various applications. The effectiveness of this approach has been demonstrated in our experiments on various datasets, where it outperformed other state-of-the-art models in terms of naturalness, fluency, and intelligibility.

In training, the ground-truth value of duration, pitch, and energy is extracted into a hidden sequence to predict the target speech and is also used to train the duration, pitch, and energy predictors which infer to synthesized target speech.

Figure 2 presents the objective results of incorporating the Arabic language into the FastSpeech2 model. A comparison of the spectrogram of the synthesized speech for both full data and limited data is provided. It is evident that the fundamental frequency (F0) and spectrogram of both speeches are consistent. As can be observed in Table 1, the loss values for the limited data are higher than those of the full data. However, this is indicative of a high-quality speech. It is worth mentioning that the limited data is still able to produce a high-quality output which is a promising result. Additionally, we have also compared the results with other state-of-the-art models on the same dataset and found that our proposed model performed better. This demonstrates the effectiveness of our approach in integrating the Arabic language into the FastSpeech2 model.

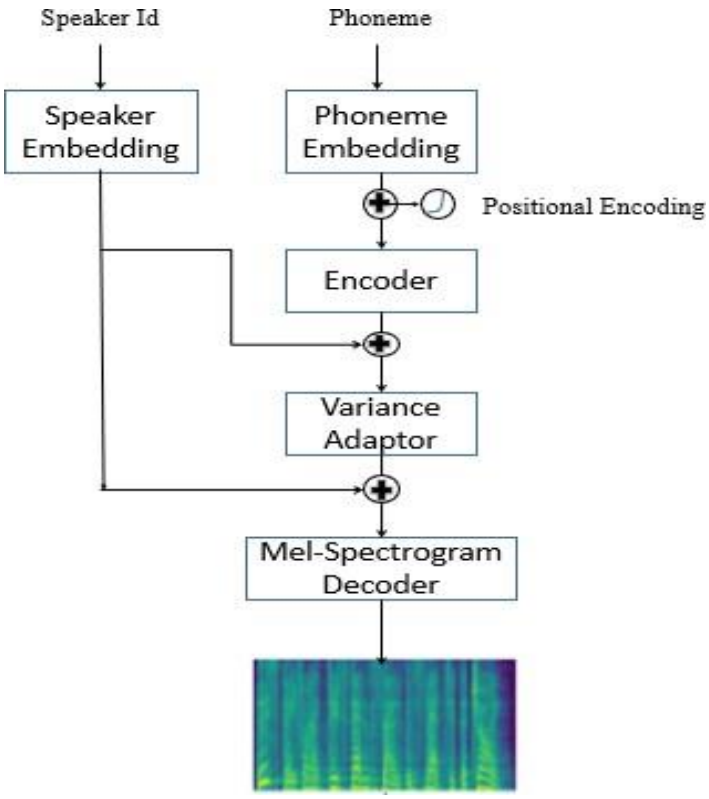


Figure 1: FastSpeech2 Architecture

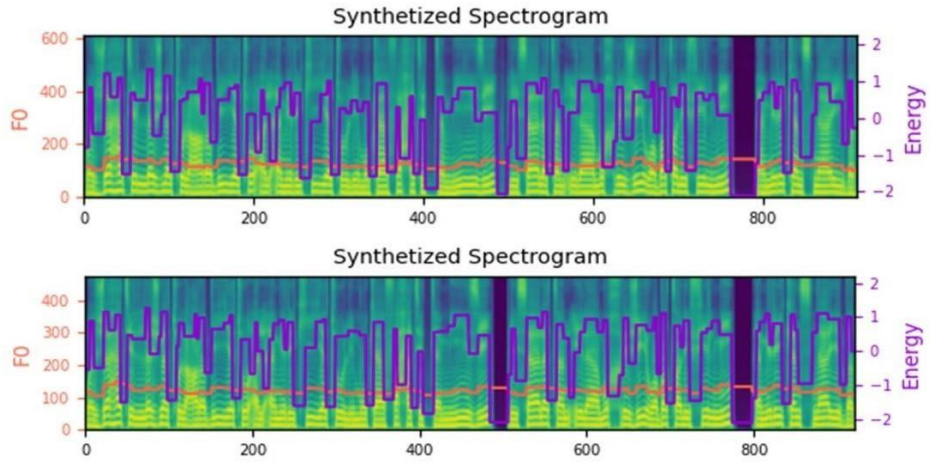


Figure 2: Objective Results with Arabic corpus: Top: Full data; Bottom: Limited data

Table 1: The performance of the FastSpeech2 model when trained using both full data and limited data.

Metrics	Full Data	Limited Data
Mel Loss	0.473	0.549
Mel PostNet Loss	0.472	0.549
Pitch Loss	0.331	0.906
Energy Loss	0.080	0.094

References

[1] Ren Y., Hu C., Tan X., Qin T., Zhao S., Zhao Z., Liu T.: FastSpeech2: Fast and High-Quality End-to-End Text to Speech. The International Conference on Learning Representations (ICLR) (2021).

[2] Halabi, N. and Wald, M.: Modern Standard Arabic Phonetics for Speech Synthesis (2016).

A beszéd temporális jellemzőinek longitudinális változása sclerosis multiplexben

Svindt Veronika,[1] Gosztolya Gábor[2,3], Hoffmann Ildikó[1,4] és Bóna Judit[5]
[1] Nyelvtudományi Kutatóközpont, Budapest
[2] SZTE Informatikai Intézet, Szeged
[3] ELKH-SZTE Mesterséges Intelligencia Kutatóközpont, Szeged
[4] SZTE Pszichiátriai Klinika, Szeged
[5] ELTE Alkalmazott Nyelvészeti és Fonetikai Tanszék, Budapest

Bevezetés

A sclerosis multiplex (SM) a központi idegrendszert érintő, jelen tudásunk szerint nem gyógyítható, autoimmun eredetű neurodegeneratív megbetegedés. A betegség következében az agyban keletkező gyulladásos gócok elsősorban a fehérállományt érintik, jelentősen lassítva az érintett területen az információáramlást. A neurális változások a fizikai és mentális tüneteken túl érinthetik a beszéd- és nyelvi folyamatokat is, mivel megjelenhetnek a beszéd- és hangképzés zavaraiban és a produkciót és/vagy a percepciót érintő nyelvi zavarokban (pl. szókeresési és szóelőhívási nehézségek, megnevezési zavar, verbális fluencia korlátozódása). Az információfeldolgozási sebesség lassulása az SM egyik első tünete, amely megjelenhet a beszédprodukciós folyamatokban is [1, 2, 3].

Noha – más neurológiai betegségek esetében – számos kutatás igazolja, hogy a beszéd temporális paramétereinek vizsgálata érzékeny eszköz lehet a háttérben zajló/kezdődő kognitív állapotromlás előrejelzésére, csak igen kevés kutatást találunk, amely SM-ben igyekezne feltárni ezeknek az összefüggéseit [4, 5]. Emellett tudomásunk szerint nem készült olyan kutatás, amely a temporális paraméterek változásait longitudinális vizsgálat során tárná fel SM-ben.

A kutatás célja

A jelen kutatásban arra keressük a választ, hogy hogyan változnak az SM-betegek beszédének temporális jellemzői az idő előrehaladtával két év távlatában, és ezek a változások milyen összefüggésben vannak a betegek állapotában bekövetkező fizikai, mentális és kognitív változásokkal.

Anyag és módszer

Vizsgálatunkban 16 SM-mel élő személy és 12 életkorban, nemben és iskolázottságban illesztett kontrollszemély vett részt. Minden résztvevővel összesen 3 időpontban készítettünk felvételt 1-1 év elteltével (0. év, 1. év, 2. év). A vizsgálatok során sztenderd mérőeszközökkel mértük a betegek (a) fizikai; (b) általános életminőségbeli; (c) kognitív állapotát,[1] valamint minden évben beszédfelvételeket készítettünk. A beszéd temporális paramétereinek mérését automatikusan végeztük (vizsgált paraméterek: artikulációs és beszédtempó, néma és kitöltött szünetek gyakorisága, hossza és aránya). Négy különböző beszédtípust vizsgáltunk, melyek a következők voltak: (a) spontán beszéd (hobbi); (b) strukturált személyes narratíva (tegnapi nap elmesélése); (c) véleménykifejtés adott témában; (d) képleírás (Boston Cookie Theft picture).

[1] Az alkalmazott eszközök: (a) Expanded Disability Status Scale (EDSS), (b) Életminőség kérdőív SM betegek részére (MSQOL), Beck depresszió skála, krónikus fáradtság mérő skála (Fatigue Impact Scale, FIS); (c) Addenbrooke's Kognitív Vizsgálat (AKV), Wisconsin Kártyaszortírozó Teszt (WCST), Stroop teszt, verbális fluencia, munkamemória (számterjedelem előre és vissza).

Eredmények

Az első bemeneti és a harmadik, vagyis utolsó mérés eredményeinek összehasonlítása során arra jutottunk, hogy a betegek 75%-ánál legalább 10%-os romlás következett be minden beszédparaméterben beszédtípustól függetlenül (a romlás átlagos mértéke: 38%) (1. ábra). A kontrollcsoportban senkinél nem találtunk az első és a harmadik mérési pont között 10%-nál nagyobb eltérést. Az eredmények azt mutatják, hogy sem a fizikai, sem a mentális állapotromlás nem függ össze a beszédparaméterekben bekövetkező változásokkal. Azt találtuk továbbá, hogy a kognitív képességeket mérő tesztekben (AKV, információfeldolgozási sebesség, verbális fluencia, munkamemória) nincs szignifikáns változás a 1. és 3. évi mérési pontok eredményei között az SM-mel élő csoportban, vagyis a beszédparaméterek szignifikáns megváltozásával nem járt együtt a kognitív tesztekben mért paraméterek romlása. Mindazonáltal fontos megjegyezni, hogy az információfeldolgozási sebesség, valamint a verbális fluencia feladatban elért teljesítményben szignifikáns különbség volt már az 1. mérési pontban is a két csoport között: az SM-mel élő csoport információfeldolgozási sebessége azonos teljesítmény mellett (vagyis a hibák számát tekintve nem volt különbség a csoportok között) szignifikánsan lassabb volt. A különbség mértéke azonban nem nőtt az első és utolsó felvétel között. Szintén közel állandó eredményt látunk mindhárom mérési pontban az általános kognitív képességfelmérésen (AKV), valamint a verbális fluencia és a munkamemória (számterjedelem előre és vissza) feladatok esetében a csoportokon belül. A csoportok között különbség, hogy a verbális fluencia feladatban a klinikai csoport tagjai szignifikánsan kevesebb szót soroltak mindhárom méréskor. A munkamemória feladatban azt tapasztaltuk, hogy nincs szignifikáns különbség a két csoport eredménye között, valamint nem tapasztalható változás a 3. mérési pontban az 1. mérési ponthoz képest.

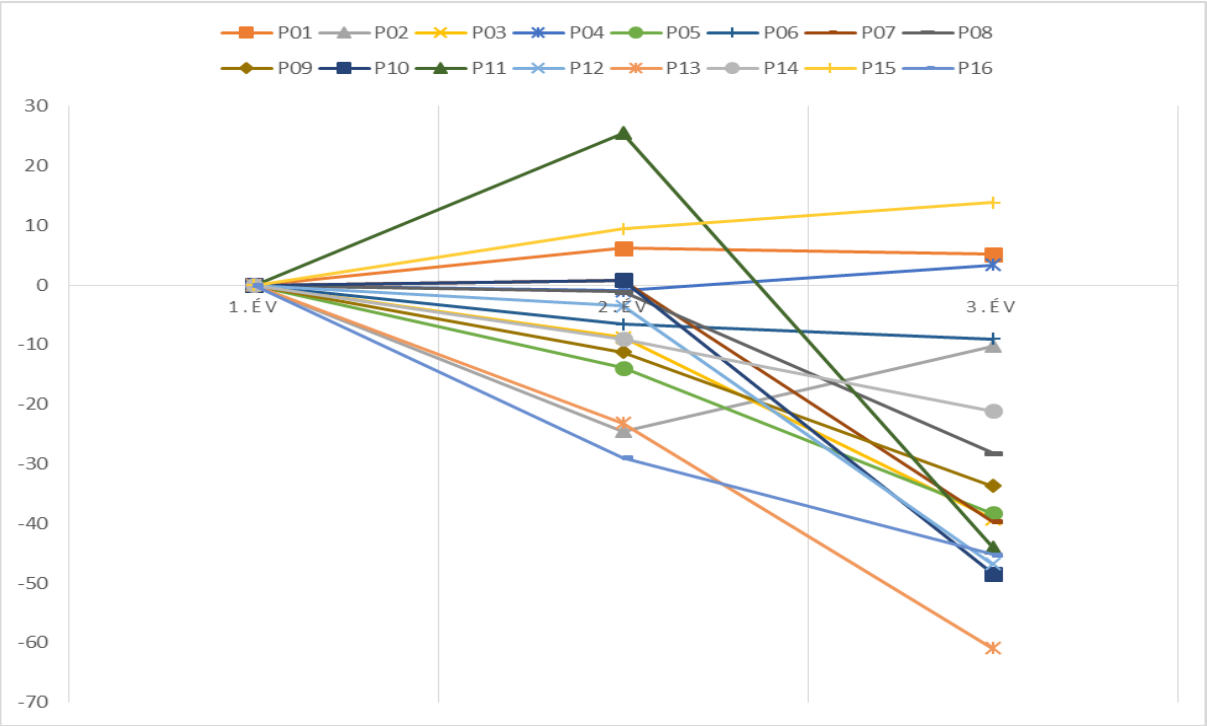
Összefoglalás

Eredményeink azt mutatják, hogy az összefüggés a fizikai, kognitív, valamint a nyelvi- és beszédképességek között korántsem triviális. Úgy tűnik, hogy a fizikai és mentális állapot nem jár együtt a beszédparaméterek megváltozásával. Mindez igaz a kognitív képességekre is: míg a kognitív képességekben nem találtunk jelentős változást az egyes mérési pontokban, a beszédtempó és az artikulációs tempó szingifikánsan lelassult, a néma és kitöltött szünetek aránya az idő elteltével szignifikánsan megnőtt az SM-mel élőknél. Eddigi eredményeink alapján valószínűsítjük, hogy a jelenséget a gondolkodás általános lassulása okozza, amely hatással van a komplex narratívaprodukcióra, a mondanivaló adekvát megformálására. A kutatás jelen állapotában nincs elegendő evidenciánk arra nézve, hogy a kapott eredmények vajon egy később bekövetkező kognitív romlás előrejelzői-e. Hipotézisünket további, több évig tartó utánkövetés támaszthatja alá vagy cáfolhatja meg.

Mindezeket együttvéve a kapott adatok jelen formájukban is hozzájárulnak ahhoz, hogy a klinikumban dolgozó szakemberek egy érzékenyebb módszer, a beszéd vizsgálata segítségével képet kapjanak betegeik állapotának változásáról, az esetleges romlásról, továbbá segítheti őket a klinikumban használt, de általában kevésbé érzékeny tesztbatériák eredményei mellett a gyakran nem feltétlenül egyértelmű állapotromlás felismerésének megerősítésében.

Köszönetnyilvánítás

A kutatás az NKFI-132460 számú projekt keretében készült.



1. ábra: A beszédtempó változása az egyes mérési pontokban az SM csoportban (az y tengelyen a változás mértéke a 0. felvételi ponthoz képest százalékban kifejezve)

Irodalom

[1] Denney, D. R., Gallagher, K. S., & Lynch, S. G. (2011). Deficits in processing speed in patients with multiple sclerosis: Evidence from explicit and covert measures. *Archives of Clinical Neuropsychology*, *26*(2), 110–119. https://doi.org/10.1093/arclin/acq104

[2] Prakash, R. S., Snook, E. M., Lewis, J. M., Motl, R. W., & Kramer, A. F. (2008). Cognitive impairments in relapsing-remitting multiple sclerosis: A meta-analysis. *Multiple Sclerosis*, *14*(9), 1250–1261. https://doi.org/10.1177/1352458508095004

[3] Rodgers, J. D., Tjaden, K., Feenaughty, L., Weinstock-Guttman, B., & Benedict, R. H. B. (2013). Influence of cognitive function on speech and articulation rate in multiple sclerosis. *Journal of International Neuropsychological Society*, *19*(2), 173–180. https://doi.org/10.1017/S1355617712001166

[4] Feenaughty, L., Tjaden, K., Benedict, R. H. B., & Weinstock-Guttman, B. (2013). Speech and pause characteristics in multiple sclerosis: A preliminary study of speakers with high andlow neuropsychological test performance. *Clinical Linguistics and Phonetics*. https://doi.org/10.3109/02699206.2012.751624

[5] Svindt, V., Bóna, J., & Hoffmann, I. (2020). Changes in temporal features of speech in secondary progressive multiple sclerosis (SPMS)–case studies. *Clinical Linguistics and Phonetics*, *34*(4), 339–356. https://doi.org/10.1080/02699206.2019.1645885

A magyar magánhangzó-minőségek térbeli variabilitásának akusztikai vizsgálata

Vargha Fruzsina Sára[1]
[1] Nyelvtudományi Kutatóközpont, Budapest
vargha.fruzsina@nytud.hu

A magyar nyelvjáráskutatás fő forrásai az impresszionisztikusan lejegyzett nyelvjárási adatok. Nagyrészt igaz ez az innovatív elemzési módszereket alkalmazó újabb kutatásokra is. Az észak-amerikai nyelvjárások atlasza ([4]) új alapokra helyezi a nyelvföldrajzi kutatásokat: a magánhangzó-minőségeket akusztikai mérésekkel vizsgálja, kiküszöbölve így a lejegyzésből adódó szubjektivitást. Eddig csupán elvétve akadnak példák a magyar nyelvjárások akusztikai jellemzőinek vizsgálatára, illetve akusztikai elemzések alkalmazására a nyelvjárási változások vizsgálatában ([1, 3, 6, 8]). A Magyar nyelvjárások akusztikai és dialektometriai vizsgálata című kutatásunkban (NKFIH FK-138396, http://nyelvfoldrajz.nlp.nytud.hu) vállaltuk, hogy A magyar nyelvjárások atlasza (MNyA., [5]) 1960 és 1964 között készült hangfelvételei alapján feltérképezzük a magyar magánhangzók ejtésének térbeli variabilitását.

A MNyA. gyűjtését több szempontból is ideális kiindulási pontnak tartjuk, egyrészt így eredményeink összevethetők lesznek az atlasz adatai alapján készült kvantitatív vizsgálatok eredményeivel, másrészt alapot adnak későbbi, a magánhangzók ejtésének változását vizsgáló kutatásoknak. A magánhangzó-minőségek megállapítására formánsméréseket végzünk, az első és a második formáns értékének meghatározásával. A formánsértékek meghatározását a PRAAT szoftverrel végezzük, PRAAT szkriptek futtatásával. Az automatikus mérések gyakran nem kielégítő minőségének javítása céljából a formánsplafon (maxFormant) paramétert a magánhangzó-minőségek figyelembevételével (vö. [2]) állítjuk be a szkripteket generáló és futtató nyelvjárási lejegyzőben. A formánsértékeket, valamint a magánhangzók időtartamát a lejegyzéseken belül, az egyes magánhangzó-realizációk fonetikus szimbólumaihoz kötötten rögzítjük, ez biztosítja számunkra, hogy ne vesszen el a kapcsolat a mérések, a hanganyag és a lejegyzés között. Az általunk alkalmazott módszer erőssége a lejegyzéseken alapuló, így a lejegyzési rendszer megengedte ejtésbeli különbségek által limitált variabilitáshoz képest, hogy minden egyes mérésünk egy folytonos skála mentén helyezkedik el, és minden egyes, az adattárba bekerülő hang bármiféle külső viszonyítási minőség (pl. köznyelvi ejtésmód) nélkül, saját inherens tulajdonságai alapján kap helyet az F1 és az F2 tengely által meghatározott síkban.

A hangfelvételek az 1960-as évek első felében készültek, magnetofonnal, zárt térben. Digitalizálásuk 2005-ben történt meg, ekkor már a szalagok állapota miatt fizikailag is restaurálni kellett azokat, hogy lejátszhatók legyenek ([7]). A felvételek hangminősége változó, még a jobb állapotban megmaradt hanganyagoknál is körültekintőnek kell lennünk a formánsmérések során. A nyelvjárási magánhangzók elemzésekor további elméleti nehézség, hogy sokszor nehezen vagy nem határozható meg, hogy egy-egy hangot milyen fonéma realizációjának tekinthetünk. A kutatási anyag jellegzetességeiből adódóan kiemelten fontos, hogy a módszertan megfelelő legyen, és ezáltal minél hatékonyabban és minél eredményesebben jussunk hozzá a kutatási céljaink megvalósításához szükséges mérési eredményekhez. Ezért a mérési adatok rögzítésének olyan rendszerét dolgoztuk ki, amely lehetővé teszi, hogy a

későbbiekben, újabb módszerek kidolgozásával akár automatizáltan újramérhessük a formánsértékeket, finomítva az elemzéseket.

Előadásom elsősorban módszertani jellegű. Bemutatom a készülő adattár munkamódszerét, jellegzetességeit, elemzési lehetőségeit. Az *e*, a zárt *ë*, a hosszú *á* és az *a* formánsértékeinek térbeli variabilitásán keresztül kitérek arra is, hogyan helyezi új megvilágításba a magánhangzók ejtését, ha nem diszkrét módon, hanem ugyanabban az akusztikai térben, tehát azonos viszonyrendszerben, egymáshoz képest vizsgáljuk azokat.

Irodalom

[1] Both Csaba Attila (2016). A magánhangzók ejtésváltozatai két háromszéki településen, Bölönben és Dálnokon. In: Czetter Ibolya, Hajba Renáta, Tóth Péter. Szerk. *VI. Dialektológiai Szimpozion. Szombathely, 2015. szeptember 2-4.* Szombathely–Nyitra: Nyugat-magyarországi Egyetem Savaria Egyetemi Központ – Konstantin Filozófus Egyetem, Közép-Európai Tanulmányok Kara. 179–192.

[2] Escudero, Paola, Paul Boersma, Andréia Schurt Rauber, Ricardo A. H. Bion (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America*, 126(3), 1379-1393. https://doi.org/10.1121/1.3180321

[3] Kocsis Zsuzsanna (2013). Nyelvjárási *ë* hangok vizsgálata – különös tekintettel a labializáció mértékére. In: Kontra Miklós, Németh Miklós, Sinkovics Balázs. Szerk. *Elmélet és empíria a szociolingvisztikában*. Budapest: Gondolat. 432–443.

[4] Labov, William, Sharon Ash, Charles Boberg (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change.* Berlin/New York: Mouton de Gruyter.

[5] MNyA. = Deme László – Imre Samu. Szerk. (1968–1977). *A magyar nyelvjárások atlasza 1–6.* Budapest: Akadémiai Kiadó.

[6] Presinszky Károly (2016). Az *á* előtti illabiális *ȧ*-zás a Csallóközben. *Magyar Nyelv* 112: 218–227.

[7] Vargha Fruzsina Sára (2007). Nyelvi változók A magyar nyelvjárások atlasza hangfelvételeiben. In: Guttmann, Miklós; Molnár, Zoltán. Szerk. *V. Dialektológiai szimpozion.* Szombathely: Berzsenyi Dániel Főiskola. 279–288.

[8] Vargha Fruzsina Sára (2013). A hangzó adat szerepe a magyar dialektológiában. In: Szoták Szilvia – Vargha Fruzsina Sára. Szerk. *Változó nyelv, nyelvváltozatok, területiség: A VII. Hungarológiai Kongresszus nyelvészeti tanulmányai.* Kolozsvár: Egyetemi Műhely Kiadó–Bolyai Társaság. 94–204.

Demonstrations/Bemutatók

A Magyar Beszédtechnológia Múzeuma az interneten - kutatóknak, oktatóknak és hallgatóknak

Olaszy Gábor[1] és Abari Kálmán[2]
[1] Budapesti Műszaki és Gazdaságtudományi Egyetem TMIT
[2] Debreceni Egyetem

A digitális világ fejlődéséhez igazodva készítettük el a Magyar Beszédtechnológia Múzeumát, amely az internet adta lehetőségeket kihasználva egy interaktív, multimédiás, több internetes fórumot is összefoglaló magyar és angol nyelvű honlap (www.magyarbeszed.hu). A beszédtechnológia tudománya fiatal, mindössze 50 évesnek mondható. A szerzők egyrészről úgy látták, hogy meg kell örökíteni és az utókornak át kell adni e tudományág hazai fejlődési állomásait, mint egy múzeumi tárlatot, másrészről az is céljuk, hogy a honlap kapcsolódjon a mindenkori jelenhez, folyamatosan bemutatva a legújabb kutatási és fejlesztési eredményeket.
Ez a honlap elsősorban a szakmának szól: kutatóknak, oktatóknak, hallgatóknak, de bárki böngészhet benne, aki a beszéd "titkairól" többet szeretne tudni.

Szerkezeti felépítés

A honlap tartalmilag rendkívül sokszintű. Helyet kaptak letölthető, illetve közvetlenül futtatható szakmai programok, képek és táblázatok, szöveges ismertetők, megjeleníthető szakcikkek, letölthető könyvek, tudományos publikációk, disszertációk, kapcsolódó honlapok, valamint hang és video összeállítások. Mindezekből adódik, hogy az interaktív honlap a kutatásban és az egyetemi oktatásban is jól felhasználható.
Szerkezetileg öt fő témakör (menüpont) köré csoportosítottuk a gazdag tartalmat. Mindegyik számos altémakört kínál, sok esetben további szintek jelennek meg a könnyebb áttekinthetőség kedvéért. A főbb témakörök és altémakörök a következők:

Nyelv-írás-beszéd

Szakszótár - *a legfontosabb beszédtechnológiai szakkifejezések gyűjteménye*
Kiejtési szótár - *megnyitható program 1,5 millió írott magyar szóalak kiejtési formáival*
Írás és beszéd - *a beszéd és az írás kapcsolata, fonetikus átirat*
Hangszimbólumok - *IPA, SAMPA és más hangszimbólum készletek táblázatai*
Nyelvi statisztikák - *betű-, szótag-, szó- és hangkapcsolat statisztikák*

Beszédakusztika

Alapok - 6 alfejezettel
Gerjesztés; Artikuláció; Beszédhullám; Hangintenzitás;
Hangidőtartam; Beszédspektrum
Beszédadatbázisok - 4 alfejezettel
Fonetikai-; Ejtett beszéd-; BEA-; PPBA adatbázis
Mondatdallamok - *magyar mondatdallamok modellezése*
Hangsúly adatbázis - *az első magyar letölthető hangsúly adatbázis*
Interaktív programok - 6 témakörrel
CCCC adatbázis; CVVC adatbázis; Formáns adatbázis;
Formánsmozgások térképe; HIDOL hangidőtartam mérő program;
Szöveg-formáns konverter

Beszédszintézis

1980 kezdetek - 8 alfejezettel

Kempelen Farkas (1791); Bánó Miklós (1916); Hungarovox (1982); PKI (1982); BrailabPC (1983); Scriptovox C-64 (1986); PC talker (1988)

1988 BME Multivox TTS család - 3 alfejezettel

Multivox 8 nyelvű szöveg-beszéd átalakító; PC-robot kártya; Multivox-4 szoftver alapú TTS

1996 ProfiVox TTS család - 3 alfejezettel

Profivox diád-triád; ProfiVox HMM; Profivox korpusz

2002 FlexVoice TTS - *gépi tanulásra épülő, inverz formánsszintetizátor*

2020 BME Neural ProfiVox - *gépi tanulás mély neurális hálózatokkal*

Beszédfelismerés

Kezdetek - 3 alfejezettel

Tarnóczy Tamás (1971); Vicsi Klára (1985); BME TMIT (1988)

1990 BME TMIT - *statisztikai adatokkal vezérelt gépi tanulás*

2013-2022 SpeechTex - *21. századi technológia, folyamatos beszédleírat készítése*

Alkalmazások

Szöveg-beszéd átalakítás (TTS) - 10 alfejezettel

Távközlés; Képernyő olvasó; Robobraille; Stroke aid; MÁV utastájékoztató; Időjárás jelentés; Beszélő bank automaták; Filmipar; NAO robot, GOH hallásvizsgáló

Beszéd-szöveg átalakítás (ASR) - 2 alfejezettel

Bírósági beszédleíró automata; MTVA automatikus feliratozó

ASR- TTS alapú beszélgető automata - 2 alfejezettel

Gyógyszervonal telefonos automata; Voice chatbot automata

Beszélő fej - *virtuális bemondó; transzparens artikulációs modell siketek tanításához*

Varázsdoboz - *számítógépes beszédoktató beszédhibás gyermekek részére*

Prozódia oktató - *számítógépes prozódia oktató*

A honlap nyitott minden kutató fejlesztő előtt. Szívesen fogadunk új eredményeket a szakma bármely részéről és az évi rendszeres frissítéskor – egyeztetés után – beépítjük azokat a honlap szerkezetébe. Jó böngészést!

www.magyarbeszed.hu