BESZÉDTUDOMÁNY – SPEECH SCIENCE

2025 (5) 1

Szerkesztők/Editors:

Gráczi, Tekla Etelka

Mády, Katalin

HUN-REN Nyelvtudományi Kutatóközpont

HUN-REN Hungarian Research Centre for Linguistics

Budapest

Technikai szerkesztés/Typesetting: Garai, Luca

Borítóterv/Cover design: Iuga, Laura ©

Korrektúra/Proofreading: Jankovics, Julianna

A folyóiratszám kiadását a Magyar Kutatási Hálózat (HUN-REN) támogatta.
This volume was supported by the Hungarian Research Network (HUN-REN).

Szerkesztői előszó

A jelen kötet írásai Csapó Tamás Gábornak kívánnak emléket állítani. Tamás hirtelen jött halála 2024 januárjában családját, barátait és kollégáit is szó nélkül hagyta. Idén, 2025-ben töltötte volna be 40. életévét.

Tamás 1985. január 28-án született. A Budapesti Műszaki és Gazdaságtudományi Egyetemre járva már hamar nyilvánvalóvá vált, hogy tudományos irányba fordul. Minden, a beszédről szóló újdonság, amivel találkozott, foglalkoztatta. Mindezeket kipróbálta a beszédszintézisben, hogy az így előállított beszéd minél természetesebbnek hasson. Munkájában fáradhatatlanul, tele ötletekkel és elszántsággal, mindemellett nagy tudással és hatalmas tehetséggel lényegében szenvedélyre is talált.

Szakmai hozzájárulását 184 tudományos műve, 390 (ismert) független hivatkozása és a számos díj, kitüntetés, ösztöndíj elnyerése is mutatja, amelyeket már hallgató korától folyamatosan szerzett. Munkássága és pályázati tevékenysége nem csak a beszédtechnológia számára, hanem a fonetika területén is jelentős újdonságokat hozott. 2014-es bloomingtoni ösztöndíja alatt elsajátította a nyelvultrahang használatát, és ezt a tudást itthon 2016–2021 között az MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport munkatársaként kamatoztatta és adta tovább, valamint a maga vezette beszédtechnológiai „OTKA-”pályázataival (NKFIH-pályázatok). Utoljára indult pályázatában már az artikulációból készülő beszédszintézistől az agyi jelek alapján történő beszédelőállítás felé is nyitott, ezzel a legújabb nemzetközi kutatási irányokhoz csatlakozott kutatótársaival. Tamás halálával ezt a projektet Sztahó Dávid vezetésével folytatják kollégái, így a beszédtechnológiai tevékenysége és újításai tovább élnek. A Tamás által hazahozott tudás fonetikai felhasználása pedig speciális beszélői csoportok kutatásával folytatódik. Tamás díjait, kitüntetéseit és ösztöndíjait a megemlékező előszó végén közöljük.

Tamás díjai közül kiemelkedik egy, amely adományozásának indoklásában szinte mindent magában foglal az ő lényéből. A Kozma László-emlékérem szövegezésében a magas színvonalú oktatási tevékenységet (előadások, gyakorlatok, laborok tartása, szakdolgozati és PhD-témavezetés), a beszédtechnológia terén elért, kiemelkedő eredményeit és önzetlen kapcsolatépítő tevékenységét emelték ki. Tamás lelkes szervezője volt minden kisebb-nagyobb eseménynek, amelyen bármilyen területen dolgozó kutatókat összehozhatott egymással, hogy meríthessenek a másik tudásából, vagy akár csak jó emberi, kollegiális kapcsolat épüljön ki, legyen szó hétköznap egy közös ebédről, vagy

éppen egy nemzetközi workshopról. Tamás 2023 nyarán, Kimlén tartotta a Moonshine Workshopot, amire a világ minden tájáról hívta kutatótársait, beszédtechnológusokat, fonetikusokat, logopédust és zoológust, hogy akár közös területeket keressenek, akár csak a valami módon kapcsolódó kutatásaikat megismerjék?, és hogy egy virágzó kapcsolati háló kibontakozását elindítsa: https://smartlab.tmit.bme.hu/moonshine-2023/index.html. Bármilyen konferencián járt ő maga, hasonlóan minden kicsit is a beszédhez köthető kutatásokról hosszas beszélgetésekbe vegyült a jelenlévőkkel, így tudást és ismeretségeket gyűjtött. Amikor épp nem volt konferencia, egy csokor kollégát gyűjtött mindenfelől közös ebédre, hogy az eszmecsere ne álljon le.

Tamás személyisége erre, a kapcsolatépítésre kiválóan alkalmas volt. Nemcsak szakmailag, hanem emberileg is nyitott és támogató volt. Mindenben segítőkész, akár pár jó szóval, akár tevőlegesen tudott hozzájárulni a társaihoz. Minden helyzetben megtalálta, és megláttatta másokkal is, hogy mit lehet abból hazavinni, megtanulni, hogyan lehet továbblépni, fejlődni.

Tamás odaadóan vett részt a tudományos ismeretterjesztésben is. Számtalan előadást tartott, amelyben a beszédről és a beszédszintézis lehetőségeiről színesen és egyben közérthetően beszélt a Kutatók éjszakája, A hang világnapja alkalmából, vagy ezektől független meghívások keretében is (pl. https://www.tmit.bme.hu/meet-the-scientist-csapo-2015).

Írásunk, megemlékezésünk Tamásról nem lehet teljes, hiszen bizonyára ennél még több oldala volt. Akár az őt más színtérről ismerő kollégái is rengeteg további mozaikdarabbal szolgálhatnak tevékenységéről, személyiségéről.

Csapó Tamás Gábor nem csak kiváló kutató és fantasztikus ember volt, hanem odaadó apa és férj is. Felesége és négy gyermeke, valamint szülei és testvére, mellettük pedig barátai számára is hatalmas veszteség távozása.

A jelen kötetben néhány megemlékezés és írás Tamás tiszteletére készült, ezek keretében kívánván emléket állítani számára.

Németh Géza megemlékezésében Tamás életútját, személyiségét és tehetségét méltatva búcsúzik saját és kollégái nevében. Juhász Kornélia Tamásról nyitottságát, motiváló erejét, odaadását emeli ki soraiban. Ibrahim Ibrahimov a hallgató–témavezető kapcsolat mentén emlékezik meg Tamásról, akire nemcsak témavezetőként, de emberként is lehetett számítani.

A tanulmányok között két olyan területről szóló írásokkal tisztelegnek emléke előtt kollégái, amelyeken kutatásai nem a beszédtechnológia, hanem fonetikai területek felé adott kiemelkedő hozzájárulást. Markó Alexandra, Deme Andrea, Juhász Kornélia és Gráczi Tekla Etelka tanulmánya arról számol be, hogy Tamás ultrahangtudása és annak Magyarországra hozatala hogyan járult hozzá az artikulációs fonetikai vizsgálatok hazai újraéledéséhez, és milyen nyelvultrahangos fonetikai

eredmények készültek az ő munkásságának köszönhetően. Pertti Palo, Steven Lulich és Daniel Aalto ugyancsak a nyelvultrahanggal kapcsolatos munkájára reflektál, mégpedig az ultrahangfej elmozdulásának mérési lehetőségeiről írtak, amivel Tamás is foglalkozott munkájában. Steven Lulich pedig Tamás egy korábbi témájához kapcsolódó írásában a szubglottális rezonanciák területéről számol be a legújabb eredményekről. Ez utóbbi két írás esetében azt is fontos kiemelni, hogy Tamás a nyelvultrahang és a szubglottális rezonanciák kutatását is Steven M. Lulich vezetése alatt kezdte meg. A beszédtechnológia területéről Trencsényi Réka és Czap László tanulmánya Tamás artikulációalapú beszédszintéziséhez kapcsolódik. Nyelvultrahang-videók feldolgozása alapján történő beszédszintézis eredményeiről számolnak be.

A kötet végén Tamás egy kiváló szemináriumi dolgozata található. A dolgozatot Steven M. Lulich Speech Acoustics órájára írta, kiemelkedő munkát végzett. A vizsgálatból később Németh Gézával közösen született egy magyar nyelvű publikáció is: Csapó, Tamás Gábor, Németh, Géza: Mássalhangzó-magánhangzó kapcsolatok automatikus osztályozása szubglottális rezonanciák alapján. In: Tanács, Attila; Szauter, Dóra; Vincze, Veronika (szerk.) *VI. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2009* : Szeged, 2009. december 3-4. Szeged, Magyarország : Szegedi Tudományegyetem Informatikai Tanszékcsoport (2009) 427 p. pp. 226-237., 12 p.

Tamásról még nagyon sok szemszögből lehetne mesélni, izgalmas és barátságos személye szinte minden találkozáskor hozott valamit, akár példamutatásként, tanulnivalóként, akár érdekességként. Búcsúzunk Csapó Tamás Gábortól, de szívünkben megtartjuk.

Csapó Tamás Gábor

díjai, kitüntetései, ösztöndíjai

2024, posztumusz: Kozma László-emlékérem, BME VIK TMIT:

2022: NKFIH K 22 *Analysis of articulation and brain signals for speech-based brain-computer interfaces* (142163)

2018: NKFIH PD 18 *Némabeszéd-interfészek kidolgozása és a beszélőfüggőség vizsgálata nyelvultrahang és elektromágneses artikulográf eszközökkel* (127915). A kapcsolódó publikációkról itt tájékozódhat:
http://nyilvanos.otka-palyazat.hu/index.php?menuid=930&lang=EN&num=127915

2017: NKFIH FK-17 *Artikulációs mozgás alapú beszédgenerálás* (124584). A kapcsolódó publikációkról itt tájékozódhat:
http://nyilvanos.otka-palyazat.hu/index.php?menuid=930&num=124584

2016: Nemzeti Fejlesztési Minisztérium, Információs Társadalomért szakmai érem

2015: Huszty Dénes Alapítvány, Disszertáció pályázat

2014: Fulbright ösztöndíj, Department of Speech and Hearing Sciences, Indiana University, Bloomington, IN, USA; kutatási téma: *Nyelvmozgás vizsgálata beszéd közben ultrahang segítségével.*

2014: Magyar Mérnökakadémia utazási ösztöndíj, Bloomington

2013: BME kutatói pályázat, 3. helyezés

2013: honlaptervező verseny: Microsoft No Time to W8 verseny, 1. helyezés. *Időjárás mindenkinek'* Windows 8 alkalmazás (megosztva más kollégákkal a BME-TMIT-ről)

2013: Campus Hungary utazási ösztöndíj, 8th Speech Synthesis Workshop

2010: Acoustical Society of America - International Student Grant

2009: OPAKFI Diplomamunka pályázat, 3. helyezés

2009: Bizáky Puky Péter Alapítvány, utazási ösztöndíj, Interspeech 2009

2007: BME VIK Tudományos Diákköri Konferencia, 2. helyezés

2007: BME GTK Tudományos Diákköri Konferencia, 1. helyezés

2007: Országos Tudományos Diákköri Konferencia, 1. helyezés

2007: Köztársasági ösztöndíj, Oktatási Minisztérium

2007: Kari BME ösztöndíj, BME VIK

2007: Egyetemi BME ösztöndíj, BME

2007: International Speech Communication Association, utazási ösztöndíj, Interspeech 2007

2006: BME VIK Tudományos Diákköri Konferencia, 1. helyezés

Editorial foreword

The papers in this volume aim to commemorate Tamás Gábor Csapó. Tamás's sudden passing in January 2024 left his family, friends and colleagues in shock. This year, in 2025, he would have turned 40.

Tamás was born on the 28th of January 1985. While studying at the Budapest University of Technology and Economics, it quickly became evident that he was drawn towards scientific research. He was deeply engaged with every innovation related to speech that he encountered. He explored them in speech synthesis, striving to make the generated speech sound as natural as possible. His work was marked by tireless effort, abundant ideas and determination, coupled with extensive knowledge and immense talent, ultimately becominghis passion.

His professional contributions are reflected by his 184 scientific publications, at least 390 independent citations and numerous awards, honours and scholarships, which he received from his student years on. His research and project work brought significant innovations not only to speech technology but also to the field of phonetics. During his 2014 scholarship in Bloomington, he learned to master the use of ultrasound for speech studies, which he later applied and disseminated in Hungary between 2016 and 2021 as a member of the MTA-ELTE „Momentum" Lingual Articulation Research Group. He also led speech technology projects through „OTKA" (now NKFIH, National Research, Development and Innovation Office) grants. His latest project extended from articulatory speech synthesis to speech production based on brain signals, aligning with the latest international research trends. Following his passing, this project continues under the leadership of Dávid Sztahó, ensuring that Tamás's contributions to speech technology live on. His phonetic research also continues, focusing on special speaker groups.

Tamás's awards, honours and scholarships are listed at the end of this commemorative introduction. One award stands out, encapsulating his essence: the Kozma László Memorial Medal. The award's justification highlighted his high-level educational activities (lectures, practical sessions, laboratory work, MSc and PhD supervision, his outstanding achievements in speech technology, and his selfless efforts in building professional relationships. Tamás was an enthusiastic organizer of events, big and small, bringing together researchers from different fields to exchange knowledge and build professional and collegial relationships. Whether it was a casual lunch or an international workshop, he created opportunities for collaboration. In the summer of 2023, he organized the Moonshine Workshop in Kimle, inviting speech technologists, phoneticians, speech therapists and even zoologists from around the world to explore common ground and foster a thriving professional network: https://smartlab.tmit.bme.hu/moonshine-2023/index.html. At every conference he at-

tended, he engaged in deep discussions about any speech-related research, gaining both knowledge and professional connections. Even outside conferences, he regularly gathered colleagues for shared meals to keep the exchange of ideas flowing.

Tamás's personality was ideally suited for networking. He was open and supportive both professionally and personally, always willing to help, whether through kind words or active contributions. He had a knack for identifying the valuable takeaways in any situation and helping others see them, too, facilitating learning, growth and progress.

He was also dedicated to scientific outreach, giving numerous engaging and accessible talks on speech and speech synthesis at events such as Researchers' Night and World Voice Day, as well as through independent invitations (e.g., https://www.tmit.bme.hu/meet-the-scientist-csapo-2015).

This tribute to Tamás cannot be complete, as he had many more facets. Colleagues who knew him in different contexts could undoubtedly provide additional pieces to the mosaic of his work and personality.

Tamás Gábor Csapó was not only an excellent researcher and a fantastic person, but also a devoted father and husband. His passing is a tremendous loss to his wife, four children, parents and a sibling, as well as to his friends.

This volume contains commemorative pieces and writings in Tamás's honour, aiming to preserve his memory. Géza Németh pays tribute by highlighting Tamás's life, personality, and talent on behalf of himself and his colleagues. Kornélia Juhász emphasizes his openness, motivational energy and dedication. Ibrahim Ibrahimov reflects on their student–supervisor relationship, remembering Tamás not only as a mentor but also as a reliable and supportive individual.

Among the academic contributions in this volume, colleagues honour Tamás's memory with research focusing on areas where his work extended beyond speech technology into phonetics. Alexandra Markó, Andrea Deme, Kornélia Juhász and Tekla Etelka Gráczi discuss how Tamás's ultrasound expertise and its introduction to Hungary revitalized articulatory phonetics research and contributed to phonetic discoveries using ultrasound imaging. Pertti Palo, Steven Lulich and Daniel Aalto reflect on his work related to tongue ultrasound, specifically measuring ultrasound probe displacement – an area he actively researched. Steven Lulich also contributes a study on subglottal resonances, a topic Tamás initially pursued under his guidance. In speech technology, Réka Trencsényi and László Czap present findings related to Tamás's articulation-based speech synthesis, processing speech synthesis from ultrasound tongue videos.

At the end of this volume, one of Tamás's outstanding seminar papers is included. He wrote this paper for Steven M. Lulich's Speech Acoustics course, demonstrating exceptional work. The

research later led to a publication in Hungarian co-authored with Géza Németh: Csapó, Tamás Gábor & Németh, Géza: "Automatic Classification of Consonant-Vowel Connections Based on Subglottal Resonances." In: Tanács, Attila; Szauter, Dóra; Vincze, Veronika (eds.) *VI. Hungarian Conference on Computational Linguistics: MSZNY 2009*, Szeged, Hungary: University of Szeged, Department of Informatics (2009) 427 p. pp. 226–237.

There are many more stories to tell about Tamás. His engaging and friendly personality always brought something valuable to every encounter–whether as an example to follow, a lesson to learn, or an intriguing insight. We bid farewell to Tamás Gábor Csapó but hold his memory in our hearts.

Tamás Gábor Csapó

awards, grants

2024, posthumous award: Kozma László-emlékérem, BME VIK TMIT

2022: NKFIH K 22 research grant: *Analysis of articulation and brain signals for speech-based brain-computer interfaces* (142163)

2018: NKFIH PD 18 research grant: *Némabeszéd-interfészek kidolgozása és a beszélőfüggőség vizsgálata nyelvultrahang és elektromágneses artikulográf eszközökkel* [*Development of silent speech interface platforms and analysis of their speaker dependent results: electromagnetic articulography and tongue ultrasound imaging*] (127915).
Publications: http://nyilvanos.otka-palyazat.hu/index.php?menuid=930&lang=EN&num=127915

2017: NKFIH FK-17 research grant: *Artikulációs mozgás alapú beszédgenerálás* [*Articulatory speech synthesis*] (124584). A kapcsolódó publikációkról itt tájékozódhat: http://nyilvanos.otka-palyazat.hu/index.php?menuid=930&num=124584

2016: Award: Nemzeti Fejlesztési Minisztérium, Információs Társadalomért szakmai érem

2015: Award: Huszty Dénes Alapítvány, Disszertáció pályázat

2014: research grant: Fulbright stipendium: Department of Speech and Hearing Sciences, Indiana University, Bloomington, IN, USA; kutatási téma: *Nyelvmozgás vizsgálata beszéd közben ultrahang segítségével [Tongue ultrasound imaging during speech]*

2014: travel grant to Bloomington: Magyar Mérnökakadémia utazási ösztöndíj, Bloomington

2013: award: BME kutatói pályázat, 3. helyezés

2013: web page development competition: Microsoft No Time to W8, 1st place. *Időjárás mindenkinek' [Weather forcast for everyone]* Windows 8 application (together with colleagues from BME-TMIT)

2013: travel grant: Campus Hungary, 8th Speech Synthesis Workshop

2010: award: Acoustical Society of America - International Student Grant

2009: award: OPAKFI MSc-thesis, 3rd place

2009: travel grant: Bizáky Puky Péter Fund, Interspeech 2009

2007: student's research competition: BME VIK „TDK", 2nd place

2007: student's research competition: BME GTK „TDK", 1st place

2007: student's research competition: Országos „TDK", 1st place

2007: award: national research award for students (Köztársasági ösztöndíj, Ministry of Education)

2007: BME faculty award, BME VIK

2007: BME university award, BME

2007: travel grant: International Speech Communication Association, Interspeech 2007

2006: student's research competition, BME VIK „TDK", 1st place

Tartalomjegyzék/Table of contents

Subglottal resonances in older adults

Steven M. Lulich

Department of Speech, Language & Hearing Sciences, Indiana University, Bloomington, Indiana 47405, U.S.A.

Abstract

The subglottal acoustic input impedance partially consists of a series of resonances that are acoustically excited during voiced speech. Measurement of these subglottal resonances (SGRs) has thus far been restricted almost entirely to young adults and children. However, many aspects of the speech production system are known to change as adults age, and there is a possibility that SGR measurements might have clinical value in older adults. This study examined SGRs measured in 10 adults (5 males, 5 females) aged 50-68 years, and probed the dependence of SGR frequencies on vowel quality, posture, pulmonary function, as well as standing height, sitting height, weight, age, and gender. Previously reported data from young adults were also compared with the new data from older adults. Vowel quality affected the frequency of the first subglottal resonance (Sg1) (Sg1 was higher in [a:] than [i:] or [u:]), and age (older adult vs. younger adult) affected the frequencies of the first three subglottal resonance (Sg1, Sg2, Sg3) (SGRs were lower in older adults). Posture did not affect SGR frequencies, and no other significant relationships were found. The interaction of vowel quality with Sg1 is likely due to acoustic coupling between the subglottal and supraglottal (vocal tract) airways during phonation. The interaction between Sg1 and vowel quality was previously reported to be non-significant in younger adults, and the significant interaction in older adults could be due to age-related changes in laryngeal biomechanics and motor control. Based on previous modeling work, the interaction of age with Sg1, Sg2, and Sg3 is most likely due to age-related changes in the geometry and biomechanics of the subglottal airways, but empirical verification of this hypothesis is still needed.

1. Introduction

The subglottal airways comprised of the trachea and the bronchial tree resonate over a wide range of frequencies. Within the speech band (roughly 100 Hz - 8 kHz or more), the three lowest subglottal resonances (Sg1, Sg2, Sg3) are most likely to interact through acoustic coupling with the vocal tract resonances (Stevens, 1998; Chi & Sonderegger, 2007; Lulich, 2006, 2010; Cranen &

Email address: slulich@iu.edu (Steven M. Lulich)

Beszédtudomány – Speech Science 2025/1. 12–36. DOI: 10.15775/Besztud.2025.1.12-36

Boves, 1987; Fant et al., 1972; Klatt & Klatt, 1990; Hanson, 1996) and with the vibration of the vocal folds (Lulich & Arsikere, 2015; Zhang et al., 2006; Berry et al., 2014; Titze, 2008; Lulich et al., 2009). A number of studies since the 1960's (i.e. starting with van den Berg, 1960) have measured the frequencies of these SGRs, although almost exclusively in young adults (e.g. Lulich et al., 2012) and in children (Lulich et al., 2011b; Yeung et al., 2018). Young adults are the best-studied population in phonetics research, in no small part due to the ease of recruiting college and graduate students as participants. Children's speech has frequently been studied in order to investigate effects of growth and development. Older adults, on the other hand, are less easily recruited and have already attained their adult height and mature speech motor control, and so they have been participants in speech research much less frequently (Kent & Vorperian, 2018). With regard to SGRs, the only verifiable exceptions known to this author are Ishizaka et al. (1976), who measured SGRs in four laryngectomy patients between the ages of 40 - 72 years, and Hanna et al. (2018), whose 10 participants included four men aged 55-63 years.

Based therefore almost entirely on data from children and young adults, the following understanding of SGRs has emerged. With some caveats discussed below, the SGR frequencies are determined primarily by the effective acoustic length, $l_a = h/k_a$, of the tracheobronchial tree modeled as a simple uniform tube (Lulich et al., 2011a),

$$SgN = \frac{(2N-1)c}{4 \cdot h/k_a} \tag{1}$$

In this equation, $N = 1, 2, 3$ represents the Nth SGR, c is the speed of sound in cm/s, h is the speaker's standing height in cm, and k_a is an empirically determined scale factor relating speaker height to the length of the tracheobronchial tree's equivalent uniform tube (Yeung et al., 2018).

Several additional resonances associated with the cartilage and soft tissues of the tracheobronchial tree walls are also present in the same frequency range as Sg1-Sg3 (Lulich & Arsikere, 2015), although they are usually difficult to

identify in spectral analyses. The first soft tissue resonance, SgT, is at a frequency slightly lower than Sg1, and causes the Sg1 frequency to be somewhat raised relative to the quarter-wavelength prediction encapsulated in Equation 1. Furthermore, at higher frequencies the acoustic waves set up by the vibrating vocal folds penetrate less deeply into the tracheobronchial tree, resulting in a shortened value of l_a and a raised resonance frequency. This is observable in children's Sg3, although adults' Sg3 frequencies appear to be low enough that they are not substantially affected by the decreasing penetration depth. Finally, speech content (for example, vowel quality) does not appear to have a substantial impact on SGR frequencies, at least in young adults and children (Lulich et al., 2012; Yeung et al., 2018). For young adults, Sg1 is in the range 660 ± 47Hz for females, and 554 ± 42 Hz for males. Sg2 is in the range 1513 ± 69 Hz for females, and 1327 ± 77 Hz for males. Sg3 is in the range 2426 ± 101 Hz for females, and 2179 ± 126 Hz for males (Lulich et al., 2012).

Stevens (1998, 2002) posited that the subglottal resonances, through their interaction with vocal tract formants, form quantal 'acoustic berms' that separate the formants of front vowels from back vowels, and high vowels from low vowels. Subsequent studies investigated this in several languages (Chi & Sonderegger, 2007; Sonderegger, 2004; Lulich, 2006, 2010; Jung, 2009; Guo et al., 2014; Dogil et al., 2011; Madsack et al., 2008), including Hungarian (Csapó et al., 2009a; Gráczi et al., 2011); explored the development of this quantal hypothesis in children (Jung, 2009; Yeung et al., 2018); and extended it from vowel acoustics to the acoustic properties of stop consonant bursts (Lulich, 2010, 2008; Lulich & Chen, 2009). A single perceptual study of SGR influence on vowel identification was carried out by Lulich et al. (2007), in which a vowel-consonant (VC) formant transition crossed from above to below an antiresonance mimicking Sg2 in the microphone signal. The frequency and time at which the crossing occurred were manipulated, and this affected the rate at which listeners identified the vowel as a back vowel [a] or a central vowel [ʌ], consistent with the quantal hypothesis. Studies of speaker normalization by computers demonstrated that knowledge of Sg1 and Sg2 can improve performance of automatic speech recognition (ASR)

systems, especially when talker age (e.g. child vs. adult) and language (e.g. English vs. Mandarin) in the training and test sets are mismatched, and when limited training data is available (Wang et al., 2008a,b, 2009; Arsikere et al., 2012, cf. Morton et al. 2015, who demonstrated that humans can perform talker normalization based on exposure to a single excised vowel).

Whether the SGR quantal hypothesis applies to older adults, and whether knowledge of SGRs is helpful in ASR of older adults' speech, has not been investigated. Previous studies of vocal tract anatomy and acoustic properties have shown that the vocal tract dimensions and volume change with aging, with concomitant changes in vowel formant frequencies, especially the first formant (F1) (cf. Xue & Hao 2003 and Kent & Vorperian 2018 [section 8.2], and references therein). It has also been shown that the trachea dimensions change with aging (Sakai et al., 2010), as do the mechanical properties of the airway walls (Gibellino et al., 1985). Therefore, the relationships between subglottal resonances and vocal tract formants might differ in older adults compared with what is already known primarily from young adults and children.

A further variable that is known to affect speech production but has not been systematically studied across a large research program, and which has never been studied with regard to subglottal resonances, is body posture. The importance of learning about posture's effects on speech may be easily derived from the fact that most speech produced by most humans most of the time is not the kind of speech we typically study in laboratory settings, viz. seated upright in a quiet environment and without performing significant extraneous tasks. A specific example will suffice to make the point: During the height of the COVID-19 pandemic in 2020-2021, attempts were made to use speech data to help make diagnoses and determine severity of disease. Speech data were attractive because they can be obtained non-invasively, at a distance, inexpensively, quickly, and with little effort from patients. However, many such patients at that time were lying supine or reclining in a hospital bed, which are very different postures than the upright seated posture which informs the vast majority of our knowledge of speech, including subglottal resonances. Because the geometry of the subglottal

airways is not volitionally modifiable, unlike the vocal tract, it is not clear that the effects of posture (which includes the effects of gravity) on subglottal resonances can be compensated for in any way, and it is therefore desirable to investigate what these postural effects are.

Kent & Vorperian (2018) call for an 'ambitious program of research ... to determine age-related effects in speech' (p. 83), and the present study represents a first step in this direction with regard to subglottal resonances and their interactions with vowel formants in older adults. The main goals are to characterize the frequencies of Sg1, Sg2, and Sg3 in a sample of 10 older adults (5 males, 5 females), and to characterize relationships between the SGRs and 1) the vowel being produced, 2) the posture in which the vowel is produced, 3) measures of pulmonary function such as total lung capacity (TLC), forced vital capacity (FVC), forced expiratory volume in 1 second (FEV1), and functional residual capacity (FRC), and 4) speaker characteristics such as standing height, sitting height, weight, gender, and age.

2. Methods

Ten older adults (5 males, 5 females) aged between 50 and 68 years participated in this study (Table 1). All of the speech recordings and pulmonary function tests for this study were collected in a sound booth in the Speech Production Laboratory at Indiana University. Speech recordings were made with a SHURE KSM32 cardioid condenser microphone mounted on a stand located approximately 4 feet ($\approx 1.2m$) away from the participant's face. Subglottal acoustics recordings were made with a K&K Sound HotSpot accelerometer held against the skin of the neck below the thyroid cartilage. Pulmonary function tests (PFTs) were carried out in a Morgan Scientific whole-body plethysmograph, using Morgan Scientific's ComPAS software. Age (in years) was recorded by self-report, and standing height, sitting height, and weight were recorded using a SECA digital stadiometer in the Speech Production Laboratory. For sitting height measurements, participants were seated on a stool with a mea-

sured height of $29in$ ($73.66cm$). Sitting height was incorrectly recorded for one participant (participant 7), and this datum is therefore excluded from further analyses.

For the speech and subglottal acoustics recordings, participants produced sustained vowels [i:], [a:], and [u:] for approximately 5 seconds, in each of three postures: upright seated, supine, and left lateral decubitus (i.e. lying on one's left side). The duration of the sustained vowels was controlled by prompting each participant when to begin, providing feedback in approximately 1-second increments, and prompting when to stop. These prompts and feedback were made using simple hand gestures, and the quality of the recordings was continuously monitored for accuracy. Although vowel identity was previously shown to not influence SGR frequencies in young adults (Lulich et al., 2012), we recorded SGRs in three vowels in each posture, since we posited that either posture or age might affect the acoustic coupling between the vocal tract and subglottal airways. Each vowel/posture combination was recorded once, thus there were 3 vowels x 3 postures = 9 recordings from each participant. Vowel formants (F1, F2, F3) and fundamental frequency (f0) were measured from the microphone recordings, and SGRs (Sg1, Sg2, Sg3) were measured from the accelerometer recordings, as in previous studies (Lulich et al., 2012; Yeung et al., 2018).

The sampling rate for both the microphone and accelerometer signals was 48kHz before down-sampling to 8kHz. The formants and SGRs were measured in two ways. First, using the original recordings with 48kHz sampling rate, long-term averaged spectra (LTAS) were made from a series of 512-point hamming windows zero-padded to 1024 points, with a step size of 64 points, and the peaks in the LTAS were measured. These measurements were compared with a wide-band spectrogram as an additional accuracy check. The fundamental frequency was measured similarly, but with a 2048-point hamming window with no zero-padding, with a step size of 64 points. Second, after down-sampling to 8kHz, newly-calculated LTAS were viewed together with the average linear predictive coding (LPC) spectrum and with the wide-band spectrogram. The LPC spectra were calculated with order $p = 12$. Peaks in the LTAS and the

LPC spectra were generally in close agreement both with each other and with the wide-band spectrograms. Formant frequencies were almost always measured at the peaks of the LPC spectra, except in instances where the LPC spectrum had clearly missed the formant. SGR frequencies were more likely to be missed by the LPC spectrum, especially at low frequencies (Sg1), resulting in a larger proportion of SGR measurements being made from the LTAS. In all cases, both the formant and SGR measurements were compared with wide-band spectrograms as an additional accuracy check. Figure 1 shows an example of how the SGRs were measured from the LTAS and LPC spectra superimposed on a wide-band spectrogram. The two sets of formant and SGR measurements were made several months apart.

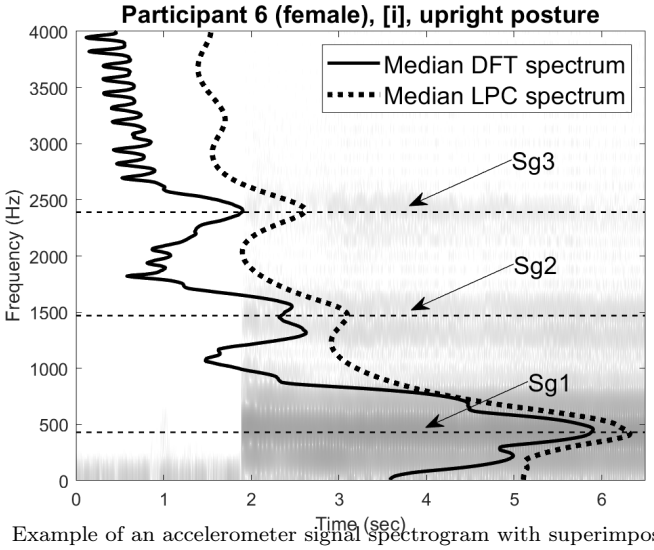


Figure 1: Example of an accelerometer signal spectrogram with superimposed LTAS and LPC spectrum (see text for details). In this example, all three SGRs are measured from the peaks in the averaged LPC spectrum, which aligns well with the peaks in the LTAS. Notice that the double peak in the LTAS around Sg2 is due to the prominence of the two nearest harmonics. In this example, there were approximately 2 seconds of rest, followed by nearly 5 seconds of the sustained vowel [i] produced by a female speaker (Participant 6) in upright posture. The sustained vowel is very stable, showing little acoustic variability throughout its duration.

The two sets of formant and SGR measurements were in excellent agreement, with Pearson correlation coefficients of $r = .9545$ for F1, $r = .9953$ for F2, $r = .8943$ for F3, $r = .6134$ for Sg1, $r = .8951$ for Sg2, and $r = .9544$ for Sg3. The relatively poorer correlation for Sg1 is consistent with previous findings that Sg1 is frequently difficult to measure accurately due to its relatively large bandwidth and its proximity to the loud low-frequency harmonics (Lulich et al., 2012; Yeung et al., 2018). For the remainder of this paper, only those formant and SGR measurements from the second set are used (i.e. from the LTAS + LPC spectra + spectrograms of the down-sampled signals).

For the PFTs, participants completed three successful vital capacity (VC) maneuvers and three successful thoracic gas volume (TGV) tasks. Success was determined automatically by the ComPAS software in accordance with American Thoracic Society (ATS) guidelines, and includes criteria such as the presence of a clear baseline of tidal breathing and absence of a breath-hold between deep inhalation and forced exhalation. The minimum number of attempts was 6 (3 successful VC maneuvers + 3 successful TGV tasks), and most participants completed the PFTs within 8 attempts. One participant was unable to complete 3 successful TGV tasks, and his TGV data are therefore excluded from analysis. From the VC maneuver, the forced vital capacity (FVC) was measured as well as the forced expiratory volume in one second (FEV1). From the TGV task, the total lung capacity (TLC) was measured as well as the functional residual capacity (FRC).

Parametric three-factor analyses of variance (ANOVA) were carried out to test for main and interaction effects of vowel, posture, and gender on fundamental frequency, formant frequencies, and SGR frequencies. Spearman correlation coefficients were calculated for the relationships between the independent variables (age, standing height, sitting height, weight, FVC, FEV1, FRC, TLC) and the dependent variables (f0, F1, F2, F3, Sg1, Sg2, Sg3).

Finally, to evaluate whether there is a difference in SGR frequencies between older and younger adults, we compared data from the present study with data reported in prior studies. Yeung et al. (2018) fit values for c and k_a in Equation 1

using young adult data from Lulich et al. (2012) and child data from Lulich et al. (2011b) and Yeung et al. (2018), and reported best-fit values of $c = 45,400 cm/s$ for Sg1 ($c = 35,900 cm/s$ for Sg2 and Sg3), and $k_a = 8.76$. Root-mean-squared errors in the prediction of Sg1, Sg2, and Sg3 from the resulting model were $49Hz$, $75Hz$, and $108Hz$, respectively, among young adults. The same model was applied to the data in the present study. Two-tailed, two-sample t-tests were used to test whether the model residuals were different in older and younger adults for Sg1, Sg2, and Sg3 ($\alpha = .05/3 \approx .01$ after Bonferroni correction). For younger adults, the mean Sg1, Sg2, and Sg3 values ($N = 50$) reported in Lulich et al. (2012) were used to calculate model residuals; for older adults, the Sg1, Sg2, and Sg3 values across all vowels and postures ($N = 90$) were used to calculate model residuals.

3. Results

The gender, age, standing height, sitting height, weight, and pulmonary function test results for each participant are presented in Table 1. Fundamental frequency (f0) and formant frequencies (F1, F2, F3) for each vowel and posture by participant are presented in Tables 2 and 3, and subglottal resonances for each vowel and posture by participant are presented in Tables 4 and 5.

Three-factor ANOVAs with vowel, posture, and gender as independent factors revealed significant main effects of gender on Sg1 ($p = .0012$), Sg2 ($p \ll .001$), and Sg3 ($p = .0086$), and a significant main effect of vowel on Sg1 ($p = .0057$), but no significant main effects of posture on any of the SGRs. There were also no significant interaction effects on the SGRs. Three-factor ANOVAs with vowel, posture, and gender as independent factors revealed significant main effects of gender on f0 ($p \ll 0.001$), F1 ($p = .0464$), F2 ($p \ll .001$), and F3 ($p \ll .001$), significant main effects of vowel on F1 ($p \ll .001$), F2 ($p \ll .001$), and F3 ($p \ll .001$), and a significant interaction effect of vowel and gender on F2 ($p = .0002$). There were no significant main effects of posture on f0 or the formants, and no additional significant interaction effects.

Table 1: Participant ID, gender, age (yrs), standing height (cm), sitting height (cm), weight (kg), FEV1 (L), FRC (L), FVC (L), and TLC (L).

Participant	Gender	Age	Standing Height	Sitting Height	Weight	FEV1	FRC	FVC	TLC
1	M	68	175.3	79.2	82	3.52	4.48	4.39	8.34
2	M	59	188.9	88.6	146	3.12	4.66	4.43	8.30
3	M	67	175.3	81.6	95	3.37	4.90	4.24	8.33
4	M	54	177.8	81.0	115	3.32	–	4.35	–
5	M	50	168.8	77.3	69	3.31	4.10	4.19	7.47
6	F	68	167.6	73.8	64	1.62	3.30	2.03	5.27
7	F	63	164.6	–	118	2.02	3.28	2.39	6.05
8	F	54	162.6	75.2	75	2.72	2.88	3.22	5.99
9	F	65	153.3	70.8	79	1.64	2.85	2.11	4.67
10	F	59	166.5	79.5	86	2.38	3.51	2.81	5.50

Table 2: Fundamental frequency (Hz) and formant frequencies (Hz) for each vowel by participant (males only) in three postures: seated upright, supine, and left lateral decubitus (lat. dec.).

		[i]				[a]				[u]			
Participant	Posture	f0	F1	F2	F3	f0	F1	F2	F3	f0	F1	F2	F3
1	Upright	141	289	2438	3102	164	836	1195	2750	164	297	1250	2320
1	Supine	141	273	2359	3109	141	703	1125	2641	141	242	1211	2164
1	Lat. Dec.	141	289	2313	3055	164	734	1000	2641	164	281	1180	2156
2	Upright	164	289	2148	2969	164	813	1094	2359	164	281	977	2180
2	Supine	164	328	2398	2602	164	1039	1164	2391	164	438	1023	2117
2	Lat. Dec.	164	305	2203	3508	164	711	1156	2211	164	328	867	2109
3	Upright	141	266	2391	3031	141	633	1078	2492	141	281	1141	1914
3	Supine	164	258	1891	2367	164	594	1078	2523	164	289	1063	2055
3	Lat. Dec.	141	250	1969	2828	141	570	992	2523	141	250	1164	2023
4	Upright	117	219	2305	2766	117	781	1133	2641	117	257	1078	2469
4	Supine	117	227	2367	2914	117	586	1055	2516	117	203	969	2461
4	Lat. Dec.	117	219	2383	2867	117	625	1070	2609	117	211	1141	2539
5	Upright	164	289	2109	2836	164	867	1273	2867	164	297	1188	2867
5	Supine	164	172	2344	2781	164	727	1203	2797	164	164	1195	2820
5	Lat. Dec.	164	148	2578	2945	141	695	1430	2977	141	391	1289	-

Table 3: Fundamental frequency (Hz) and formant frequencies (Hz) for each vowel by participant (females only) in three postures: seated upright, supine, and left lateral decubitus (lat. dec.).

		[i]				[a]				[u]			
Participant	Posture	f0	F1	F2	F3	f0	F1	F2	F3	f0	F1	F2	F3
6	Upright	234	477	2602	3172	211	602	1008	3086	211	445	1383	2594
6	Supine	211	422	2688	3164	211	750	1117	3102	211	398	1453	2797
6	Lat. Dec.	258	234	2477	3180	234	867	1070	3250	234	250	1602	2594
7	Upright	305	289	2680	3266	281	1094	1352	3008	305	383	1133	2867
7	Supine	258	266	2977	3118	258	969	1320	2922	281	258	992	-
7	Lat. Dec.	281	281	2719	3039	258	906	1227	3000	258	273	984	2914
8	Upright	258	234	2531	2773	258	953	1281	2789	258	250	1242	2727
8	Supine	281	258	2563	2977	281	1031	1289	2820	281	258	1086	2633
8	Lat. Dec.	281	250	2602	3000	281	1070	1320	2875	281	250	1094	2648
9	Upright	234	266	2563	3445	234	719	1070	2539	281	250	1102	2711
9	Supine	234	234	2664	3133	234	641	1234	2984	515	461	1141	2672
9	Lat. Dec.	258	242	2781	3273	258	711	1133	2250	281	266	883	2773
10	Upright	234	234	2477	3023	211	852	1219	2508	211	234	992	2867
10	Supine	211	195	2648	3570	211	719	1219	2633	234	211	1008	-
10	Lat. Dec.	234	227	2594	3523	211	586	1305	2352	211	258	1234	2766

Table 4: Subglottal resonance frequencies (Hz) for each vowel by participant (males only) in three postures: seated upright, supine, and left lateral decubitus (lat. dec.).

		[i]			[a]			[u]		
Participant	Posture	Sg1	Sg2	Sg3	Sg1	Sg2	Sg3	Sg1	Sg2	Sg3
1	Upright	516	1320	2398	477	1164	2141	508	1328	2297
1	Supine	477	1203	2313	484	1102	2609	266	1336	2102
1	Lat. Dec.	555	1274	2242	508	1273	2617	508	1219	2086
2	Upright	477	1094	-	406	1086	-	414	1086	-
2	Supine	477	1039	-	297	1102	-	469	1219	2133
2	Lat. Dec.	398	1047	-	289	1117	2063	359	1117	2148
3	Upright	523	1305	-	508	1344	-	500	1320	-
3	Supine	547	1211	2102	609	1234	2070	547	1227	-
3	Lat. Dec.	234	1234	2016	625	1219	2039	422	1227	2148
4	Upright	523	-	-	438	1125	2320	414	1164	2383
4	Supine	242	1164	-	602	1117	-	258	1141	-
4	Lat. Dec.	211	1141	1891	430	1125	1898	219	1148	2180
5	Upright	516	1336	1961	625	1438	1961	523	1344	1961
5	Supine	445	1234	2133	438	1430	-	461	1266	2320
5	Lat. Dec.	398	1289	1969	539	1344	2320	383	1305	-

Table 5: Subglottal resonance frequencies (Hz) for each vowel by participant (females only) in three postures: seated upright, supine, and left lateral decubitus (lat. dec.).

		[i]			[a]			[u]		
Participant	Posture	Sg1	Sg2	Sg3	Sg1	Sg2	Sg3	Sg1	Sg2	Sg3
6	Upright	414	1469	2406	781	1484	2500	398	1383	2516
6	Supine	406	1203	2352	336	1484	2352	406	1406	2344
6	Lat. Dec.	461	1359	2258	625	1227	2297	594	1414	2328
7	Upright	586	1383	-	547	1094	2406	586	1391	2406
7	Supine	531	1313	-	719	1438	-	539	1398	-
7	Lat. Dec.	531	1320	-	734	1453	2422	531	1406	-
8	Upright	469	1563	2445	484	1422	2484	484	1445	2406
8	Supine	469	1531	2477	781	1539	2547	484	1570	2609
8	Lat. Dec.	484	1305	2344	797	1563	2359	594	1594	2609
9	Upright	461	1398	2250	406	1469	2305	438	1398	2172
9	Supine	461	1547	2359	430	1438	2461	469	1563	2438
9	Lat. Dec.	477	1398	2398	422	1492	2242	477	1375	-
10	Upright	484	1258	2047	602	1406	2117	570	1258	2109
10	Supine	391	1289	2039	563	1336	2023	406	1406	2094
10	Lat. Dec.	422	1266	1992	586	1336	2086	633	1281	2000

All demographic and pulmonary function variables except for age were significantly correlated with Sg2. The Spearman correlation coefficients were $r = -.9240$ for standing height ($p < .00015$), $r = -.9167$ for sitting height ($p = .0013$), $r = -.6848$ for weight ($p = .0351$), $r = -.6606$ for FEV1 ($p = .0440$), $r = -.8061$ for FVC ($p = .0082$), $r = -.9333$ for FRC ($p < .00075$), and $r = -.7500$ for TLC ($p = .0255$). No other correlations between demographic or pulmonary function variables and SGRs were significant ($p > .065$). Fundamental frequency (f0) was significantly correlated with standing height ($r = -.8616$, $p = .0014$), FEV1 ($r = -.7853$, $p = .0071$), FVC ($r = -.7117$, $p = .0210$), FRC ($r = -.8984$, $p = .0019$), and TLC ($r = -.7289$, $p = .0316$). The third formant (F3) was significantly correlated with sitting height ($r = -.7333$, $p = .0311$), FEV1 ($r = -.6485$, $p = .0490$), and FVC ($r = -.7939$, $p = .0098$).

The f0, F1, F2, and F3 values (pooled across speakers and postures, but separated by gender; Figure 2, top panels) for the vowels [i], [a], and [u] were similar to previous reports. Whether pooled across speakers and vowels (Figure 2, middle panels) or across speakers and postures (Figure 2, bottom panels), all three SGRs were lower than expected in comparison with young adults. This was further reflected when standing height was accounted for (Figure 3). Two-tailed 2-sample t-tests with Bonferroni correction ($\alpha = .05/3 \approx .01$) showed that model residuals (Figure 4) for Sg1, Sg2, and Sg3 were all significantly lower in older adults than in younger adults ($p \ll .001$).

4. Discussion

In contrast to prior studies of SGRs by Lulich et al. (2012) and Yeung et al. (2018), which reported data from 50 and 43 participants, respectively, the present study is based on a smaller sample of 10 participants. Most of these participants (males and females) were of similar stature, with standing heights ranging primarily between $162cm$ and $178cm$. Participants were recruited from the community by word of mouth, and although some were relatives of other participants, they were related only through marriage and not genetically (e.g. two

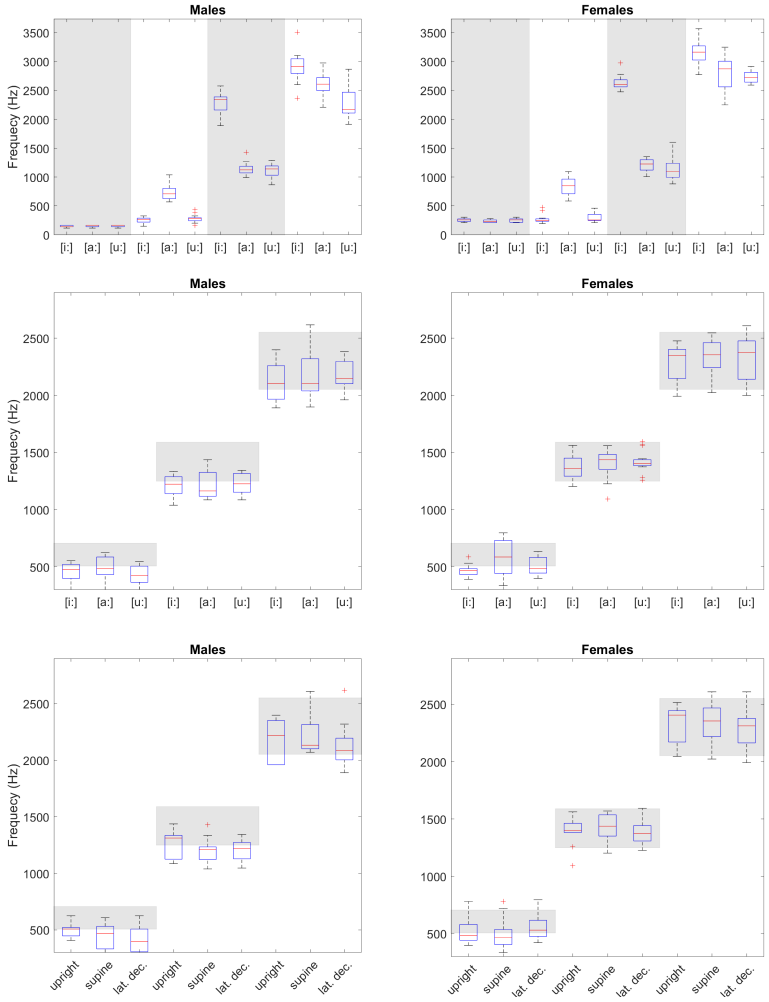


Figure 2: Distributions of fundamental frequency and formant measurements (Top Panels), and distributions of subglottal resonance measurements by vowel (Middle Panels) and by posture (Bottom Panels) for older adults. In the Middle and Bottom Panels, the shaded regions correspond to the expected frequency ranges based on data from Lulich et al. (2012).

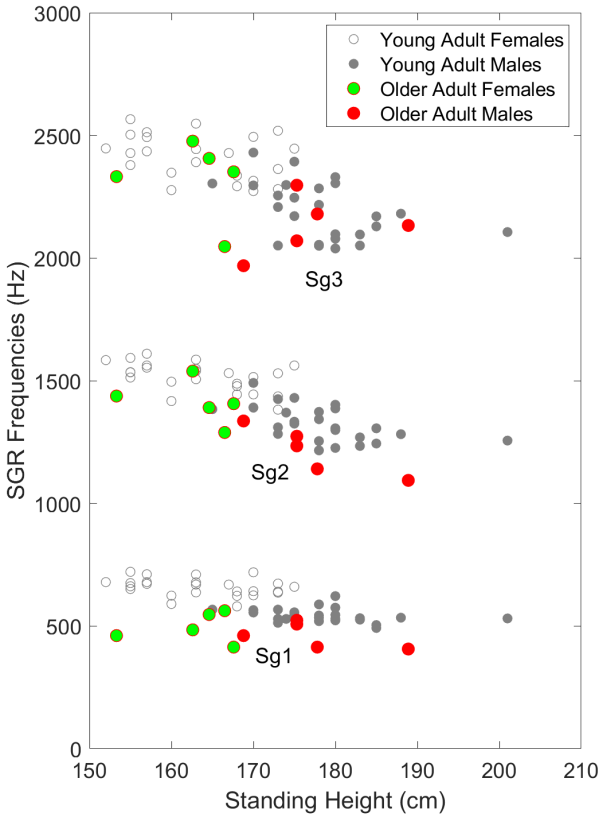


Figure 3: SGR frequencies (Sg1, Sg2, Sg3) as a function of standing height for the 50 young adults in Lulich et al. (2012) and the 10 older adults in the present study. Gray-scale symbols represent data from young adults; colored symbols represent data from older adults. The filled gray and red symbols represent data from male speakers; the open gray symbols and the filled green symbols represent data from female speakers.

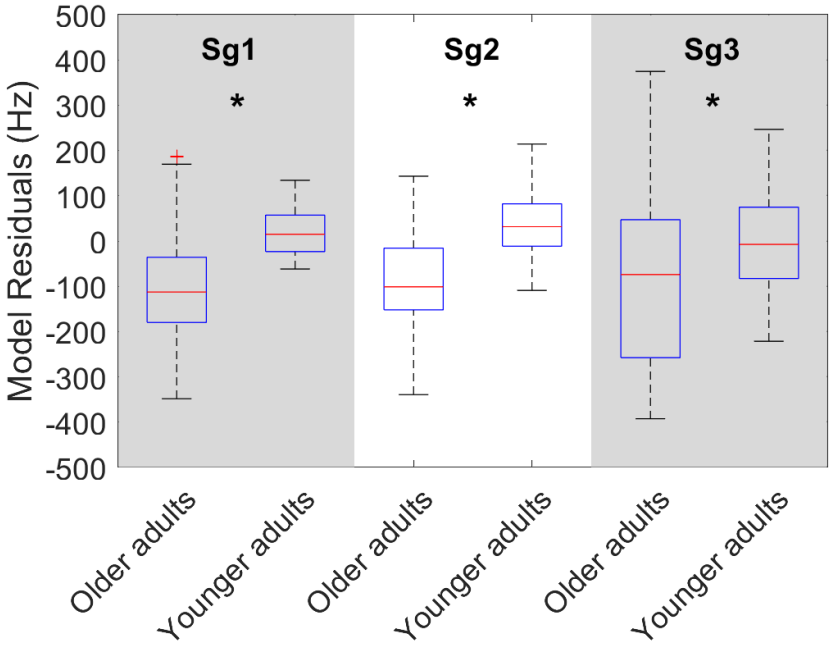


Figure 4: Residuals from applying the best-fit model of Yeung et al. (2018) to both younger adults (Lulich et al., 2012) and older adults (present study). The model used Equation 1 to predict SGR frequencies from standing height, an empirically determined scale factor, $k_a = 8.76$, and a speed of sound, $c = 35,900 cm/s$ for Sg2 and Sg3 frequencies and $c = 45,400 cm/s$ for Sg1 frequencies. The difference between older and younger adults was significant for all three SGRs.

participants were brother-in-law and sister-in-law). The total range of standing heights among the 10 participants is similar to the total range among participants in the study of young adults by Lulich et al. (2012), with the earlier study including only one male participant with a substantially greater standing height (cf. Figure 3). The small sample size in the present study was therefore sufficient to demonstrate the value of further investigations in this topic. Nevertheless, while the present study was intended as an initial exploration of the possible effects of age and posture on SGRs, future studies should recruit larger and more representative samples of older adults.

Posture was found not to have a significant effect on SGRs, f0, or formant frequencies in this study of healthy older adults in seated upright, supine, and left lateral decubitus postures. Vowel identity affected formant frequencies, as expected, but it also had a significant effect on Sg1. Vowel identity did not affect Sg2, Sg3, or f0. Gender affected all of the SGRs, f0, and formants. Some significant relationships were revealed between SGRs and formant frequencies on the one hand, and speaker characteristics (standing height, sitting height, weight, and age) and pulmonary function test results (TGV, FRC, FVC, FEV1) on the other hand, especially those involving Sg2 and f0. All three SGRs were found to be lower than expected in older adults compared with young adults. An explanation for this cannot be provided by the data presented here, but one possibility is that the decrease in SGR frequencies reflects aging-related changes in the geometry and biomechanics of the subglottal airways.

Subglottal resonances have shown promise in ASR and other speech-based technologies, such as speaker normalization and height estimation. The fact that SGRs are significantly lower in older adults than in younger adults suggests that technologies trained on data from younger adults may suffer performance degradation when applied to older adults, and training data from older adults is therefore desirable for future applications that target this demographic.

5. Conclusion

Subglottal resonances in older adults are not dependent on posture, and only Sg1 was dependent on vowel quality. All of the SGRs were lower in older adults than in younger adults, perhaps due to age-related changes in the geometry and biomechanics of the subglottal airways. With the caveat that the SGRs are lower in older adults than in younger adults, it appears that measurements of SGRs are otherwise comparable in healthy older and younger adults, regardless of posture. Future studies should continue to investigate the effects of healthy aging on subglottal acoustics with larger samples, and extend the sample characteristics to older adults with cardiopulmonary illnesses such as COVID, influenza, pneumonia, asthma, congestive heart failure, emphysema, and others. Not only will such studies contribute to our ability to use speech data in novel clinical settings, but they will also inform our understanding of speech production, phonetics, phonology, and motor control in a broader range of embodied conditions.

Acknowledgments

This research was supported in part by an Undergraduate Research Grant from the Indiana University Department of Speech, Language & Hearing Sciences awarded to Nicole Neuenschwander, who helped collect much of the data from the older speakers included in this study. Special thanks also go to Carey Smith for help with all of the data collection from the older speakers. Many thanks to the reviewer and the editors who provided feedback on the manuscript and thus helped to make it substantially better than it originally was; any remaining errors are, of course, my own.

I am grateful to the editors for organizing this special issue in memory of Dr. Tamás Gábor Csapó. I first met Dr. Csapó in 2009 when our mutual friend, Dr. Tamás Mihály Bőhm (1981-2019), arranged for me to teach a graduate course on speech acoustics via Skype at the Budapest Institute of Technology and Economics (many thanks to Dr. Klára Vicsi for facilitating this). Dr. Csapó

was one of the students enrolled in the course (as was Dr. Tekla Etelka Gráczi, one of the editors). Dr. Csapó's final paper in the course investigated subglottal resonances (SGRs) and their relationship with vowel formants in Hungarian. The principal data set for this final paper was a set of speech and subglottal (accelerometer) recordings from Dr. Bőhm, which I have used as pedagogical materials in several of my speech courses over the years. Dr. Csapó's class project led to a series of three papers on SGRs in Hungarian (Csapó et al., 2009a,b; Gráczi et al., 2011). Subsequently, Dr. Csapó spent 6 months as a Fulbright Scholar in my laboratory at Indiana University in 2014, where we worked together to develop 3D/4D ultrasound as a research tool for speech production, culminating in a paper on tongue surface segmentation errors in 2D ultrasound (Csapó & Lulich, 2009). Over the years, and through a handful of visits to Budapest that were never long enough, I was fortunate enough to continue my friendship with Dr. Csapó and his family, and I took a degree of personal pride in his many successes as a scholar and teacher. I was deeply affected to hear of his passing on January 31, 2024. I am grateful to have had the opportunity to join him for lunch on November 17, 2023, when I was in Budapest for the Symposium on Vowel Harmony. My last communication with him was when we wished each other a Merry Christmas on December 25, 2023. It is with sorrow over his passing, mixed with gratitude for his friendship and pride in his accomplishments that I wish to dedicate this paper to his memory.

References

Arsikere, H., Leung, G. K., Lulich, S. M., & Alwan, A. (2012). Automatic estimation of the first two subglottal resonances in children's speech with application to speaker normalization in limited-data conditions. In *Thirteenth Annual Conference of the International Speech Communication Association* (pp. 1267–1270). Portland, OR.

van den Berg, J. (1960). An electrical analogue of the trachea, lungs and tissues. *Acta Physiol. Pharmacol. Neerlandica.*, *9*, 361–385.

Berry, D., Neubauer, J., & Zhang, Z. (2014). Impact of subglottal resonances on bifurcations and register changes in laboratory models of phonation. *The Journal of the Acoustical Society of America*, *136*, 2259–2259.

Chi, X., & Sonderegger, M. (2007). Subglottal coupling and its influence on vowel formants. *The Journal of the Acoustical Society of America*, *122*, 1735–1745.

Cranen, B., & Boves, L. (1987). On subglottal formant analysis. *The Journal of the Acoustical Society of America*, *81*, 734–746. URL: `http://link.aip.org/link/?JAS/81/734/1`.

Csapó, T. G., Bárkányi, Z., Gráczi, T. E., Bőhm, T., & Lulich, S. M. (2009a). Relation of formants and subglottal resonances in Hungarian vowels. In *Tenth Annual Conference of the International Speech Communication Association* (pp. 484–487). Brighton.

Csapó, T. G., Gráczi, T. E., Bárkányi, Z., Beke, A., & Lulich, S. M. (2009b). Patterns of Hungarian vowel production and perception with regard to subglottal resonances. *The Phonetician*, (pp. 7–28).

Csapó, T. G., & Lulich, S. M. (2009). Error analysis of extracted tongue contours from 2d ultrasound images. In *Sixteenth Annual Conference of the International Speech Communication Association* (pp. 2157–2161).

Dogil, G., Lulich, S. M., Madsack, A., & Wokurek, W. (2011). Crossing the quantal boundaries of features: Subglottal resonances and Swabian diphthongs. *Tones and Features: Phonetic and Phonological Perspectives*, (pp. 137–148).

Fant, G., Ishizaka, K., Lindqvist, J., & Sundberg, J. (1972). Subglottal formants. *STL-QPSR*, *1*, 1–12.

Gibellino, F., Osmanliev, D. P., Watson, A., & Pride, N. B. (1985). Increase in tracheal size with age: implications for maximal expiratory flow. *American Review of Respiratory Disease*, *132*, 784–787.

Gráczi, T. E., Lulich, S. M., Csapó, T. G., & Beke, A. (2011). Context and speaker dependency in the relation of vowel formants and subglottal resonances-evidence from Hungarian. In *Twelfth Annual Conference of the International Speech Communication Association* (pp. 1901–1904). Florence.

Guo, J., Liu, A., Arsikere, H., Alwan, A., & Lulich, S. M. (2014). The relationship between the second subglottal resonance and vowel class, standing height, trunk length, and f0 variation for mandarin speakers. In *Fifteenth Annual Conference of the International Speech Communication Association* (p. Singapore). 930–934.

Hanna, N., Smith, J., & Wolfe, J. (2018). How the acoustic resonances of the subglottal tract affect the impedance spectrum measured through the lips. *The Journal of the Acoustical Society of America*, *143*, 2639–2650.

Hanson, H. M. (1996). Measurements of subglottal resonances and their influence on vowel spectra. *The Journal of the Acoustical Society of America*, *100*, 2656–2656. URL: `http://link.aip.org/link/?JAS/100/2656/2`.

Ishizaka, K., Matsudaira, M., & Kaneko, T. (1976). Input acoustic-impedance measurement of the subglottal system. *The Journal of the Acoustical Society of America*, *60*, 190–197. URL: `http://link.aip.org/link/?JAS/60/190/1`.

Jung, Y. (2009). *Acoustic articulatory evidence for quantal vowel categories: The features [low] and [back]*. Ph.D. thesis Massachusetts Institute of Technology.

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of communication disorders*, *74*, 74–97.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*, 820–857. URL: `http://link.aip.org/link/?JAS/87/820/1`.

Lulich, S. M. (2006). *The role of lower airway resonances in defining vowel feature contrasts.*. Ph.D. thesis Massachusetts Institute of Technology.

Lulich, S. M. (2008). On the relation between locus equations and subglottal resonances. In *Proceedings of Meetings on Acoustics* (pp. 1–10). AIP Publishing volume 5.

Lulich, S. M. (2010). Subglottal resonances and distinctive features. *Journal of Phonetics*, *38*, 20–32.

Lulich, S. M., Alwan, A., Arsikere, H., Morton, J. R., & Sommers, M. S. (2011a). Resonances and wave propagation velocity in the subglottal airways. *The Journal of the Acoustical Society of America*, *130*, 2108–2115.

Lulich, S. M., & Arsikere, H. (2015). Tracheo-bronchial soft tissue and cartilage resonances in the subglottal acoustic input impedance. *The Journal of the Acoustical Society of America*, *137*, 3436–3446.

Lulich, S. M., Arsikere, H., Morton, J. R., Leung, G. K. F., Alwan, A., & Sommers, M. S. (2011b). Analysis and automatic estimation of children's subglottal resonances. In *Twelfth Annual Conference of the International Speech Communication Association* (pp. 2817–2820). Florence.

Lulich, S. M., Bachrach, A., & Malyska, N. (2007). A role for the second subglottal resonance in lexical access. *The Journal of the Acoustical Society of America*, *122*, 2320–2327.

Lulich, S. M., & Chen, N. (2009). Automatic classification of consonant-vowel transitions based on subglottal resonances and second formant frequencies. In *Proceedings of Meetings on Acoustics* (pp. 1–8). AIP Publishing volume 6.

Lulich, S. M., Morton, J. R., Arsikere, H., Sommers, M. S., Leung, G. K., & Alwan, A. (2012). Subglottal resonances of adult male and female native speakers of american english. *The Journal of the Acoustical Society of America*, *132*, 2592–2602.

Lulich, S. M., Zanartu, M., Mehta, D. D., & Hillman, R. E. (2009). Source-filter interaction in the opposite direction: Subglottal coupling and the influence of vocal fold mechanics on vowel spectra during the closed phase. In *Proceedings of Meetings on Acoustics* (pp. 1–14). AIP Publishing volume 6.

Madsack, A., Lulich, S. M., Wokurek, W., & Dogil, G. (2008). Subglottal resonances and vowel formant variability: A case study of high german monophthongs and swabian diphthongs. *Proc. LabPhon*, *11*, 91–92.

Morton, J. R., Sommers, M. S., & Lulich, S. M. (2015). The effect of exposure to a single vowel on talker normalization for vowels. *The Journal of the Acoustical Society of America*, *137*, 1443–1451.

Sakai, H., Nakano, Y., Muro, S., Hirai, T., Takubo, Y., Oku, Y., Hamakawa, H., Takahashi, A., Sato, T., Chen, F. et al. (2010). Age-related changes in the trachea in healthy adults. In *Oxygen Transport to Tissue XXXI* (pp. 115–120). Springer.

Sonderegger, M. A. (2004). *Subglottal coupling and vowel space: An investigation in quantal theory*. Ph.D. thesis Massachusetts Institute of Technology.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*, 1872–1891. URL: `http://link.aip.org/link/?JAS/111/1872/1`.

Titze, I. R. (2008). Nonlinear source–filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America*, *123*, 2733–2749.

Wang, S., Alwan, A., & Lulich, S. M. (2008a). Speaker normalization based on subglottal resonances. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4277–4280). IEEE.

Wang, S., Lulich, S. M., & Alwan, A. (2008b). A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation. In *Ninth Annual Conference of the International Speech Communication Association* (pp. 4277–4280). Brisbane.

Wang, S., Lulich, S. M., & Alwan, A. (2009). Automatic detection of the second subglottal resonance and its application to speaker normalization. *The Journal of the Acoustical Society of America*, *126*, 3268–3277.

Xue, S. A., & Hao, G. J. (2003). Changes in the human vocal tract due to aging and the acoustic correlates of speech production. *Journal of Speech, Language, and Hearing Research*, *46*, 689–701.

Yeung, G., Lulich, S. M., Guo, J., Sommers, M. S., & Alwan, A. (2018). Subglottal resonances of american english speaking children. *The Journal of the Acoustical Society of America*, *144*, 3437–3449.

Zhang, Z., Neubauer, J., & Berry, D. A. (2006). The influence of subglottal acoustics on laboratory models of phonation. *The Journal of the Acoustical Society of America*, *120*, 1558–1569.

A nyelvgesztusok ultrahangos vizsgálata a magyar beszédben

Markó Alexandra[1], Deme Andrea[2,1,3], Juhász Kornélia[3,1,2], Gráczi Tekla Etelka[3,1]

[1] *MTA–HUN-REN NYTK Lendület Neurofonetikai Kutatócsoport*
[2] *Eötvös Loránd Tudományegyetem*
[3] *HUN-REN Nyelvtudományi Kutatóközpont*

Abstract

Tamás Gábor Csapó, as one of the founding members of MTA–ELTE "Lendület" Lingual Articulation Research Group, introduced the technique of ultrasound tongue imaging (UTI) into articulatory research in Hungary. He mastered this technique during a Fulbright Scholarship at Indiana University, Bloomington. He did pioneering work in the development of UTI data analysis methodology, silent speech interfaces, and other digital applications of UTI. His achievements are acknowledged and his work is referenced worldwide. In the present paper, after a brief outline of the technique of 2D ultrasound tongue imaging, we review the research, results, and application areas in which Tamás Gábor Csapó (with his co-authors) made his marks within the fields of articulatory research and articulatory-to-acoustic conversion.

1. Bevezetés

Csapó Tamás Gábor (1985–2024) honosította meg Magyarországon az ultrahangos képalkotó technikát a beszéd artikulációs működéseinek vizsgálatára, amelyet Fulbright-ösztöndíjasként az Indianai Egyetemen, Bloomingtonban sajátított el, Steven Lulich vezetése alatt. Csapó Tamás Gábor úttörő munkát végzett a nyelvultrahangos elemzési módszertan, a némabeszéd-interfész és egyéb beszédtechnológiai alapú digitális alkalmazások fejlesztése terén, nemzetközi viszonylatban is. Tanulmányunkban a nyelvultrahang módszertanának vázlatos ismertetése után áttekintjük azokat a kutatási és alkalmazási területeket, ame-

Email addresses: marko.alexandra.phd@gmail.com (Markó Alexandra), deme.andrea@btk.elte.hu (Deme Andrea), juhasz.kornelia@nytud.hun-ren.hu (Juhász Kornélia), graczi.tekla.etelka@nytud.hu-ren.hu (Gráczi Tekla Etelka)

Beszédtudomány – Speech Science 2025/1. 37–73. DOI: 10.15775/Besztud.2025.1.37-73

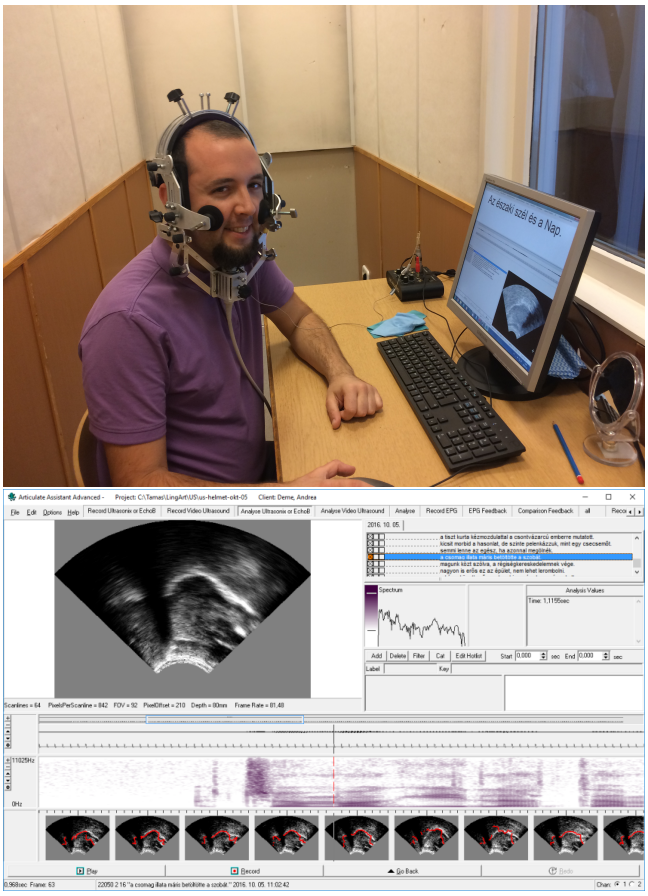
lyeken Csapó Tamás Gábor – mindazon kutatócsoportokkal, amelyekben dolgozott – maradandót alkotott.

2. A nyelvultrahang mint artikulációs vizsgálati eszköz

Az akusztikai elemzés holisztikusságával ellentétben az artikulációs vizsgálati módszerek általában csak részleges információt nyújtanak: szinte mindegyik az artikulációs gesztusoknak csak egy kis részhalmazához fér hozzá, és gyakran csak korlátozott térbeli és/vagy időbeli felbontású információt szolgáltatnak. A teljes artikulációs csatorna vizsgálata még mindig technológiai kihívást jelent, költséges, és a legtöbb műszer bizonyos mértékig beavatkozik a beszédképzés folyamatába.

A nyelvultrahang (ultrasound tongue imaging: UTI, magyarul jellemzően UH) egyetlen aktív beszédszerv vizsgálatára szolgál: a nyelv felszínét képezi le (bár bizonyos feltételekkel – lásd alább – a szájpad is kirajzoltatható a segítségével, illetve esetenként a gége helyzetének követésére is használják, vö. Kochetov, 2020). Az ultrahang olyan képalkotó technika, amely ultramagas frekvenciájú hanghullámot bocsát ki egy piezoelektromos kristály gerjesztése révén. Ez a hanghullám áthalad a szöveten és visszaverődik annak felületéről (Stone, 2005). Az áll alatt elhelyezkedő ultrahangos jelátalakító így készít képet a nyelvről. A (kétdimenziós) ultrahangfelvételek tipikus eredménye szürkeárnyalatos képek sorozata, amelyeken a nyelv felszíni kontúrja a környező szöveteknél és a levegőnél nagyobb fényerősséggel látszik (1. ábra alsó panel).

Az ultrahangos technika alkalmazását a beszédszervek működésének vizsgálatára az 1960-as évek végén Kelsey, Minifie és Hixon (1969) javasolták, elsősorban a röntgennel végzett vizsgálatok alternatívájaként, amelyet egészséges emberek esetében a sugárterhelés miatt indokolatlanul kockázatosnak tartottak. Az ultrahang ezzel szemben – ahogy írják (564) – biztonságos, gyors, és nem okoz kényelmetlenséget a vizsgálati személy számára. Ennek megfelelően egyre gyakrabban alkalmazzák a fonetikai kutatásban, különösen az elmúlt húsz évben (Kochetov, 2020). Az artikulációs kutatások módszertanának nemzetközi nép-



1. ábra. Csapó Tamás Gábor az ultrahangfejet rögzítő sisakban (fent, a sisak az Articulate Instruments Ltd. terméke) és az Articulate Assistant Advanced (AAA) szoftver képernyőképe (lent) (Csapó et al., 2017a: 343; Csapó & Xu, 2020: 3735).

szerűségi listáján a második helyet a nyelvultrahang foglalja el az elektromágneses artikulográfia (EMA) után: a nemzetközi fonetikai folyóiratokban 2000 és 2019 között megjelent publikációkban az EMA a tanulmányok 29%-ában, az ultrahang 17%-ukban szerepelt (vö. Kochetov, 2020).

A nyelvultrahang előnye, hogy a teljes nyelvfelszín működéséről képet ad, nem csak a nyelv (és más beszédszervek) néhány meghatározott pontjáról (mint az EMA). Egyszerűen használható, elérhető árú, valamint nagy felbontású (akár 800×600 pixel) és nagy sebességű (akár 100 képkocka/s) felvétel készíthető a

segítségével (Csapó et al., 2017a). A jó térbeli felbontás azért fontos, hogy a nyelv alakjáról minél pontosabb képet kapjunk, míg a jó időbeli felbontás ahhoz szükséges, hogy a beszédhangok képzésének gyors változását (pl. zárfelpattanás, koartikuláció) is vizsgálni lehessen. Előnye még, hogy a beszélők döntő többségénél jól alkalmazható, nem okoz különösebb nehézséget a használata, mivel jól tolerálható. Az ultrahang további előnyei közé tartozik az is, hogy mérete és súlya miatt hordozható, így például a gyermekkorú vagy a más sérülékeny populációba tartozó beszélőket nem szükséges fonetikai laboratóriumban vizsgálni, a természetes közegükben is felvehetők a kísérletek. Mindezen túlmenően más műszerekkel, elsősorban elektroglottográffal (EGG, más néven laringográf) is kombinálható, így egyszerre rögzíthető és vizsgálható a nyelv artikulációs tevékenysége és a hangszalagműködés.

Mindezzel szemben hátránya az, hogy kizárólag egyetlen beszédszerv, a nyelv mozgását teszi megfigyelhetővé, és meglehetősen érzékeny a beszélő szöveteinek állapotára (pl. zsírosság, hidratáltság), továbbá bizonyos mértékben a testméretekre is (mert a nagyobb üregekben kevésbé ad jó képet, mint a kisebbekben) (vö. pl. Stone, 2005; Csapó et al., 2017a). További nehézséget jelent az ultrahangfej rögzítése a mozgó artikulátoron (állkapocs), valamint az adatkinyerés (a nyelvkontúrok dokumentálása és elemzése).

Artikuláció közben a szájpad nem látható az ultrahangképen, mert az ultrahangsugár visszaverődik a szájüregbeli levegőről, ugyanakkor a nyelvhelyzetet illetően a nyelv és a szájpad távolsága alapvető fontosságú információ lehet. Nyelés során azonban a nyelv szinte teljes mértékben érintkezik a szájpaddal, így a szájpad kontúrja megfigyelhetővé válik (Epstein & Stone, 2005). A szájpad vonala ily módon rögzíthető az ultrahangképen, illetve referenciaként használható a nyelvhelyzet vizsgálatában.

További viszonyítási alapot jelent a harapási sík felvétele (Scobbie et al., 2011). A harapási síkot úgy rögzíthetjük, hogy a két fogsor között egy merev lemezt helyezünk el, amelyre a kísérleti személy ráharap. A nyelvet a lemez aljához támasztva a nyelvkontúr kirajzolja a harapási síkot.

A kétdimenziós nyelvultrahang (amelyet a magyar beszéd vizsgálatára eddig használtak) a nyelv szagittális (oldalnézeti keresztmetszeti) középvonalának mozgását rögzíti (Stone, 2010). Az ultrahangfejnek az áll alatt történő pozicionálására sisak használatos (erre látható példa az 1. ábra felső képén). Az ultrahangfelvétellel egyidejűleg a beszédet általában a sisakra csíptetett mikrofonnal veszik fel. A videó és a hang szinkronizált felvételét, a felvételek megjelenítését és az elemzést valamilyen szoftver teszi lehetővé, illetve támogatja. Az 1. ábra alsó képén erre látunk példát: az Articulate Assistant Advanced (AAA) szoftver bizonyos ablakaiban egyszerre látszik az ultrahangkép, a beszéd hullámformája, FFT-spektruma és spektrogramja. Emellett az AAA szoftver az ábra alján látható módon automatikus nyelvkontúrkövetésre is alkalmas, tehát (valamennyi hibával) képes az ultrahangképeken automatikusan letapogatni a nyelv felszínének körvonalát.

Ugyan az ultrahang láthatóvá teszi a nyelv felszíni kontúrját (1. ábra alsó panel), de ez a kontúr közvetlenül még nem elemezhető. Ahhoz, hogy a nyelvalakot és annak változásait vizsgálni lehessen, a rögzített képsorozatból valamilyen módszerrel ki kell nyerni a nyelv körvonalát (illetve ennek adatait). Ez kivitelezhető manuálisan: ennek során a képernyőn megjelenő ultrahangképre egy célszoftver segítségével az egér mozgatásával berajzoljuk a nyelv kontúrját. Ez azonban rendkívül időigényes. A másik lehetőség, hogy automatikus módszerekkel kíséreljük meg a nyelv körvonalának azonosítását (vö. az 1. ábra alsó képsorán a piros vonalakat). Ezek pontossága azonban még nem éri el a manuális kontúrkinyerését, ezért mindenképpen szükség van emberi ellenőrzésre és javításra, ha automatikus módszereket alkalmazunk (Whalen et al., 2019).

3. A magyar beszéd vizsgálata ultrahangos technikával: rövid történeti áttekintés

A magyar nyelvre vonatkozóan az ultrahang hazai megjelenését megelőzően kevés olyan vizsgálat született a lingvális artikulációról, amely dinamikus adatokon (azaz nem csak statikus állóképeken) alapult, és amely valós időben rögzí-

tette a beszédszervek mozgását (nem pedig utólag dokumentálta a lenyomatot, mint például a fotópalatográfia). Lotz János az 1960-as években (1966), Szende Tamás az 1970-es években (1974), majd Bolla Kálmán az 1980-as években (1981a; 1981b) röntgenfilm (ún. röntgenogram/kinoröntgenografikus vizsgálat) technológiával vizsgálta a nyelvgesztusokat. A kétezres évek elején elektromágneses artikulográfiával készültek különféle vizsgálatok külföldi laboratóriumokban (Mády, 2008; Deme et al., 2016, 2017).

Habár a nemzetközi szakirodalomban a nyelvultrahang alkalmazása az artikulációs kutatásokban az 1960-as évektől dokumentálható (Kelsey et al., 1969), az általunk ismert első olyan ultrahangos vizsgálat, amelyet a magyar beszédre vonatkozóan végeztek el, Stefan Beňuš és Adamantios I. Gafos 2007-ben publikált kísérlete, amelyet a Haskins Laboratóriumban elektromágneses artikulográf (3 beszélő) és nyelvultrahang (1 beszélő) segítségével rögzítettek. A kutatók a magyar nyelv egy sajátos jelenségét, a magánhangzó-harmóniát, illetve azon belül az áttetsző magánhangzók artikulációs jellemzőit tanulmányozták. Noha a kísérlet anyagával kapcsolatban felmerülnek kérdések (vö. Markó et al., 2019a), Beňuš és Gafos (2007) arról számolnak be, hogy az eredményeik szerint az antiharmonikusan viselkedő magánhangzók esetén hátulsóbb nyelvhelyzet volt mérhető, mint a harmonikusan viselkedők esetében.

A hazai tudományban a nyelvultrahangos vizsgálatok 2016-tól kezdődtek meg, amikor az MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport pályázati támogatása lehetővé tette nagyobb értékű artikulációs műszerek beszerzését. Csapó Tamás Gábor a kutatócsoport alapító tagjaként jelentős részt vállalt a nyelvultrahangos technika itthoni implementálásában. Az első kísérletek nagy része módszertani-feltáró jellegű volt (ezek többségét nem publikálták, de később lásd pl. Csapó & Xu, 2020; Csapó et al., 2021), majd a folyamatos magyar beszéd olyan jelenségeire összpontosítottak (a teljesség igénye nélkül), mint például a gemináta és szingleton zár- és réshangok artikulációs eltérései a nyelvhelyzet tekintetében (Percival et al., 2020a,b,c); a zöngétlen szibilánsok artikulációs mintázatának longitudinális változása gyermekeknél (Gráczi et al.,

2021a); a zöngés és zöngétlen obstruensek ejtési sajátosságainak eltérései (Gráczi et al., 2020a,b, 2021b).

A kezdetektől fogva zajlottak kísérletek az UH-adatok alapján beszédtechnológiai alkalmazások fejlesztésére is (pl. Csapó et al., 2020), majd Csapó Tamás Gábor több egyéni és csoportos pályázati projektje is ezt tűzte ki célul, elsősorban az ún. némabeszéd-interfész témakörében: 2017–2021 NKFIH FK-17 (124584) „Artikulációs mozgás alapú beszédgenerálás"; 2018–2021 NKFIH PD-18 (127915) „Némabeszéd-interfészek kidolgozása és a beszélőfüggőség vizsgálata nyelvultrahang és elektromágneses artikulográf eszközökkel"; 2022–2026 NKFIH FK-22 (142163) „Az artikuláció és az agyi jelek elemzése beszéd alapú agy-gép interfészhez".

A Csapó Tamás Gábor által az MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport laboratóriumában rögzített ultrahangfelvételek alapján a Miskolci Egyetem és a Debreceni Egyetem kutatói is végeztek méréseket, összehasonlításokat az MRI- és az ultrahangképek harmonizációjára, összekapcsolására (lásd pl. Trencsényi, 2020b,a; Trencsényi & Czap, 2021).

4. Módszertani kérdések a nyelvultrahang használata és az adatfeldolgozás kapcsán

4.1. A nyelvultrahang alkalmazása kísérleti helyzetben

Tekintettel arra, hogy az ultrahangkészülék beszerzésekor az MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport tagjainak nem volt még tapasztalata a technika használatával kapcsolatban, és korábban Csapó Tamás Gábor is csak tipikus felnőtt beszélőkkel és az ultrahanggal magában (más műszerrel való összekapcsolás nélkül) végzett kísérleteket, kezdetben számos megvalósíthatósági tanulmány elkészítésére volt szükség. Csapó Tamás Gábor vezetésével zajlottak próbakísérletek a gyermeki artikuláció vizsgálatára vonatkozóan (ennek konklúzióit lásd Markó et al., 2017); az ultrahang és az elektroglottográf (EGG) összekapcsolásával (ennek tapasztalatait a későbbi publikációkban használták fel, pl. Percival et al., 2020a); illetve az ultrahang és az EGG kapcsolt

alkalmazásával egy fiatalkori Parkinson-kórban szenvedő beteg beszédének elemzésére is sor került (Gráczi et al., 2018).

4.2. A nyelvkontúr kinyerése

A kétdimenziós ultrahangfelvételek szürkeárnyalatos képsorozatok, amelyeken a nyelv felszíni kontúrja a környezetéhez képest nagyobb fényerősségű vonalként jelenik meg. Ez azonban önmagában nem alkalmas a kvantitatív elemzésre. Ahhoz, hogy a nyelv kontúrján elemzéseket lehessen végezni (például összehasonlítani a nyelvformákat, méréseket végezni a nyelv görbületére vonatkozóan), a nyelvkontúrt (pontosabban annak számszerű adatait, illetve az egyes pontok koordinátáit) célszoftver segítségével ki kell nyerni. Mint említettük, a manuális kontúrrajzolás, bár pontos, rendkívül időigényes, az automatikus kontúrkinyerés ugyanakkor gyors, azonban hibák torzítják. A kézi kontúrrajzolás és a különféle automatikus módszertanok összehasonlítása, hibaelemzése és ennek révén a gépi technikák továbbfejlesztése a kezdetektől foglalkoztatja a kutatókat. Csapó Tamás Gábor egyik első, az ultrahangos technikára vonatkozó, Steven Lulichcsal végzett kutatása is erre a kérdéskörre irányult (Csapó & Lulich, 2015). A vizsgálat célja a kétdimenziós ultrahangos képsorozatokból történő midszagittális (középmetszeti) nyelvkontúrok meghatározásakor fellépő hibák jellemzése volt. A kísérletben hét személy – két szakértő és öt egyetemi hallgató – végezte el négy beszélő ejtéséből származó, összesen 1145 nyelvultrahangkép kézi berajzolását, és a kontúrokhoz kapcsolódó variabilitást számszerűsítették a kutatók. Ugyanezekből a felvételekből automatikusan is nyertek ki nyelvkontúrokat az EdgeTrak (Li et al., 2005), a TongueTrack (Tang et al., 2012) és az AutoTrace (Fasel & Berry, 2010) algoritmusok segítségével, majd ezeket kvantitatív módszerrel összehasonlították a kézileg berajzolt nyelvkontúrokkal. Az eredmények szerint az ultrahangkép minősége nagyobb mértékben befolyásolta a kézi kontúrrajzolás pontosságát, mint a kontúrt berajzoló személy gyakorlottsága, szakértelme. Emellett a kézi kontúrok pontosabbak voltak a gépi kontúroknál, illetve az algoritmusok közül az AutoTrace 3.5 bizonyult a legpontosabbnak. A szerzők három alapvető hibafajtát azonosítottak: 1. a program képtelen nyo-

mon követni a gyorsan mozgó nyelv kontúrját; 2. a zajból adódó hibák, amelyek a kép rossz minőségére vezethetők vissza; valamint 3. amikor a nyelv egyes részeit nem követi le a program. Ezek a hibatípusok jelzik a javításra szoruló területeket a jövőbeli fejlesztések során. A hibamértékek azonosítására további tesztek is készültek (vö. pl. Csapó & Csopor, 2015.

4.3. A nyelvkontúr adatainak feldolgozása

Csapó Tamás Gábor többféle módszerrel kísérletezett a nyelvkontúradatok számszerűsítése érdekében. A legközelebbi szomszéd távolsága (nearest neighbour distance, NND) módszerrel a nyelvkontúrok közötti távolság számszerűsíthető (Zharkova & Hewlett, 2009) a következőképpen: U-val és V-vel jelöljük a két nyelvkontúrt, illetve $U = u_1, u_2, \ldots u_n$ és $V = v_1, v_2, \ldots v_m$ a nyelvkontúrok pontjai. Első lépésként sorban meghatározzuk mindegyik u_i ponthoz a hozzá legközelebbi v_j pont távolságát; majd sorban meghatározzuk mindegyik v_i ponthoz a hozzá legközelebbi u_j pont távolságát; végül kiszámítjuk az összes fenti módon számított távolság átlagértékét. A legközelebbi szomszéd távolság számítási módjából következik, hogy amennyiben a két nyelvkontúr látszatra közel van egymáshoz, és hasonló hosszúságúak, akkor a köztük lévő távolság kicsi lesz; illetve amennyiben a két nyelvkontúr látszatra távol van egymástól, és hasonló hosszúságúak, akkor a távolságuk nagy lesz. Ugyanakkor ha a két nyelvkontúr látszatra ugyan közel van egymáshoz, de nem egyforma hosszúságú, akkor ott a számított távolság ugyancsak viszonylag nagy lesz. Csapó Tamás Gábor ezt a módszertant alkalmazta például a gyermeki artikuláció vizsgálatában (pl.Markó et al., 2017, 2019b).

Más kutatásokban (pl. Gráczi et al., 2021a) a nyelvkontúrok összehasonlításának alapjául Csapó Tamás Gábor a $\mathrm{LOC_{a\text{-}i}}$ indexet (Zharkova et al., 2015) alkalmazta. Ebben a módszerben a nyelvkontúr két szélső pontját összekötik egy egyenessel, majd az egyenes harmadolópontjaira merőlegest illesztenek, és kiszámítják a nyelvkontúr és az egyenes közötti távolságot ennél a két harmadolópontnál. Az index értéke ezeknek a távolságoknak az aránya, és ezzel számszerűen jellemezhető a nyelvkontúr.

Egy másik fajta, a nyelvkontúrok összehasonlítására alkalmas módszertant fejlesztett ki Csapó Tamás Gábor és Maeda Percival a magyar szingleton és geminá̧ta mássalhangzók összehasonlítására (lásd 5.1; Percival et al., 2020a,b,c). Ebben a módszerben a nyelvkontúrokat a beszélő harapási síkja által definiált koordináta-rendszerre forgatták, majd koronális, veláris és faringális régiókra osztották őket (Recasens & Rodríguez, 2016 alapján). Minden régió esetében a poláris koordinátákat lineáris kevert regressziós modellekkel elemezték annak meghatározására, hogy az ultrahangfej és a nyelv felszíne közötti radiális (sugárirányú) távolság mely régióban tér el a mássalhangzók között (tükrözve a nyelv alakjának és helyzetének különbségeit).

Felmerül a kérdés, hogy a nyelvkontúrok összehasonlítására mely módszertanok a legalkalmasabbak, illetve melyik módszer melyik kutatási kérdés(ek) megválaszolására a legalkalmasabb. A különböző szakirodalmi források különféle mérési módszertanokat alkalmaztak, így ha a saját eredményeinket egy adott tanulmány eredményeivel tervezzük összevetni, hasonló módszert is kell alkalmaznunk. A legközelebbi szomszédok közötti távolság becslésénél, mint említettük, fontos szempont, hogy a két nyelvkontúr hasonló hosszúságú-e. Lényeges az is, hogy a teljes nyelvkontúr elemezhető legyen, amit azonban az alsó állkapocs és a nyelvcsont árnyéka befolyásolhat. A szibilánsok elsajátításának vizsgálatában Zharkova és munkatársai (2015) a nyelvgörbületet számszerűsítették: a $\mathrm{LOC}_{\mathrm{a\text{-}i}}$ és más hasonló görbületmérőszámok a nyelvformáról közvetetten, egy számadattal adnak becslést. Mivel ebben az eljárásban a nyelv legelülső és leghátsó (látható) pontját kötjük össze egy egyenessel, és az erre az egyenesre állított merőlegesekkel dolgozunk, az eredmény ismét nagymértékben függ attól, hogy mekkora rész látszik a nyelvből. A Lingvális Artikuláció Kutatócsoport mindezek miatt azokban az esetekben, amikor nem egy-egy tanulmánnyal/szerző eredményével összevetve, hanem általánosságban kíván vizsgálni egy jelenséget a teljes nyelvkontúrok összehasonlításával, akkor az SS-ANOVA vagy a GAMM (általánosított kevert lineáris modellek) statisztikai elemzési módszertanokat alkalmazza, mivel az ezek által adott eredmények pontosabban képezik le a valóságot. Ezekben a módszerekben több nyelvkontúrra statisztikai eljárással gör-

bét illesztünk, amely egy becsült nyelvkontúrt eredményez a kérdéses jelenség leírására. Ezek esetében is célszerű a valószerűtlenül rövid nyelvkontúrokat elhagyni, ugyanakkor a statisztikai modellezés során a kisebb bizonytalanságok kellő számú adat esetében kisebb eséllyel okoznak téves eredményt. Mindezeken túl azonban további módszerek is léteznek, valamint hangsúlyozzuk, hogy ezek az általunk vizsgált kérdésekre legjobb módszertanok, nem végeztünk általános összehasonlító kutatást.

4.4. A beszélők és a felvételek eltéréséből adódó variancia kezelése az adatokban

Mint láttuk, az ultrahangfelvételek számos okból variábilisak, és ezek közül csak egy az artikulációra inherensen jellemző változatosság. Magából az ultrahangos módszertanból is következik, hogy az ultrahangfej helyzete még rögzítő sisak használata esetén is eltér két felvétel között, de egy felvétel közben is változhat egyazon beszélőn belül is. Az ultrahangból adódó további variabilitási tényező a felvételek minősége és ezáltal a nyelvkontúrok minősége (például hossza). A beszélők közötti változatosság szintén közhelynek számít a fonetikában. Mindezek a tényezők korrumpálhatják a nyelvkontúrok számszerűsítésére használatos mérőszámok (lásd például 4.3) érvényességét is.

Máig nem áll rendelkezésre olyan módszer, amely megoldaná ezeket a problémákat, azaz a különböző alkalmakkor és/vagy különböző beszélőkkel rögzített ultrahangfelvételek tekintetében lehetővé tenné a normalizálást, az illesztést vagy az adaptálást. Csapó Tamás Gábor Kele Xuval dolgozott olyan módszereken, amelyek segíthetik e problémák megoldását. 2020-ban publikált tanulmányukban mindenekelőtt az ultrahangfej pozíciójának eltéréséből adódó hibák kvantifikálására tettek kísérletet, és ennek kapcsán a harapási sík felvételét látták megoldásnak. Meglátásuk szerint a harapási sík alapján egy referencia koordináta-rendszert kell felvenni (vö. Scobbie et al., 2011), amelyhez a nyelvkontúrok a megfelelő szögben illeszthetők (beforgathatók), és így összehasonlíthatóvá válnak. Ennek a módszernek kétségkívül hátránya az, hogy a nyelvkontúrok kinyerése mellett további, nagyrészt csak manuálisan elvégezhető munkával jár, a tanulmány írásakor ugyanis erre nem állt rendelkezésre automa-

tikus módszertan. További nehézséget jelent véleményünk szerint az, hogy az ultrahangfej felvétel közbeni elmozdulására nem jelent megoldást, ha a felvétel készítője nem észleli ezt azonnal, és nem rögzíti újra a harapási síkot, amely alapján a referencia koordináta-rendszer újból létrehozható.

A témában további elemzések is készültek (Csapó et al., 2021), és Csapó Tamás Gábor tervezte a jövőben egy automatikus osztályozó kifejlesztését, amely a nyelvkontúr elemzése során figyelmeztet, ha az ultrahangfej helyzete eltérő egy felvételen belül; illetve amely konfidenciaintervallumokat rendel az adatok megbízhatóságához. Úgy gondolta, hogy a nyelv semleges pozícióban való felvétele (az egyes megnyilatkozások elején vagy végén), és e referenciakép változásának a regisztrálása segítség lenne a felvétel közben adódó problémák automatikus észleléséhez is.

5. Alapkutatások: a nyelv artikulációs működésének tanulmányozása

Az MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport Csapó Tamás Gábornak a nyelvultrahanggal kapcsolatos szakértelme révén több olyan alapkutatást is végzett, illetve kezdett meg, amelyek hozzáférhetővé tettek a magyar nyelvvel és beszéddel kapcsolatban korábban nem ismert adatokat, illetve bővítették a tudásunkat. Ezek a kutatások magánhangzók és mássalhangzók artikulációjára, valamint az ezek kapcsolatában jelentkező koartikulációs jelenségekre irányultak, nemcsak felnőtt beszélők, hanem gyermekek esetében is.

Az ultrahangfelvételeken alapuló alább bemutatandó vizsgálatokban Csapó Tamás Gábornak meghatározó szerepe volt számos tekintetben. Egyfelől fontos kiemelni, hogy ezek a kutatások a Csapó Tamás Gábor által kiépített szakmai kapcsolatokon (és erre épülő kutatói mobilitáson) alapultak. Továbbá az is hangsúlyozandó, hogy ezeknek a kutatásoknak az ultrahangos módszertanában is újszerű megoldásokat eredményezett Maeda Percival és Csapó Tamás Gábor közös munkája (5.1): közösen dolgozták ki az elemzés technikáját, és Magyarországon, illetve a magyar nyelvre irányuló elemzésben elsőként alkalmaztak EGG és UH eszközöket összekapcsoltan adatfelvételre. A további kutatásokban (5.2,

5.3) is ehhez hasonlóan a Csapó Tamás Gáborral közös munka eredménye a felvételi és elemzési módszertanok tesztelése, összevetése, majd az adekvát eljárás kiválasztása és alkalmazása (vö. még 4.3).

5.1. A nyelv működése a gemináta és szingleton zár- és réshangok artikulációjában

A magyar nyelvben a szingleton-gemináta oppozíció, azaz a nyelvileg rövid és hosszú mássalhangzók kontrasztja fonológiai, azaz jelalak-megkülönböztető szerepű szembenállás, pl. *ép* vs. *épp, tol* vs. *toll, ara* vs. *arra.* A leírások szerint a magyar mássalhangzórendszerben a szingleton-gemináta mássalhangzó-oppozíció legfőbb akusztikus kulcsa az ejtés időtartama. Emellett hagyományosan azt is feltételezik, hogy geminátak csak bizonyos pozíciókban, csak intervokálisan, illetve megnyilatkozás végén (magánhangzó után és szünet előtt) fordulhatnak elő a magyarban. Így nem állhatnak például szó elején, illetve nem fordulnak elő egy további mássalhangzóval szomszédosan. Amennyiben (például toldalékolás vagy szóképzés miatt) mégis előállna az utóbbi eset, a gemináta mássalhangzó (a legtöbb esetben) rövidül, azaz degeminálódik (Siptár & Törkenczy, 2000/2007).

A geminátak és a degemináció fonetikai vonatkozásait korábban már több idegen nyelvben is vizsgálták, és a magyarra vonatkozóan is napvilágot látott néhány, az akusztikai szerkezetre fókuszáló elemzés. Ezek fő kérdése jellemzően az volt, hogy mi különbözteti meg a gemináta és szingleton mássalhangzókat a hangok akusztikai megvalósításában. A kérdést felpattanó zárhangok és zár-rés hangok körében elemezték, és azt találták, hogy csakúgy, mint más nyelvekben (ehhez vö. pl. Ridouane, 2007), a magyarban is az időtartam, a felpattanó zárhangok esetében pedig különösen a zárképzés időtartama a fonológiai szingleton-gemináta oppozíció legfontosabb akusztikai korrelátuma (Olaszy, 2006; Pycha, 2009, 2010; Neuberger, 2015; Siptár & Gráczi, 2014).

Az MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport megalakulásával és az új artikulációs eszközök beszerzésével újabb kérdések nyíltak, amelyekre a csoport vizsgálatokat tervezett. Egyrészt elemezték azt, hogy a gemi-

náta felpattanó zárhangok rövidüléséből (degeminálódásából) előálló rövid hang ejtésében megegyezik-e a további (elvben azonos minőségű) szingleton hangokkal, különös tekintettel az artikulációs mozdulatok időbeli szerveződésére, időzítésére (ennek elemzésére EMA-t alkalmaztak az akusztikai elemzés mellett, vö. Deme et al., 2019b,a). Másrészt vizsgálták, hogy a gemináta felpattanó zárhangok és réshangok a szingletonokhoz képest nagyobb artikulációs erőfeszítéssel képzettek-e, amelyekben a nyelv inkább eléri az artikulációs célt – ez a vizsgálat ultrahanggal készült (Percival et al., 2020b,c). Harmadrészt pedig elemezték azt is, hogy hogyan alakul a gemináta felpattanó zárhangok és réshangok zöngéssége, tekintettel arra, hogy a zönge fenntartása nagyobb időtartamban artikulációsan nehéz – ennek vizsgálatára ultrahangot és elektroglottográfiát is használtak összekapcsoltan (Percival et al., 2020a).

A hosszú, degeminálódó hosszú, intervokális rövid és más rövid mássalhangzó mellett álló mássalhangzók EMA-n alapuló vizsgálata azt mutatta, hogy a rövidülő geminátáknak és a rövid mássalhangzóknak artikulációs és akusztikai tekintetben is eltérnek: a degemináció folyamata nem a megfelelő intervokális szingleton időtartamára redukálta a rövidülő gemináta mássalhangzókat, hanem sokkal inkább azokhoz a szingletonokhoz tette őket analóggá, amelyek a degeminációs pozícióhoz hasonlóan jobbról egy eltérő képzési helyű mássalhangzóval szegélyezettek, azaz kéttagú kapcsolatok tagjai. Az artikulációs adatok továbbá azt is felfedték, hogy a rövidülő geminátáknak és a két rövid mássalhangzóból álló kapcsolatoknak az említett időtartam-paraméterek tekintetében egy, a szingletonoknak és a geminátáknak közti átmeneti kategóriát képeznek. A zárképzés artikulációs korrelátumaként értelmezett artikulációsgesztus-platók átfedése a hangkapcsolatokban ezeken túlmenően azt is megmutatta, hogy a rövidülő geminátáknak és a szingletonkapcsolatoknak csak a lingvális-labiális képzéshelyi sorrend esetén térnek el időzítésüket tekintve (azaz a /tp/ gesztusok időzítése nem volt azonos a /ttp/ gesztusokéval, mégpedig olyan módon, hogy a /ttp/-ben megjelenő zárképzési szakaszok közti késleltetés nagyobb volt, mint a /tp/ esetében). A labiális-lingvális (azaz a /pt/ és /ppt/) kapcsolat gesztusidőzítése ezzel szemben nagyon hasonló volt, és átfedést mutatott a zárképzési szakaszokat mutató

artikulációs „platók" között. Ezt úgy értelmezhetjük, hogy a gemináták rövidülésének, azaz a degeminációnak a folyamata a szomszédos mássalhangzók képzési helyének függvényében rövidíti a hosszú mássalhangzókat kapcsolatbeli rövid mássalhangzókká – „hatékonyabban" a „könnyebb" ejtésű, eleve nagyobb artikulációs átfedést mutató /ttp/, és kevésbé hatékonyan a „nehezebben" ejthető /ppt/ esetében. A fentebbi összefüggéseknek a leírása nem csak a magyar nyelv vonatkozásában számít újszerűnek. Végül azt is megállapították, hogy a gemináták és a két szingletonból álló mássalhangzó-kapcsolatok előtt hosszabb a magánhangzók időtartama, ami abból következik, hogy ezekben az esetekben a nyelv lassabban emelkedik a mássalhangzó zárképzéséhez, mint a szingletonok esetében (Deme et al., 2019b,a). Ez az eredmény megcáfolta a más nyelvekre, kisebb beszélőszám elemzésével kapott adatok alapján megfogalmazott feltételezéseket.

Ultrahanggal kerestek választ arra a kérdésre, hogy hogyan alakul az artikulációs célok megközelítése a magyar szingleton és gemináta mássalhangzókban. Más nyelvekre, így például a japánra, a koreaira, az olaszra és a (keleti) oromóra kapott korábbi eredmények ugyanis azt mutatták, hogy a geminátákban nagyobb a kontaktusfelszín, magasabb és laposabb a nyelv, és előrébb van a nyelvtest, tehát összességében a nyelv jobban eléri a feltételezett artikulációs célját, mint a szingletonokban (Payne, 2006; Kochetov & Kang, 2017; Percival et al., 2018). A magyar nyelvre kapott adatok jelentősen árnyalták az addig kialakított képet: a zöngés és zöngétlen bilabiális, alveoláris és veláris felpattanó és réses képzésű szingletonok és gemináták elemzése azt mutatta, hogy nem egyértelmű a nyelv előrébb tolulása, tehát az artikulációs cél jobb megközelítése a geminátákban. Ez pedig arra utal, hogy nyelvfüggő lehet az, hogy a gemináták létrehozása fortizációként, azaz erősítésként jelentkezik a szingleton megfelelőikhez képest, hiszen kétféle következtetés vonható le mindebből. Vagy arról van szó, hogy a korábban elemzett nyelvekben erősebb a fortizáció a geminátákban és/vagy a lenizáció (az artikuláció lazulása) a szingletonokban, mint a magyarban, vagy pedig arról, hogy a gemináták fortizációja a magyarban más

természetű, mint a további elemzett nyelvekben, és ez más artikulációs mérőeszközökkel lehet inkább feltárható (Percival et al., 2020b,c).

A világ nyelveiben viszonylag ritkák a zöngés gemináta mássalhangzók, valószínűleg azért, mert a zöngésség fenntartása hosszabb időtartamban nehezebb. Ennek megfelelően egyes eredmények a zöngés geminátáknak részleges zöngétlenedésére utalnak olyan nyelvekben (pl. a tokiói japánban), ahol ezek a ritkább hangok egyáltalán előfordulnak (Kawahara, 2015). A magyarban léteznek zöngés geminátáknak, amelyek közül a bilabiális, alveoláris és veláris felpattanó és réses képzésűeket elemezték, illetve ezek szingleton párjait, valamint mindezek zöngétlen megfelelőit. Az EGG-adatok elemzése szerint nem volt statisztikailag kimutatható hatása a zöngésség időarányára annak, hogy az adott hang gemináta vagy szingleton-e, tehát a vizsgálat nem támasztotta alá azt a feltevést, hogy a zöngés geminátáknak részben vagy egészen zöngésednének vagy teljes mértékben zöngések lennének (fonetikai tekintetben). Ugyanakkor a zöngés hangok jelentős (és a zöngétlenekben látottnál nagyobb) változatossággal valósultak meg ebben a tekintetben. A nyelv helyzetét leképező UH-adatok pedig azt mutatták, hogy a beszédhangok fonológiai zöngéssége nem jelzi előre a nyelv előre mozdulását, sokkal inkább a fonetikai zöngésség. Ebből a szerzők arra következtettek, hogy a magyarban nem valószínűsíthető, hogy a nyelvgyök előre tolása a zöngésség fenntartásának előre tervezett mozzanata lenne – de ennek a feltevésnek az alátámasztása még további vizsgálatokat igényel (ehhez vö. 5.3).

5.2. Gyermekek lingvális artikulációjának vizsgálata

Az anyanyelv-elsajátítás vizsgálata a magyar gyermekek körében régóta népszerű kutatási téma, ugyanakkor megfelelő műszerek híján az artikulációs elemzések az említett kutatócsoport megalakulása előtt teljesen hiányoztak. Mivel a nyelvultrahang a nemzetközi kutatások alapján jól alkalmazható a gyermeki artikuláció regisztrálására is (vö. pl. Zharkova et al., 2015, 2018), a magyar gyermekek ejtéséről is készültek elemzések ezzel a módszerrel: vizsgálták a magánhangzók koartikulációs jellemzőit és a zöngétlen réshangok ejtésének longitudinális változását egyazon beszélőn belül.

Az artikulációs csatorna születéstől felnőttkorig jelentős változásokon megy keresztül, mind méreteit, mind arányait, mind pedig az egyes beszédképző szervek működési precizitását és összehangoltságát tekintve. A különböző életkori szakaszokban az érintett szervek méretbeli változása eltérő ütemű (Seikel et al., 2010), és mindezek a változások természetszerűleg hatnak a toldalékcső rezonátorsajátosságaira is. Az egyes képzőszervek koordinációjának fejlődése sem egyenletes: például az állkapocsgesztusok korábban érnek (azaz válnak hasonlóbbá a felnőtt működéshez, a gesztus téri-idői kivitelezéséhez), mint az ajakmozgás, és még később rögzülnek a nyelvgesztusok. Ez utóbbinak egyrészt az az oka, hogy a nyelvnek mint beszédszervnek a beszédbeli kontrollálása komplexebb feladat; másrészt pedig az, hogy a nyelv mozgása beszéd közben nem látható, így a nyelvmozgás felnőtt mintájának imitálása is nehezebb a gyermekek számára (Goffman, 2015). Habár a gyermekek úgynevezett fonetikai jártassága, a beszédhangoknak az elvárásoknak megfelelő ejtésére való képesség 8 éves korra kialakul, a finommotorikus ügyesség még később is jelentősen fejlődik, és ez a folyamat tizenéves korra is kitolódik (Vorperian & Kent, 2007). A beszéd motoros rendszerének érésével az orofaciális struktúrák közötti dinamikus koordináció egyre konzisztensebbé válik, ennélfogva az artikulációs gesztusok egyre kevésbé variábilisak az életkor előrehaladtával (egészséges fejlődés esetén) (vö. pl. Terband et al., 2011).

Tekintettel arra, hogy a magánhangzók nem akadályhangok, beszéd közben nem kapunk jól megfigyelhető taktilis visszajelzéseket a magánhangzók létrehozását eredményező közelítés vagy szűkület létrejöttének helyéről, azaz a magánhangzó képzéshelyéről. Egyes magánhangzók képzésekor a nyelv érintkezhet a nem mozgatható artikulátorokkal – például a nyelvperem a felső nyelvállású magánhangzók esetében hozzáér a felső fogsorhoz (vö. Stevens, 1998) – és ezek a szenzoros visszajelzések feltehetőleg hozzájárulnak ahhoz, hogy a beszélő az egyes magánhangzók ejtéséhez szükséges artikulációs beállításokat pontosan eltalálja, ez azonban magánhangzófüggő. Emellett a magánhangzók izolált ejtése statikus, így kinesztetikus információk sem segítik a magánhangzók képzéshelyének azonosítását, miközben a nyelv igen nagy variabilitással mozgat-

ható és alakítható (Ashby, 2011). A magánhangzók ejtése rendkívül variábilis is (egyes magánhangzóké nagyobb, másoké kisebb mértékben, bizonyos képzési és akár nyelvfüggő sajátosságokkal összefüggésben, vö. Deme, megj.), hiszen a nyelv alakjának kontrollálása (a nyelv izmainak komplex szerkezete és a motoros ügyesség szükséges foka miatt) igen nehéz feladat.

A gyermekek koartikulációjával számos nemzetközi tanulmány foglalkozik, magyarra azonban kevés adatunk van. Artikulációs szempontú elemzés pedig kettő, Markó és munkatársai (2019b; 2019c) vizsgálatából. A gyermekek beszédelsajátításában az egyik kulcsfontosságú feladatnak egyes szerzők annak a megtanulását tartják, hogy egy adott szegmentum létrehozásához szükséges artikulációs gesztusokat együttesen, egyre inkább szinkronban hozzák létre a gyermekek – a téri-idői átfedés pontos mértéke teszi az artikulációt felnőttszerűvé (Nittrouer et al., 1996). Általánosságban a gyermekek ejtésében erőteljesebb koartikulációs hatásokat adatoltak, mint a felnőtteknél (Terband et al., 2011). A magyar vizsgálatokban a 9 magyar magánhangzó-minőséget elemezték VCVCVC szerkezetben; a résztvevő gyermekek életkora 7 év 10 hónap és 14 év 8 hónap között szóródott. A két vizsgálatban kétféle elemzés történt, az elsőben az NND-t (a legkisebb szomszéd távolságot) határozták meg, azaz az azonos logatomok ismétlései között a nyelvkontúrok hasonlóságát elemezték. A második elemzésben pedig a nyelv és a szájpad közötti legkisebb távolságot számították ki. Az eredmények alapján általános tendenciaként az életkor előrehaladtával csökkent az ejtési variabilitás, ugyanakkor jelentősek voltak az egyéni különbségek. A nyelvkontúrbeli variabilitás nem függött össze az időtartambeli variabilitással a gyermekeknél.

A nemzetközi irodalomban két zöngétlen szibiláns, az /s/ és az /ʃ/ gyermeki artikulációját vizsgálta Zharkova több publikációban. Bár hasonló nyelvformamintázatokat kapott 10–12 éves gyermekek ejtésében, mint a felnőtteknél (Zharkova, 2016), mégis még serdülőkorban is talált artikulációs eltéréseket a felnőtt beszélők produkciójától (Zharkova et al., 2018). Emellett a koartikulációs rezisztencia (a hangkörnyezet hatásának való ellenállás) és az akusztikai eredmények eltérőek voltak a felnőtt ejtéstől (Zharkova et al., 2011, 2012). Ezen

eredményekre alapozva egy longitudinális esettanulmányban elemezték két magyar anyanyelvű testvér /s/- és /ʃ/-ejtését nyelvultrahang segítségével. A felvételek négy időpontban készültek, a lány 7;5, 7;11, 8;5, 8;11 éves és a fiú 11;0, 11;6, 12;0, 12;5 éves korában. A gyermekek nyelvkontúradatait az édesanyjuk (43 év) ejtésével vetették össze (Gráczi et al., 2021a). Az artikuláció számszerűsítésére az$LOC_{a\text{-}i}$ módszert választották Zharkova és munkatársai (2015) alapján (lásd 4.3). A felnőtt nő ejtésében a $LOC_{a\text{-}i}$ eltért a két mássalhangzó között. A lánygyermek ejtésében mind a négy időpontban, míg a fiúgyermek ejtésében az utolsó felvétel idejében találtak szignifikáns eltérést a két mássalhangzó között. Mivel a lánygyermek fiatalabb volt a felvételek készültekor, így fordított mintázatot várnánk. Ugyanakkor a beszédképző szervek nem egyenletes és egymással sem azonos ütemű növekedése (Vorperian et al., 2009), illetve az ehhez való folyamatos artikulációs alkalmazkodás miatt is előfordulhatott, hogy az idősebb gyermeknél egy ideig kevésbé jelent meg a szembenállás a nyelvkontúr (pontosabban annak az $LOC_{a\text{-}i}$ indexének a) vonatkozásában, holott a két mássalhangzó megkülönböztetése az ő esetében is stabilan és konzekvensen észlelhető volt. A nyelvkontúrokra kapott eredmények nem mutattak jelentős eltérést a mássalhangzók kezdete, közepe és vége között, tehát végeredményben a mássalhangzó teljes időtartamán hasonló tendenciákat találtak a mássalhangzók elkülönítését illetően.

A gyermekek artikulációjának kapcsán nyelvultrahanggal végzett vizsgálatok eredményei anyanyelvi tantárgymódszertani nézőpontból is szolgáltak hozadékokkal. A közoktatásban 5. és 9. osztályban tananyag a magánhangzók képzése, amelyet részben percepciós (vö. magas, mély), részben artikulációs (vö. nyelvállású, ajakkarekítéses, ajakréses) alapú terminusokkal jellemeznek a tankönyvek. Az absztrakt képzési jegyek tanítása – különösen, ha ezeket olyan jegyeknek tartjuk, amelyek külön-külön egy az egyben megfeleltethetők artikulációs mozdulatoknak, képzőszervi helyzeteknek – egészen bizonyosan nehézséget okoz, hiszen az artikulációs és vizuális tapasztalat legalábbis részben ellentmond az absztrakt képzési jegyeknek (Markó et al., 2020). Egy 5. osztályos fiú (a felvétel időpontjában 11;5 éves volt) és egy 9. osztályos fiú (a felvétel

időpontjában 14;8 éves volt) ultrahang-felvételeiről a 9 magyar sztenderd nyelvi magánhangzó-minőséget (VpVpV szekvenciák közepéről) reprezentáló kontúrok alapján az volt leolvasható, hogy a különböző korú beszélők között igen nagy az eltérés a szájpad és a nyelv közötti távolságot tekintve a szájtér és a nyelv méretkülönbségeiből adódóan. A felső nyelvállású magánhangzók esetében jól elkülönült a hátul képzett /u/ és az elöl képzett /i/ és /y/ mindkét beszélő esetében, az elöl képzettek kontúrja viszonylag fedte egymást. A középső nyelvállásúak esetében mind a nyelv vízszintes, mind a nyelv függőleges helyzete nagyobb variabilitást mutatott, mint ami a tankönyvi kategóriák alapján várható lett volna. Mindkét beszélőnél az látszott, hogy az /o/ és az /eː/ magasabb nyelvhelyzettel képződött, mint az /ø/, valamint hogy a nyelv vízszintes helyzetét tekintve az /ø/ nyelvkontúrja köztes pozíciót foglalt el az /o/ és az /eː/ nyelvkontúrja között. A vízszintes nyelvhelyzetet tekintve az 5. osztályos beszélő esetében az /aː/ nyelvkontúrja köztes helyzetet foglalt el az /ɒ/ és az /ɛ/ nyelvkontúrja között; a 9. osztályos beszélő esetében ez kevésbé látszott jól, mivel a nyelv hátulsó része szinte egymást átfedve jelent meg az /ɒ/ és az /aː/ nyelvkontúrján, de a nyelv formáját tekintve az /aː/ esetében az ív a szájtér közepén is magasan volt még, amiből következtetni lehet az /aː/ előrébb képzett voltára. Mindezek alapján artikulációs alapon mind a nyelvállásfok, mind az elöl-hátul képzettség kategorizálása bizonytalan, a határok elmosódottak, és kérdéseket vetnek fel. Ha feltételeznénk is, hogy a beszélő pontos fiziológiai visszacsatolással rendelkezik a nyelvhelyzetről a különböző magánhangzók kiejtése közben, és ezeket össze tudja hasonlítani egymással, majd ezen összevetések alapján képes kategóriákat elkülöníteni, ezek a kategóriák a lingvális képzési jegyek esetében korántsem mutatnák az elméleti kategóriák alapján csoportosítható magánhangzó-ejtési mintázatokat. Ehhez hozzáadódik még az életkorral összefüggésben a szájüreg méretbeli (és egyéb morfológiai) eltérése (például a szájpad alakja), valamint az egyének közötti és az egyénen belüli ejtési variancia is, amely ezeknek a kategóriáknak a határait igencsak elmosódottá teszi.

5.3. A zöngés és zöngétlen obstruensek, valamint a megelőző magánhangzók ejtési sajátosságai

A magyar nyelvre vonatkozó ultrahangos vizsgálatok további vonulata a zöngés és zöngétlen obstruensek ejtési sajátosságaira irányul. A fonetikában közismert tény, hogy a fonáció fenntartása szájüregi akadály jelenlétében különféle artikulációs manővereket igényel, mivel a szupralaringális és szájüregi nyomásnövekedés a hangszalagrezgés leállását okozhatja (vö. pl. Stevens, 1998). Az egyik ilyen artikulációs manőver a nyelvgyök előrébb mozdítása a zöngés szegmentumok ejtésekor a zöngétlen párhoz képest – főként lingvális obstruensek ejtésekor – a garatüregi régió megnagyobbítása érdekében (Narayanan et al., 1995). Egyes tanulmányok szerint a zöngés obstruensekben (különösen a réshangokban) előrébb tolódott a nyelvgyök e mássalhangzók zöngétlen párjához képest, de olyan mássalhangzópárt is találtak, amelyre ez nem volt jellemző (Ahn & Davidson, 2016). A magyarra vonatkozóan eddig a /z/ és az /s/ ejtésének összehasonlítása történt meg, azt a kérdést vetve fel, hogy van-e összefüggés a nyelvgyökhelyzet variabilitása és i) a fonációs arány, illetve ii) a nyelvhegy mozgása között (Gráczi et al., 2020a,b, 2021a), illetve, hogy a képzési mód és hely, valamint a kontextus milyen hatással van erre a jelenségre. Az első eredmények (a szibilánsokra vonatkozóan) összefüggést mutattak a mássalhangzók képzési helye, a fonáció beszélőspecifikus aránya és a nyelvgyök pozíciója között (Gráczi et al., 2023). Jelenleg is zajlik a nyelvgyök pozíciójának, a mássalhangzó képzési módjának és helyének, valamint a szomszédos magánhangzók artikulációs jellemzőinek kapcsolatát vizsgáló kutatás, kiterjesztve az összes magyar zöngés és zöngétlen obstruensre.

6. Alkalmazott kutatások: a nyelvultrahang-adatok felhasználása a beszédszintézisben

A beszédszintézis, vagyis a mesterséges beszéd előállítása hagyományosan a gépi szövegfelolvasást jelenti (Csapó, 2022), de újabban már nemcsak írott szövegből, hanem artikulációs mozgásból, vagy akár agyi jelekből is kísérletez-

nek beszéd létrehozásával. Artikuláció-akusztikum becslésnek (Articulatory-to-Acoustic Mapping, AAM) nevezzük azokat az eljárásokat, amelyekben a beszédszintézis bemenete az emberi artikuláció valamilyen reprezentációja, és a kimenete a hangzó beszéd (Csapó, 2022). A beszédszintézis területén a 21. század első két évtizedében a rejtett Markov-modell (Hidden Markov Model, HMM) számított kurrens módszertannak. A 2010-es évek közepétől ennek a helyébe a mély neurális hálózat (Deep Neural Network, DNN) lépett. A DNN magukból a nyers adatokból „tanulja meg", hogy azok milyen absztrakcióval írhatók le: a betanítás során megtanulja, hogy a bemeneti adatokhoz (a beszédszintézis esetén ez lehet például az írott szöveg) milyen kimeneti adatokat (a beszédszintézisben ezek a beszédparaméterek) rendeljen (Csapó, 2022).

Az a kérdés, hogy hogyan lehetne gépi úton akusztikumot generálni az artikulációról készült vizsgálati adatok (pl. ultrahangképek) alapján, elsősorban az ún. némabeszéd-interfész (Silent Speech Interface, SSI) rendszerek előállítása kapcsán foglalkoztatják a kutatókat és a fejlesztőket (Denby et al., 2010). Az SSI megvalósításához az artikulációs szervek hangtalan mozgását kell rögzíteni valamely erre alkalmas technikával (ez lehet ultrahang, EMA vagy más alkalmas módszertan is), majd ebből a felvételből a gépi rendszer beszédet szintetizál. Egy ilyen technológia jelentősége abban áll, hogy az eszköz használójának nem kell képesnek lennie arra, hogy hangot adjon ki. A leggyakrabban említett példa a felhasználók körére azok a beszédükben akadályozott emberek, akiknek az artikulációjuk ép, de zöngeképzésre valamilyen okból (pl. a gége teljes vagy részleges eltávolítása miatt) nem képesek. Ugyanakkor szokás hivatkozni a zajos környezetben zajló kommunikációra is, valamint olyan helyzetben történő üzenetátadásra, ahol a beszélőnek csendben kell lennie, de írásban nem tud kommunikálni (néhány hétköznapi helyzet mellett akár hadászati, titkosszolgálati akciók is említhetők itt).

Csapó Tamás Gábor saját egyéni kutatásaiban és az általa vezetett kutatócsoportokban egyaránt a beszédszintézis állt a fókuszban, különösen az a kérdés foglalkoztatta, hogy a beszédben akadályozott populáció számára hogyan lehetne jól használható és könnyen hozzáférhető alkalmazást fejleszteni akár az

artikulációs felvételek (UH) felhasználásával, majd később az agyi jelek (EEG) detektálása révén. Csapó Tamás Gábor alkalmazásorientált kutatásainak a teljes körű bemutatására itt nincs módunk, néhány általunk jelentősebbnek ítélt vizsgálatot és kezdeményezést foglalunk össze röviden.

Az első magyar nyelvű ultrahangalapú SSI-kísérletben (Csapó et al., 2017b) egy beszélő közel 200 szinkronizált beszéd- és nyelvultrahang-felvételét használták fel. Az akusztikai jelből kinyerték az alapfrekvencia- és a spektrális paramétereket. Ezután mélyneurálisháló-alapú gépi tanulást alkalmaztak, melynek bemenete a nyelvultrahang volt, kimenete pedig a beszéd spektrális paraméterei. Az eredeti beszédből származó f0-paraméterekkel és a gépi tanulás által becsült spektrális paraméterekkel a szerzők mondatokat szintetizáltak, amelyekben szavak, néhol akár teljes mondatok is érthetők lettek.

Az ultrahangfelvételekből előállított akusztikai szerkezet értelemszerűen (mivel a nyelv a szupraglottális artikuláció egyik főszereplője) elsősorban a spektrális jellemzők becslésére alkalmas, ugyanakkor korábbi kísérletek szerint az alapfrekvencia becslése sem lehetetlen az ultrahanggal rögzített kép segítségével. Ennek hátterében az áll, hogy a zöngésség összefügg az artikulációs jellemzőkkel (lásd az 5. pontban írtakat). Grósz és munkatársai (2018) erre vonatkozó kísérleteikben két mély neuronhálót tanítottak párhuzamosan, egyet a zöngésség (zöngés vs. zöngétlen), egyet pedig a konkrét alapfrekvencia-érték és a spektrális paraméterek becslésére. A becsült és az eredeti alapfrekvencia-görbe között 0,74 korrelációs együtthatót kaptak, a szubjektív meghallgatási tesztek során a tesztek résztvevői az eredeti, illetve a prediktált f0-görbével szintetizált mondatokat közel egyformán természetesnek minősítették. Az f0-ra vonatkozóan további kísérleteket is folytattak, például a Wave-Glow (Prenger et al., 2019) használatával, amely az egyik legfrissebb fejlesztésű neurális vokóder, előnye, hogy relatíve egyszerű, mégis a valós idejűnél gyorsabb szintézist valósít meg. A percepciós tesztek alapján négyből három beszélő esetén a Wave-Glow segítségével a korábbiaknál természetesebb hangzású szintetizált beszédet sikerült előállítani az ultrahangfelvételek alapján készült artikuláció-akusztikum konverzióban (Csapó et al., 2020).

Egy tanulmányban az SSI kapcsán azt vizsgálták, hogy a különböző felvételek közötti variabilitás (session dependency) hogyan befolyásolja az artikuláció-akusztikum konverzió sikerességét a nyelvultrahang-alapú beszédszintézisben (Gosztolya et al., 2020). A korábbi kutatási eredmények alapján a szerzők azt várták, hogy ha olyan felvételeket használnak, amelyek között az ultrahangfejet eltávolították, majd újból visszahelyezték a beszélő álla alá, az nagymértékben rontja az előre tanított rendszer pontosságát. Azt találták, hogy a szintetizált beszéd érthetetlen volt, amennyiben a rendszert az egyik felvételen tanították, és a másikon értékelték ki (adaptálás nélkül). Még egyazon felvétellel is nagy különbségek mutatkoztak a DNN-modellek performanciáját tekintve, felvételfüggő módon. Végül az adott felvételre készült DNN-modell DNN-adaptáció esetén jobban teljesített, illetve ezen túlmenően a DNN-adaptáció gyorsabb működéshez is vezetett.

A fejlesztések az ellenkező irányban is megkezdődtek, azaz az akusztikum-artikuláció inverzió (AAI) területén. Az ilyen rendszerek például számítógéppel segített nyelvoktatásban lehetnek felhasználhatóak, ahol a nyelvultrahang-felvétel vizualizációja alapján begyakorolható az egyes hangzók ejtése. Ebben a kutatásban Csapó Tamás Gábor és munkatársai szintén mély neurális hálón alapuló gépi tanulást alkalmaztak: ezúttal a beszéd spektrális paraméterei alkották a bemenetet, a kimenet pedig a generált nyelvultrahang-videó volt (pl. Porras et al., 2019).

Vitán felül áll a jelentősége Csapó Tamás Gábor azon kezdeményezésének, amely a kommunikációs agy-gép interfész (brain-computer interface, BCI) hazai fejlesztését célozta meg. A BCI-alkalmazások természetes vagy ahhoz közeli minőségű kommunikációt tehetnek lehetővé olyan személyek számára, akik fizikai vagy neurológiai károsodás miatt nem képesek hangzó beszédre. A beszéd valós idejű szintézisének alapjául a mért idegi aktivitás szolgál. Csapó Tamás Gábor és munkatársai elektroenkefalográf (EEG) segítségével rögzített agyi aktivitásból kíséreltek meg beszédet szintetizálni (pl. Arthur & Csapó, 2022); illetőleg ugyancsak EEG-jelből ultrahangképet előállítani (Csapó et al., 2023). 2023-ban e téma köré Csapó Tamás Gábor egy workshopot rendezett Moonshine fantázia-

névvel (https://smartlab.tmit.bme.hu/moonshine-2023/), hogy szorgalmazza a nemzetközi kapcsolatokat az ekkor már kidolgozás alatt álló projekthez.

7. Összegzés

Csapó Tamás Gábor több területen katalizálta a hazai kutatásokat az által, hogy az ultrahangos képalkotó technikát megismertette a magyar kutatókkal, módszertani és alkalmazásfejlesztéseket végzett és indított el. Az említetteken túlmenően számos egyetemi hallgató és doktorandusz különböző kutatásait is motiválta és támogatta a nyelvultrahangos vizsgálatok különféle területein is, tevékenységének iskolateremtő hatása volt. A széleskörű nemzetközi ismertség és elismertség, a pályázati sikerek jelzik az általa végzett tudományos munka értékét.

Zárszó

Csapó Tamás Gábor mindig érdeklődő, minden új ismeretre nyitott és rendkívül pozitív szemléletű és személyiségű kutató volt, aki nem félt kérdezni, és mindezzel nem csak hallgatói, hanem kollégái előtt is követendő példával járt. Inspiráló volt vele együtt dolgozni. Tamás mindenben támogató volt: akár szakmai kérdések, akár konfliktusok, akár szervezési problémák merültek fel, ott volt, hogy segítsen, megoldja, együtt gondolkodjon, vagy csak jó szóval támogassa társait. Az általa szervezett Moonshine workshopon olyan légkört teremtett, amely hosszútávú nemzetközi együttműködéseket alapozott meg a BCI területén, és ezek Tamás váratlan távozása nélkül talán a legrangosabb pályázatok elnyeréséig vezethettek volna.

Köszönetnyilvánítás

A tanulmány elkészítését a Magyar Tudományos Akadémia Bolyai János Kutatási Ösztöndíja (Bolyai$_9$18$_2$3) (G. T. E.), az Innovációs és Technológiai Minisztérium ÚNKP-23-5 Új Nemzeti Kiválósági Programja a Nemzeti Kutatási, Fejlesztési és Innovációs Alap forrásából (G. T. E.), a TKA-DAAD 177375

számú ösztöndíja (D. A., J. K.), valamint az NKFIH FK128814 számú és az EKÖP-24 számú pályázata (J. K.) támogatta.

Hivatkozások

Ahn, S., & Davidson, L. (2016). Tongue root positioning in english voiced obstruents: Effects of manner and vowel context. *Journal of the Acoustical Society of America*, *140*, 3221–3221. URL: https://doi.org/10.1121/1.4970161. doi:10.1121/1.4970161.

Arthur, F., & Csapó, T. (2022). Deep learning alapú agyi jel feldolgozás és beszédszintézis előkészítő munkálatai. In G. Berend, G. Gosztolya, & V. Vincze (Eds.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022* (p. 185–198). Szeged: Szegedi Tudományegyetem Informatikai Intézet.

Ashby, P. (2011). *Understanding phonetics*. London: Hodder Education.

Beňuš, S., & Gafos, A. (2007). Articulatory characteristics of hungarian 'transparent' vowels. *Journal of Phonetics*, *35*, 271–300. URL: https://doi.org/10.1016/j.wocn.2006.11.002. doi:10.1016/j.wocn.2006.11.002.

Bolla, K. (1981a). A magyar hosszú mássalhangzók képzése. (kinoröntgenografikus vizsgálat számítógéppel. *Magyar Fonetikai Füzetek*, *7*, 7–55.

Bolla, K. (1981b). A magyar magánhangzók és rövid mássalhangzók képzési sajátságainak dinamikus kinoröntgenográfiai elemzése. *Magyar Fonetikai Füzetek*, *8*, 5–62.

Csapó, T. (2022). A nyelvmozgás ultrahangos vizsgálata és az automatikus elemzés alkalmazási lehetőségei a beszédtechnológiában. In A. Deme, & A. Kuna (Eds.), *Tanulmányok a nyelvészet alkalmazásainak területéről* (p. 197–219). Budapest: ELTE Eötvös Kiadó.

Csapó, T., Arthur, F., Nagy, P., & Boncz, A. (2023). Towards ultrasound tongue image prediction from eeg during speech production. In *Procee-*

dings of the 24th International Speech Communication Association, INTERSPEECH 2023 (p. 1164–1168). Dublin: International Speech Communication Association. URL: https://doi.org/10.21437/Interspeech.2023-40. doi:10.21437/Interspeech.2023-40 iSCA).

Csapó, T., & Csopor, D. (2015). Ultrahangos nyelvkontúrkövetés automatikusan: a mély neuronhálókon alapuló autotrace eljárás vizsgálata. *Beszédkutatás*, (p. 177–187).

Csapó, T., Deme, A., Gráczi, T., Markó, A., & Varjasi, G. (2017a). Szinkronizált beszéd- és nyelvultrahang-felvételek a sonospeech rendszerrel. In V. Vincze (Ed.), *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017* (p. 339–346). Szeged: Szegedi Tudományegyetem Informatikai Intézet. URL: http://rgai.inf.u-szeged.hu/project/mszny2017/files/kotet.pdf.

Csapó, T., Grósz, T., Tóth, L., & Markó, A. (2017b). Beszédszintézis ultrahangos artikulációs felvételekből mély neuronhálók segítségével. In V. Vincze (Ed.), *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017* (p. 181–192). Szeged: Szegedi Tudományegyetem Informatikai Intézet. URL: https://rgai.inf.u-szeged.hu/sites/rgai.sed.hu/files/kotet.pdf.

Csapó, T., & Lulich, S. (2015). Error analysis of extracted tongue contours from 2d ultrasound images. In *Proceedings of Interspeech 2015* (p. 2157–2161). Drezda: ISCA. URL: https://doi.org/10.21437/Interspeech.2015-486. doi:10.21437/Interspeech.2015-486.

Csapó, T., & Xu, K. (2020). Quantification of transducer misalignment in ultrasound tongue imaging. In *Proc. Interspeech 2020* (p. 3735–3739). Sanghaj: Interspeech. URL: https://doi.org/10.21437/Interspeech.2020-1672. doi:10.21437/Interspeech.2020-1672.

Csapó, T., Xu, K., Deme, A., Gráczi, T., & Markó, A. (2021). Transducer misalignment in ultrasound tongue imaging. In M. Tiede, D. Whalen, & V. Gracco (Eds.), *Proceedings of the 12th International Seminar on Speech Production* (p. 166–169). New Haven (CT: Haskins Press.

Csapó, T., Zainkó, C., Tóth, L., Gosztolya, G., & Markó, A. (2020). Ultrasound-based articulatory-to-acoustic mapping with waveglow speech synthesis. In H. Meng, B. Xu, & T. Zheng (Eds.), *Interspeech 2020* (p. 2727–2731). Sanghaj: International Speech Communication Association. URL: https://doi.org/10.21437/Interspeech.2020-1031. doi:10.21437/Interspeech.2020-1031 iSCA).

Deme, A. (megj). *A magánhangzók változatossága a magyarban: Eredmények a koartikulációs ellenállásról és a koartikulációs agresszióról a magánhangzók közötti koartikulációban artikulációs és akusztikai adatok alapján*. Budapest: Akadémiai Kiadó.

Deme, A., Bartók, M., Gráczi, T., Csapó, T., & Markó, A. (2019a). Articulatory organization of geminates in hungarian. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences 2019* (p. 1739–1743). Canberra: Australasian Speech Science and Technology Association Inc.

Deme, A., Bartók, M., Gráczi, T., Csapó, T., & Markó, A. (2019b). Geminátak artikulációs szerveződése a magyarban. *Beszédkutatás*, *27*, 54–74.

Deme, A., Greisbach, R., Markó, A., Meier, M., Bartók, M., Jankovics, J., & Weidl, Z. (2016). Tongue and jaw movements in high-pitched soprano singing: A case study. *Beszédkutatás*, (p. 121–138).

Deme, A., Greisbach, R., Meier, M., Bartók, M., Jankovics, J., Weidl, Z., & Markó, A. (2017). Tongue and jaw articulation of soprano singers at high pitch in hungarian and german. In J. Dang, & A. Li (Eds.), *11th International Seminar on Speech Production (ISSP 2017) Book of Abstracts* (p. 110–113). Tianjin: ISSP.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., & Brumberg, J. (2010). Silent speech interfaces. *Speech Communication*, *52*, 270–287. URL: https://doi.org/10.1016/j.specom.2009.08.002. doi:10.1016/j.specom.2009.08.002.

Epstein, M., & Stone, M. (2005). The tongue stops here: Ultrasound imaging of the palate (l. *The Journal of the Acoustical Society of America*, *118*, 2128–2131.

Fasel, I., & Berry, J. (2010). Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. In *Proceedings of ICPR* (p. 1493–1496). Istanbul: ICPR. URL: https://doi.org/10.1109/ICPR.2010.369. doi:10.1109/ICPR.2010.369.

Goffman, L. (2015). Effects of language on motor processes in development. In M. Redford (Ed.), *The handbook of speech production* (p. 555–577). Chichester: Wiley Blackwell. URL: https://doi.org/10.1002/9781118584156.ch24. doi:10.1002/9781118584156.ch24.

Gosztolya, G., Grósz, T., Tóth, L., Markó, A., & Csapó, T. (2020). Applying dnn adaptation to reduce the session dependency of ultrasound tongue imaging-based silent speech interfaces. *Acta Polytechnica Hungarica*, *17*, 109–124. URL: https://doi.org/10.12700/APH.17.7.2020.7.6. doi:10.12700/APH.17.7.2020.7.6.

Gráczi, T., Bóna, J., Csapó, T., Deme, A., Bartók, M., & Markó, A. (2018). Medicational effect on acoustic and articulatory vowel and voice parameters in young onset parkinson's disease. case study. In *Előadás. 17th International Clinical Phonetics and Linguistics Association Conference (ICPLA* (p. 23–26). St Julians, Malta.

Gráczi, T., Csapó, T., Bartók, M., Deme, A., & Markó, A. (2020a). The realization of voicing opposition in alveolar fricatives in hungarian: Preliminary study on articulation and acoustics. *Beszédtudomány / Speech Science*, *1*, 22–56.

Gráczi, T., Csapó, T., Bartók, M., Deme, A., & Markó, A. (2021a). Articulatory and acoustic differentiation of /s/ and /ʃ/ in children's speech: longitudinal case studies. In J. Bóna (Ed.), *Dis)fluencies in children's speech*. Budapest:

Akadémiai Kiadó. URL: https://mersz.hu/dokumentum/m873dfics__35/. doi:10.1556/9789634547099.4.

Gráczi, T., Csapó, T., Deme, A., Juhász, K., & Markó, A. (2020b). A réshangok zöngésségével összefüggő nyelvpozíciós jellemzők a megelőző magánhangzóban. *Nyelvtudományi Közlemények*, *116*, 155–190.

Gráczi, T., Csapó, T., Deme, A., Juhász, K., & Markó, A. (2021b). Tongue root position in vc sequences with regard to the phonetic realization of obstruent voicing: A preliminary study on hungarian. In M. Tiede, D. Whalen, & V. Gracco (Eds.), *Proceedings of the 12th International Seminar on Speech Production* (p. 198–201). New Haven (CT: Haskins Press.

Gráczi, T., Juhász, K., Csapó, T., Deme, A., & Markó, A. (2023). Dynamic articulatory and acoustic features of hungarian sibilants as a function of phonological voicing. In R. Skarnitzl, & J. Volín (Eds.), *Proceedings of 20th International Congress of Phonetic Sciences (ICPhS* (p. 1077–1081). Praha: Guarant.

Grósz, T., Tóth, L., Gosztolya, G., Csapó, T., & Markó, A. (2018). Kísérletek az alapfrekvencia becslésére mély neuronhálós, ultrahang-alapú némabeszéd-interfészekben. In V. Vincze (Ed.), *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2018* (p. 196–2005). Szeged: Szegedi Tudományegyetem Informatikai Intézet.

Kawahara, S. (2015). The phonetics of obstruent geminates, sokuon. In H. Kubozono (Ed.), *The Mouton Handbook of Japanese Language and Linguistics*. Berlin: Mouton Gruyter. URL: https://doi.org/10.1515/9781614511984.43. doi:10.1515/9781614511984.43.

Kelsey, C., Minifie, F., & Hixon, T. (1969). Applications of ultrasound in speech research. *Journal of Speech and Hearing Research*, *12*, 564–575. URL: https://doi.org/10.1044/jshr.1203.564. doi:10.1044/jshr.1203.564.

Kochetov, A. (2020). Research methods in articulatory phonetics i: Introduction and studying oral gestures. *Language and Linguistics Compass*, *14*, 1–29. URL: https://doi.org/10.1111/lnc3.12368. doi:10.1111/lnc3.12368.

Kochetov, A., & Kang, Y. (2017). Supralaryngeal implementation of length and laryngeal contrasts in japanese and korean. *Canadian Journal of Linguistics*, *62*, 18–55. URL: https://doi.org/10.1017/cnj.2016.39. doi:10.1017/cnj.2016.39.

Li, M., Kambhamettu, C., & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics*, *19*, 545–554. URL: https://doi.org/10.1080/02699200500113616. doi:10.1080/02699200500113616.

Lotz, J. (1966). Egy magyar röntgen-hangosfilm és néhány fonológiai kérdés. *Magyar Nyelv*, *62*, 257–266.

Markó, A., Bartók, M., Csapó, T., Gráczi, T., & Deme, A. (2019a). Az /iː/ artikulációs és akusztikai sajátosságai harmonikusan és antiharmonikusan toldalékolódó tövekben. *Nyelvtudományi Közlemények*, *115*, 233–254.

Markó, A., Csapó, T., Bartók, M., Gráczi, T., & Deme, A. (2019b). Patterns of lingual cv coarticulation in hungarian children's speech: The case of stops. In R. Rose, & R. Eklund (Eds.), *Proceedings of DiSS 2019. The 9th Workshop on Disfluency in Spontaneous Speech ELTE Eötvös Loránd University Budapest, Hungary 12–13 September, 2019* (p. 93–94). Budapest: ELTE Faculty of Humanities.

Markó, A., Csapó, T., Deme, A., Bartók, M., & Gráczi, T. (2020). A magánhangzók lingvális képzési jegyeinek tanítása az artikulációs tapasztalat, a kutatási eredmények és az elméleti kategóriák tükrében. In I. Zs. Ludányi, & A. Domonkosi (Eds.), *A nyelv perspektívája az oktatásban. Válogatás a PeLiKon2018 oktatásnyelvészeti konferencia előadásaiból* (p. 71–82). Eger: EKE Líceum Kiadó. URL: https://doi.org/10.17048/Pelikon2018.2020.71. doi:10.17048/Pelikon2018.2020.71.

Markó, A., Csapó, T., Deme, A., Gráczi, T., & Bartók, M. (2019c). Gyermekek lingvális artikulációjának variabilitása magánhangzós nyelvkontúrok alapján. In J. Bóna, & V. Horváth (Eds.), *Az anyanyelvelsajátítás folyamata hároméves kor után* (p. 165–190). Budapest: ELTE Eötvös Kiadó.

Markó, A., Csapó, T., Deme, A., Gráczi, T., & G, V. (2017). A gyermeki artikuláció vizsgálata – Új lehetőségek a hazai kutatásban. In J. Bóna (Ed.), *Új utak a gyermeknyelvi kutatásokban* (p. 65–95). Budapest: ELTE Eötvös Kiadó.

Mády, K. (2008). Magyar magánhangzók vizsgálata elektromágneses artikulográffal normál és gyors beszédben. *Beszédkutatás*, (p. 52–66).

Narayanan, S., Alwan, A., & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *98*, 1325–1347. URL: `https://doi.org/10.1121/1.413469`. doi:`10.1121/1.413469`.

Neuberger, T. (2015). Durational correlates of singleton-geminate contrast in hungarian voiceless stops. In *Proc. 18th ICPhS*. paper: ICPHS0422.

Nittrouer, S., Studdert-Kennedy, M., & Neely, S. (1996). How children learn to organize their speech gestures: Further evidence from fricative-vowel syllables. *Journal of Speech and Hearing Research*, *39*, 379–389. doi:`10.1044/jshr.3902.379`.

Olaszy, G. (2006). Hangidőtartamok és időszerkezeti elemek a magyar beszédben. *Nyelvtudományi Értekezések*, *155*.

Payne, E. (2006). Non-durational indices in italian geminate consonants. *Journal of the International Phonetic Association*, *3*, 83–95. URL: `https://doi.org/10.1017/S0025100306002398`. doi:`10.1017/S0025100306002398`.

Percival, M., Csapó, T., Bartók, M., Deme, A., E., T., & Markó, A. (2020a). Tongue root and voicing in hungarian singleton and geminate

obstruents. In *12th International Seminar on Speech Production*. URL: https://issp2020.yale.edu/S08/percival_08_15_156_abstract.pdf.

Percival, M., Csapó, T., Markó, A., Deme, A., Gráczi, T., & Bartók, M. (2020b). Gemination as fortition? articulatory data from hungarian. In *Előadás. LabPhon17 virtual conference* (p. 6–8). Vancouver. URL: https://labphon.org/sites/default/files/previous_conferences/LP17/abstracts/LabPhon_17_paper_308.pdf.

Percival, M., G., C. T., Bartók, M., Deme, A., Gráczi, T., & Markó, A. (2020c). Ultrasound imaging of hungarian geminates. előadás. ultrafest ix. *Ultrasound Imaging for Speech and Language. Virtual Conference*, (p. 21–24). URL: https://ultrafest2020.indiana.edu/abstracts/UltraFest_IX__Percival_Csapo_Bartok_Deme_Graczi_Marko_Hungarian.pdf.

Percival, M., Kochetov, A., & Kang, Y. (2018). An ultrasound study of gemination in coronal stops in eastern oromo. *Proceedings of Interspeech*, (p. 1531–1535). URL: https://doi.org/10.21437/Interspeech.2018-2512. doi:10.21437/Interspeech.2018-2512.

Porras, D., Sepulveda-Sepulveda, A., & Csapó, T. (2019). Dnn-based acoustic-to-articulatory inversion using ultrasound tongue imaging. In *Proceedings of the International Joint Conference on Neural Networks* (p. 1–8). New Jersey: IEEE. URL: https://doi.org/10.1109/IJCNN.2019.8851769. doi:10.1109/IJCNN.2019.8851769.

Prenger, R., Valle, R., & Catanzaro, B. (2019). Waveglow: A flow-based generative network for speech synthesis. In *Proc ICASSP* (p. 3617–3621). Brighton: ICASSP. URL: https://doi.org/10.1109/ICASSP.2019.8683143. doi:10.1109/ICASSP.2019.8683143.

Pycha, A. (2009). Lengthened affricates as a test case for the phonetics–phonology interface. *Journal of the International Phonetic Association*, *39*, 1–31.

Pycha, A. (2010). A test case for the phonetics–phonology interface: Gemination restrictions in hungarian. *Phonology*, *27*, 119–152.

Recasens, D., & Rodríguez, C. (2016). A study on coarticulatory resistance and aggressiveness for front lingual consonants and vowels using ultrasound. *Journal of Phonetics*, *59*, 58–75. URL: https://doi.org/10.1016/j.wocn.2016.09.002. doi:10.1016/j.wocn.2016.09.002.

Ridouane, R. (2007). Gemination in tashlhiyt berber: an acoustic and articulatory study. *Journal of the International Phonetic Association*, *37*, 119–142.

Scobbie, J., Lawson, E., Cowen, S., Cleland, J., & Wrench, A. (2011). A common co-ordinate system for mid-sagittal articulatory measurement. In *Working Paper WP-20*. Edinburgh: Queen Margaret University. doi:https://core.ac.uk/reader/161924590.

Seikel, J., King, D., & Drumright, D. (2010). *Anatomy & physiology for speech, language and hearing*. (4th ed.). Delmar: Cengage Learning.

Siptár, P., & Gráczi, T. (2014). Degemination in hungarian: Phonology or phonetics? *Acta Linguistica Hungarica*, *61*, 443–471. URL: https://doi.org/10.1556/aling.61.2014.4.4. doi:10.1556/aling.61.2014.4.4.

Siptár, P., & Törkenczy, M. (2000/2007). *The Phonology of Hungarian*. Oxford: Oxford University Press.

Stevens, K. (1998). *Acoustic Phonetics*. Massachusetts: MIT Press. URL: https://doi.org/10.7551/mitpress/1072.001.0001. doi:10.7551/mitpress/1072.001.0001.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, *19*, 455–501. URL: https://doi.org/10.1080/02699200500113558. doi:10.1080/02699200500113558.

Stone, M. (2010). Laboratory techniques for investigating speech articulation. In W. Hardcastle, J. Laver, & F. Gibbon (Eds.), *The Handbook*

of Phonetic Sciences (p. 9–38). Oxford: Willey-Blackwell. URL: https://doi.org/10.1002/9781444317251.ch1. doi:10.1002/9781444317251.ch1.

Szende, T. (1974). A magyar hangrendszer néhány összefüggése röntgenográfiai vizsgálatok tükrében. *Magyar Nyelv*, *70*, 68–77.

Tang, L., Bressmann, T., & Hamarneh, G. (2012). Tongue contour tracking in dynamic ultrasound via higher-order mrfs and efficient fusion moves. *Medical Image Analysis*, *16*, 1503–1520. URL: https://doi.org/10.1016/j.media.2012.07.001. doi:10.1016/j.media.2012.07.001.

Terband, H., Maassen, B., Lieshout, P., & Nijlan, D. (2011). Stability and composition of functional synergies for speech movements in children with developmental speech disorders. *Journal of Communication Disorders*, *44*, 59–74. URL: https://doi.org/10.1016/j.jcomdis.2010.07.003. doi:10.1016/j.jcomdis.2010.07.003.

Trencsényi, R. (2020a). Mri- és uh-felvételek geometriai elemzése a beszédszintézisben. *Acta Medicinae et Sociologica*, *11*, 55–65.

Trencsényi, R. (2020b). A nyelvkontúrkövető algoritmusok és a gépi tanulás összekapcsolhatóságának vizsgálata. In G. Berend, G. Gosztolya, & V. Vincze (Eds.), *XVI. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2020* (p. 23–24 233–244). Szeged: Szegedi Tudományegyetem Informatikai Intézet.

Trencsényi, R., & Czap, L. (2021). Possible methods for combining tongue contours of dynamic mri and ultrasound records. *Acta Polytechnica Hungarica*, *18*, 143–160.

Vorperian, H., & Kent, R. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*, *50*, 1510–1545. URL: https://doi.org/10.1044/1092-4388(2007/104). doi:10.1044/1092-4388(2007/104).

Vorperian, H., Wang, S., Chung, M., M., S. E., B., D. R., D., K. R., Ziegert, A., & Gentry, L. (2009). Anatomic development of the oral and pharyngeal

portions of the vocal tract: An imaging study. *The Journal of the Acoustical Society of America*, *125*, 1666–1678. URL: https://doi.org/10.1121/1.3075589. doi:10.1121/1.3075589.

Whalen, D., Kang, J., Iwasaki, R., Shejaeya, G., Kim, B., Roon, K., & Tiede, M. (2019). Accuracy assessments of hand and automatic measurements of ultrasound images of the tongue. In S. Calhoun, P. M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (p. 542–546). Canberra: Australasian Speech Science and Technology Association Inc.

Zharkova, N. (2016). Ultrasound and acoustic analysis of sibilant fricatives in preadolescents and adults. *The Journal of the Acoustical Society of America*, *139*, 2342–2351. URL: https://doi.org/10.1121/1.4947046. doi:10.1121/1.4947046.

Zharkova, N., Gibbon, F., & Hardcastle, W. (2015). Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. *Clinical Linguistics and Phonetics*, *29*, 249–265. URL: https://doi.org/10.3109/02699206.2015.1007528. doi:10.3109/02699206.2015.1007528.

Zharkova, N., Hardcastle, W., & Gibbon, F. (2018). The dynamics of voiceless sibilant fricative production in children between 7 and 13 years old: An ultrasound and acoustic study. *The Journal of the Acoustical Society of America*, *144*, 1454–1466. URL: https://doi.org/10.1121/1.5053585. doi:10.1121/1.5053585.

Zharkova, N., & Hewlett, N. (2009). Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using english /t/ and /ɑ/. *Journal of Phonetics*, *37*, 248–256. URL: https://doi.org/10.1016/j.wocn.2008.10.005. doi:10.1016/j.wocn.2008.10.005.

Zharkova, N., Hewlett, N., & Hardcastle, W. (2011). Coarticulation as an indicator of speech motor control development in children: An ultrasound study.

Motor Control, *15*, 118–140. URL: https://doi.org/10.1123/mcj.15.1.118. doi:10.1123/mcj.15.1.118.

Zharkova, N., Hewlett, N., & Hardcastle, W. (2012). An ultrasound study of lingual coarticulation in /sv/ syllables produced by adults and typically developing children. *Journal of the International Phonetic Association*, *42*, 193–208. doi:10.1017/S0025100312000060.

Interaction of phonetic materials and simulated probe position on the mean square error metric of tongue ultrasound probe alignment – a methodological case study

Pertti Palo[1], Steven M. Lulich[2], Daniel Aalto[1]

[1] *University of Alberta*
[2] *Indiana University*

Abstract

In tongue ultrasound imaging shifts in probe placement can cause problems in data interpretation if they go undetected. We analyse a promising metric – the mean squared error (MSE) of the mean ultrasound images – as a metric of probe stability. The metric's performance is evaluated against systematically varied speech materials (fronted articulation versus backed articulation) and probe displacement. The speech materials consist of 54 different /$C_1VC_2VC_3V$/ utterances in random order produced by one native speaker of Finnish and recorded with a Micro ultrasound setup using Articulate Assistant Advanced. In the fronted condition the vowel is /i/ and consonants are varied systematically among /n,s,t/. In the backed condition the vowel is /o/ and the consonants are varied among /h,k,ŋ/. The probe displacement is both simulated and produced intentionally in the real world. For the latter the 54 utterances were repeated in a second block in a different random order. The differences between the results of the two displacement methods indicate that this dual approach merits further study. The results also indicate that varying speech materials may overshadow probe displacement which leads to a tentative recommendation of comparing like with like in speech materials when using MSE to detect probe movement.

Keywords: Ultrasound, simulated transducer rotation, transducer misalignment, data quality control.

1. Introduction

When recording any kind of speech data, acquiring data that is comparable with recordings from other speakers let alone the same speaker, is of utmost importance. Otherwise, it is difficult to make reliable inferences based on the data. To state this

more formally, we have two desirable qualities: intra-speaker comparability and inter-speaker comparability. In recording articulatory data, the first quality is satisfied by recording the same anatomical area throughout the session (or keeping the sensors or pellets in place for point tracking methods). Satisfying the second quality involves tackling the problem of anatomical normalisation. This can be much more difficult as in general it requires a method of anatomical normalisation.

When using Ultrasound Tongue Imaging (UTI) to acquire articulatory speech data from a given speaker, the intra-speaker comparability condition stipulates that we want to capture the same region of the tongue for all recordings. This means that we need some method of first setting the system up to record the correct area and of consequently assuring that this remains true across the recording session. Satisfying the second quality on a general level is not easy with tongue ultrasound as it does not readily provide data on dimensions of the vocal apparatus beyond the tongue. However, many methods circumvent this problem by quantifying the geometrical qualities of articulatory positions instead (Ménard et al., 2012; Stolar & Gick, 2013; Zharkova et al., 2015; Dawson et al., 2016). However, rotating out of the mid-sagittal plane can cause problems with even these methods as the central groove of the tongue may appear as an extra feature in the images that could lead to false conclusions about the actual articulatory position. This makes it important to be able to minimise or track probe movement during recording and to check for it in already recorded data.

In satisfying the intra-speaker comparability condition, there have been various efforts to either reduce variability in collecting tongue ultrasound data or to reduce variability by postprocessing. The first kind have mainly taken the form of stabilising the ultrasound probe in different ways, while the second kind have consisted of tracking the probe's position in relation to the speakers head, and using a biteplate and/or anatomical measurements to orient the images and resulting extracted tongue surfaces.

The most basic approach to probe stabilisation is for either an experimenter or the participant to hold it in their hand. This has the obvious drawbacks that the inherent unsteady nature of this method potentially causes frequent significant probe misalignment, and that documenting the imaging position is challenging. It is, however, often the only method that will work with very young children as they may not be able to tolerate wearing a relatively heavy headset for the duration of the experiment (Zharkova et al.,

2017).

It can also be beneficial in a clinical treatment setting because it gives more flexibility in what is actually imaged (Adler et al., 2007). The reliability of data from hand stabilisation can be improved with stability measures like laser pointers attached to the probe and the speaker's head (Gick, 2002). And if employing analysis methods that do not rely on an unmoving frame of reference such as those from for example Dawson et al. (2016) and others mentioned above the data will be readily analysable as long as the imaging plane of the probe has been relatively stable.

The probe can also be stabilised by attaching it to something very immobile and this can be combined either with asking the speaker to hold still with their chin resting on the probe or by immobilising the speaker's head as well (Stone & Davis, 1995). While this solution is simple, it is hardly portable. These days, stabilising probe position is perhaps most usually done with helmets and headsets which can be either rigid (Articulate Instruments Ltd, 2008; Spreafico et al., 2018) or elastic (Derrick et al., 2018). These have the advantage of being fairly portable, but still allow the speaker's head to move in relation to the probe.

Another way of tackling the problem is to relate the image data to anatomic markers. This can take the form of locating anatomical features in the images after recording and can involve using special calibration recordings to capture them. The former utilises usually the tendon of the genioglossus, but can also include the palate (Stone, 2005; Wrench & Balch-Tomes, 2022; Aalto et al., 2024). Perhaps the most used calibration recording is obtaining a tongue depressor or biteplate trace (Stone & Davis, 1995). These give a common reference line – the occlusal plane of the speaker – that can be used to rotate the images to the same orientation across sessions and speakers. This comes with the obvious caveat that if the participant has malocclusion, the occlusal plane is not well-defined – which might be a potential problem even when there is no malocclusion because the teeth do **not** generally form a perfect plane. Furthermore, a biteplate does not guarantee that probe positioning stays the same in a given session nor that the probe is correctly positioned in neither translational nor rotational sense.

To guarantee or at least record the probe position in relation to the head, we need to employ some form of tracking. This can be done with optical point tracking methods (Whalen et al., 2005), tracking aids attached to the speakers head and the ultrasound

probe combined with image processing (Mielke et al., 2005), accelerometers (Hueber, 2013), or electromagnetic articulography (Kirkham et al., 2023). All of these methods require extra equipment and potentially heavy post-processing to detect head/probe movement and/or to make corrections.

An alternative light-weight method was first developed by Csapó & Xu (2020); Csapó et al. (2020). It involves using Mean Squared Error (MSE) of mean ultrasound images as a metric to identify potential probe movement. The method is computationally relatively light and does not require any extra equipment and does not require any special setup in the data recording. On top of this the MSE analysis results are easy to interpret as they provide a holistic view to the data from a given recording session in one glance.

The lack of special requirements on the data makes the MSE method very useful as it can be applied to any existing tongue ultrasound data. Given the ease of interpreting the results, it can also be used for quick assessment of data quality right after recording, allowing for prompt re-recording in case misalignment is detected.

In this study we provide a first step towards evaluating the MSE method's performance by using controlled data as well as simulated data. In the data we varied the speech materials to assess their effect on the metric, and we also intentionally changed the probe position during the recording session. Furthermore, we simulated probe rotation via a bootstrap method.

2. Materials and Methods

2.1. Audio and ultrasound recordings

Synchronised audio and ultrasound were recorded with Articulate Assistant Advanced (Articulate Instruments Ltd, 2024) using a Micro ultrasound system from Articulate Instruments Ltd. Audio was sampled at 22.050 kHz and ultrasound was acquired at 63.5 fps with Field of View of $\approx 102.7°$, 64 scanlines and depth of 120 mm / 898 pixels.

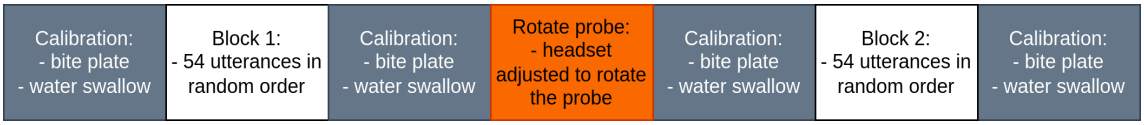


Figure 1: Block design of the experiment.

Figure 1 illustrates the over all design of the experiment. The speech materials consisted of /$C_1VC_2VC_3V$/ utterances produced by one male native speaker of Finnish.

The vowels and consonants were chosen to elicit fronted and backed articulations. Overall we chose to keep the vowel constant within each condition and vary the consonant by choosing a nasal, fricative and a plosive that have target constrictions corresponding to the condition. In the fronted condition the vowel was the fronted high corner vowel /i/ and consonants (C_1, C_2, andC_3) were varied among /n,s,t/ which all have frontal articulatory targets. In the backed condition the vowel was the back round vowel /o/ and the consonants were varied among /h,k,ŋ/. While the latter two have velar targets, in Finnish, /h/ has allophonic variation and can be produced as a pharyngeal fricative (Suomi et al., 2008). The participant was instructed to aim for the pharyngeal allophone of /h/. In each condition, the vowel remained constant, and the consonant was systematically varied to produce all combinations (including e.g., in the fronted condition /ninini/ and /tinisi/ among others). Combining the vowels and consonants resulted in $2 * 3 * 3 * 3 = 54$ unique /C_1VC_2VC_3V/ utterances.

The utterances were repeated in two independently randomised blocks, each consisting of the full set of 54 utterances. Between the blocks the headset was adjusted with the aim to rotate the probe by approximately 5-10°. The first block was used as the basis for the simulated rotation described below. Both blocks contained some speech errors including false starts, disfluencies, and repeated productions of the target utterance. These were not removed from the data but rather included to see how they affected the results. Calibration tasks were recorded before and after each block. These included a bite plate recording with a wooden tongue suppressor and a water swallow.

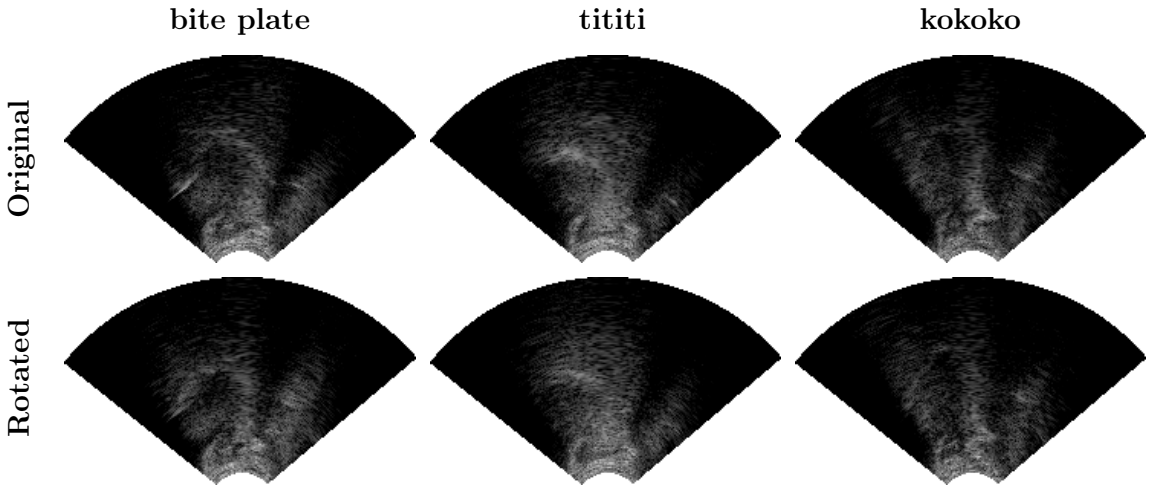


Figure 2: Representative tongue profiles.

Figure 2 shows tongue profiles from 6 different conditions. The images in the first row are from the first block and images in the second row are from the second block. In all of the images in this study the tongue tip is to the left and pharynx is on the right. The first column shows frames captured with the tongue depressor, while the tongue depressor does not appear to change position between these two images, the hyoid bone shadow does do so. See discussion for a further comment on this. The other two columns give examples of fronted – here a /tititi/ utterance – and backed – here a /kokoko/ utterance – productions. These frames are from the middle vowel of each utterance.

2.2. Probe alignment measurement with MSE

The probe alignment measurement with Mean Squared Error (MSE) is based on raw ultrasound frames. Raw frames are the uninterpolated or probe return frames illustrated in Figure 3. In our data each raw ultrasound frame has 64×898 pixels.

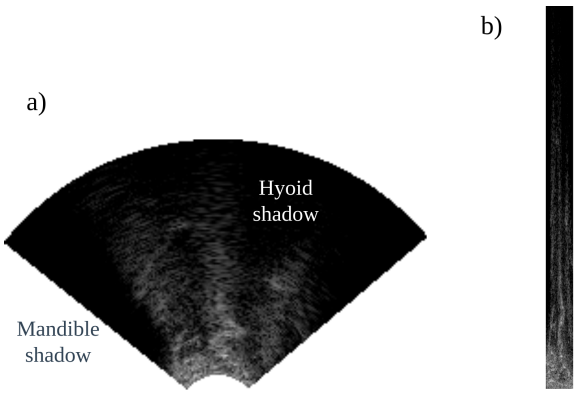


Figure 3: a) Interpolated (human-readable) ultrasound frame with mandible and hyoid shadows indicated, and b) the corresponding raw frame (probe return data).

In the following we will treat each raw ultrasound frame in a recording as a vector of l pixels indexed with k, use j to index the m consecutive images in a recording, and i to index the n consecutive recordings. So, the k^{th} individual pixel in the j^{th} frame of the i^{th} recording is $im(i, j, k)$. To calculate the MSE we will first calculate the average frame $\overline{im}$ of each recording by averaging each individual pixel over the recording:

$$\overline{im(i,k)} = \frac{1}{m}\sum_{j=1}^{m} im(i,j,k) \tag{1}$$

Figure 4 shows the mean images for the same recordings that were sampled in Figure 2. Interestingly, in Figure 4 rather than rotated, the tongue contour seems to be slightly lower in all of the images for the second block than for the first block.

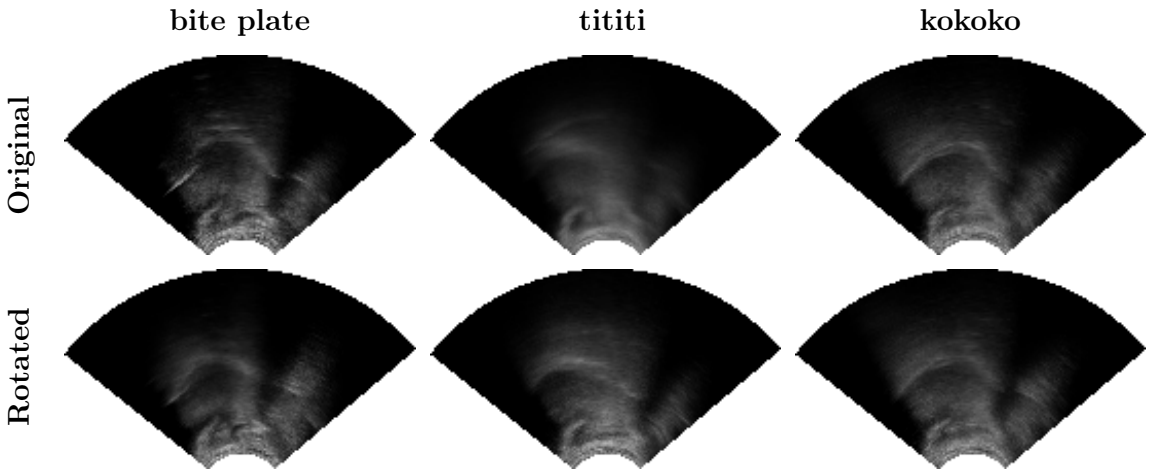


Figure 4: Examples of mean tongue profiles.

After this we calculate the MSE between each pair of average images (s, t) as follows:

$$MSE(s,t) = \frac{1}{l}\sum_{k=1}^{l}\left(\overline{im(s,k)} - \overline{im(t,k)}\right)^2. \quad (2)$$

The results of the MSE calculation can be represented as a $n \times n$ distance matrix where each individual element gives a measure of likeness between the recordings corresponding to the row and the column of the element.

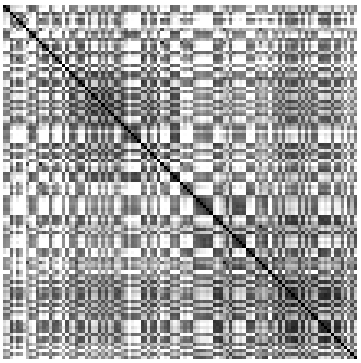


Figure 5: MSE matrix calculated on our whole (unsorted) data set. The dark line on the diagonal is the result of comparing each recording also to itself.

Figure 5 illustrates a distance matrix produced with MSE on the whole (unsorted) data after removing the calibration recordings (biteplate and water swallow). In it both rows and columns correspond to the individual recordings and each pixel is produced by calculating the MSE between the mean images of the recordings corresponding to the pixel's row and column. The closer to white a pixel is the more different the recordings are while darker pixels mark more similar recordings. The recordings appear in order of recording. The pixel in row 1, column 1 is comparison of the first recording to itself and

the dark pixels on the diagonal are comparisons of recordings to themselves. Importantly the left top quarter contains within-block comparisons of recordings in block one and right bottom quarter is comparisons within block two. Bottom left, and top right quarters are mirrored as they both are cross-block comparisons between recordings of blocks one and two.

2.3. Simulating probe rotation

We also simulated rotating the probe under a speaker's chin by producing artificially rotated data from the first (unadjusted) block of recordings. First we selected a subsector of each mean image to use in the analysis. This is done by excluding scanlines from either the front or the back of the image producing a continuous sector which is part of the original data as shown in Figure 6. We then compare two differently selected sectors over all the baseline recordings in our data. This simulates the probe having been turned by the amount that the sectors differ within the data.

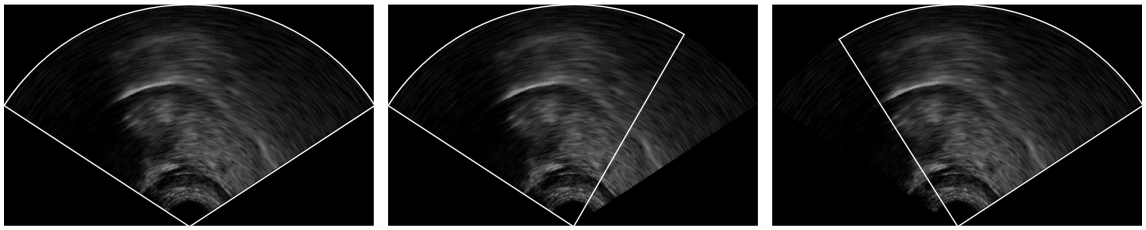


Figure 6: Simulating probe rotation by selecting sectors from the data. Panel on the left illustrates the original scanned sector (tongue tip on the left), middle panel shows a selection on the extreme left and panel on the right on as far right as possible. Intermediate selections can be made easily by moving the selection by one scanline at a time.

2.4. Code availability

The analysis was implemented in Python as part of the PATKIT software package (Palo et al., 2025) and stimulus lists were constructed in R (R Core Team, 2013) with the Randomise AAA Stimulus List scripts (Palo, 2024).

3. Results

Figure 5 shows the MSE matrix for the whole data without any sorting with the calibration recordings (biteplate and water swallow) removed. While there does appear to be some structure in it, it is very difficult to see a difference between blocks 1 and 2 in that image.

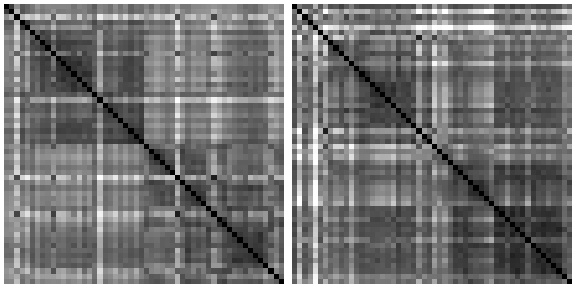


Figure 7: MSE matrix calculated on the fronted speech samples (left) and backed speech samples (right).

3.1. Varying speech materials and rotating the probe physically

Running the analysis with only the fronted or backed utterances, the situation becomes much clearer. In Figure 7 the effect of probe rotation can be seen fairly clearly in the fronted samples and less clearly in the backed samples. It appears as the two darker quadrants bisected by the diagonal and two lighter quadrants off the diagonal. This means that within block/rotation condition the recordings appear to be more similar.

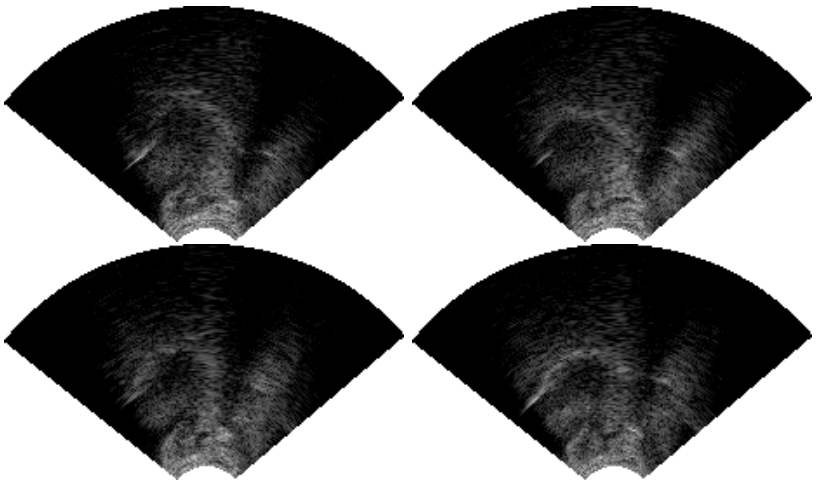


Figure 8: Bite plate frames. Top row left: block one begin, top right: block one end, bottom left block: two begin, and bottom right: block two end.

Interestingly, while the probe attachment was turned approximately 5-10° between the blocks, the biteplate traces are very much in the same direction, as seen in Figure 8. Instead, it seems that the angle of the hyoid shadow changes, but that is unfortunately more difficult to measure. And as mentioned already above in connection to Figure 4, instead of rotating it seems that the tongue contour is lowering from Block 1 to Block 2.

3.2. Simulation

The simulated data was produced from the first (unadjusted) block of original recordings. To simulate two probe positions the data was sampled as explained above in

Section 2.3. Table 1 lists the step lengths and corresponding rotation angles in radians and degrees used. For each step length condition we get two probe positions – position 1 and 2 in the following figures – rotated by the angle in Table 1. Figures 9-11 display the results from simulating probe rotation. In the Figures the recordings have been sorted within each simulated block (positions 1 and 2). Each position shows first the frontal – containing /i/ and then the backed articulations – containing /o/, and are sorted alphabetically within those sub-blocks. The squares under each step length heading contain cross-comparisons of each recording within that step length condition. No comparisons were made between non-matching step-lengths as this would mean comparing images of different size which is not a defined operation for the MSE metric.

In Figure 9 the first two vowel rows correspond to the first simulated rotation position – simulated Block 1 – and the other two to the second simulated position – simulated Block 2. The columns repeat this same pattern nesting within each step length the simulated positions (i.e., rotations), and within the positions the front and back articulatory conditions marked here by vowel qualities /i/ and /o/, respectively.

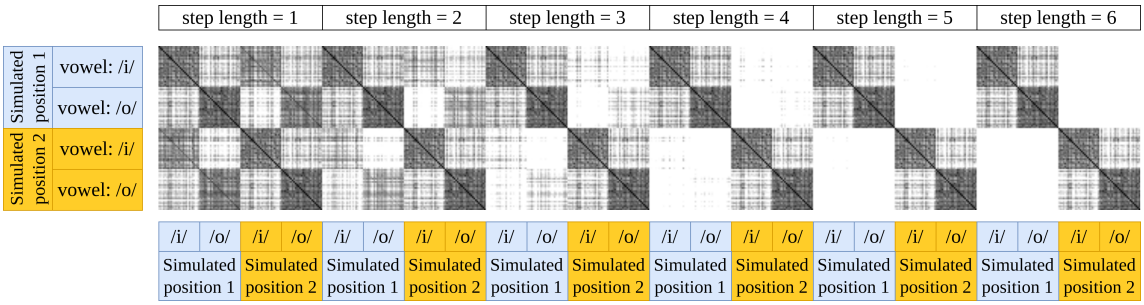


Figure 9: Effects of simulated probe rotation.

Looking at the columns where rotation step length = 1, we can see a black line on the diagonal. These are formed by pixels which correspond to recordings being compared to themselves resulting in MSE = 0. After that the darkest part of the figure are the areas where the articulation condition (/i/ or /o/) matches in the same rotation. These are the dark squares that surround the black pixels on the diagonal. Next darkest are comparisons of matching articulation types but differing rotation – for example third

Table 1: Rotation angles used in simulating probe rotation.

Step length	1	2	3	4	5	6
Radians	0.028	0.056	0.084	0.112	0.14	0.168
Degrees	1.60	3.21	4.81	6.42	8.02	9.63

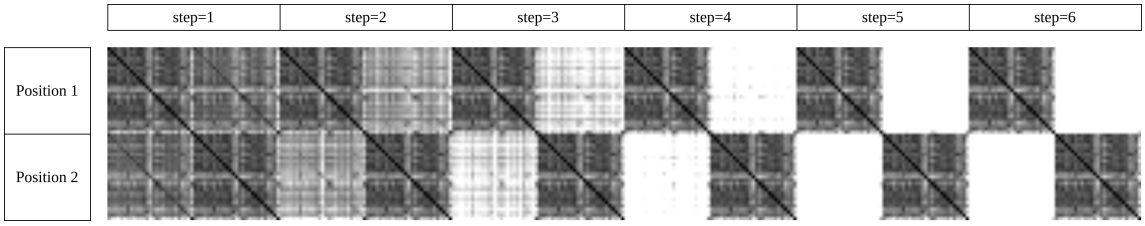


Figure 10: Simulated probe rotation for only the fronted utterances, vowel /i/ in Figure 9.

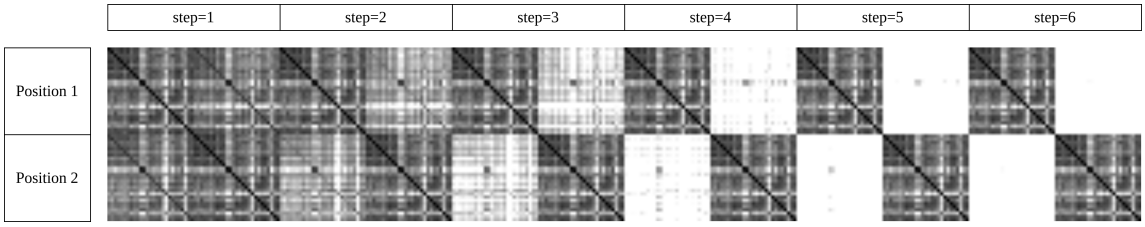


Figure 11: Simulated probe rotation for only the backed utterances, vowel /o/ in Figure 9.

quarter 'row' (Simulated position 2, vowel /i/) counting from the top in the first quarter 'column' (Simulated position 2, vowel /i/) – and only after that we get the same rotation but differing articulation type. In step = 2 the order is less clear and from step = 3 onwards the darkest squares are found in matching rotation with steps 5 and 6 having effectively totally white squares for mismatching rotation.

This gradient effect of increasing simulated rotation can be more easily observed in Figures 10 and 11. In these figures the MSE results are plotted only for fronted and backed utterances respectively. As contrast changes with subsetting the data, the effect of individual recordings is much clearer compared with Figure 9.

In Figure 10 the most prominent light lines are utterances /sisini/ and /sisisi/ in the middle of the block, and /tititi/ at the end of the block. It is unclear what sets the other two apart but /tititi/ has large non-speech movement after the acoustic utterance.

In Figure 11 the prominent utterances are /kohoho/ and /kohoko/ (in the middle of the block) which both have a similar pattern of non-speech movement to /tititi/, and (in the bottom half of the block) /ŋohoho/ – which does not have anything immediately different about the articulation, and /ŋoŋoho/ – which has a repeated utterance. It should be noted that all of the repeated utterances, false starts or too early starts were in the backed utterances in the first block, but only /ŋoŋoho/ shows any prominence in Figure 11.

4. Discussion

Both Figure 5 and the simulated rotation results show that probe alignment changes may be totally overshadowed by variation in speech materials. This means that there is reason to prefer comparing like articulations with like articulations when using MSE to evaluate probe movement. Simple means of mitigating this effect include sorting the data and plotting only part of the data at a time.

It should be noted that while /h/ was included in the speech materials in the hope of eliciting the pharyngeal fricative allophonic variant – and thus a backed tongue configuration – from the speaker, the fact that there is allophonic variation in /h/ resulted in the speaker mainly producing glottal variants. A way to avoid this problem in the future is to recruit speakers whose phonological system does not include this sort of variation, but rather specifically has a phonemic pharyngeal fricative or more generally does not include wide variation in realisations of the target sounds.

The attempt at physical rotation of the probe proved trickier than we expected. While changing probe position resulted in apparent change in tongue position, it was not the expected simple rotational change. This might be due to speaker adaptation and/or just complex mechanical interaction between the headset, probe and the speaker's anatomy. We originally intended to quantify the angle of probe adjustment based on the biteplate traces, but this proved to not be possible due to the biteplate traces not really changing.

It is easy to think that this is less of a problem in regular studies where all that matters is that the probe stays in a stable position with the desired structures in view. However, in the current context as well as more widely there are two points to consider. First, adjusting the probe holder in a given way does not necessarily have the expected result, which may affect not just how a participant makes contact with the probe but also how they articulate. Second, for purposes of simulating probe movement physical experiments involving a headset may have extra complexity in the form of speaker adaptation and physical interactions of the measurement setup and the speaker. In this sense, physically turning the probe and simulating it can be very different. This does not invalidate either approach. Rather it makes the point for using both methods for studying probe alignment evaluation methods. Physical experiments are an essential tool in ensuring that findings are grounded in reality and simulation experiments are an essential tool in checking how different manipulations affect the results in principle.

It should also be noted that the current study is limited in its use of regular 2D ultrasound. A very interesting expansion would be using 3D/4D ultrasound (Lulich et al., 2018) as the basis of the simulation. Taking 2D slices of 3D data would allow us to also examine the effects of the probe rotating position outside the sagittal plane.

Finally, we did not examine the effect of how much speech there is in a sample. Some of the bright lines visible in Figures 10 and 11, might be mitigated by limiting the analysis to just the speech. Then again depending on what we are trying to find in the data, an opposite selection – excluding the speech part from analysis – might be preferable.

5. Conclusion

We conclude that comparing like with like is the safest approach in trying to detect probe alignment issues. Furthermore, this type of evaluation method shows promise but to get closer to ground truth we need 3D/4D data as basis of the simulations and more participants.

Mean squared error as introduced for evaluating ultrasound probe alignment (Csapó & Xu, 2020), is a relatively simple but very informative tool. It is sensitive to articulatory position, probe orientation and other factors. Its power lies in combining all of these into a relative measure of similarity which can be used to flag recordings and parts of recording sessions for more fine-grained scrutiny. And as such, the method itself definitely merits further study.

Acknowledgements

This study was inspired by the work of our late colleague and friend Dr. Tamás Gábor Csapó. Pertti Palo and Steven M. Lulich remember with warmth the conversations they had with him over the years and the collaborations we were fortunate to work on with him. He is dearly missed.

Pertti Palo and Daniel Aalto work at the University of Alberta which is located on Treaty 6 territory, a traditional gathering place for diverse Indigenous peoples including the Cree, Blackfoot, Métis, Nakota Sioux, Iroquois, Dene, Ojibway/Saulteaux/Anishinaabe, Inuit, and many others whose histories, languages, and cultures continue to influence our vibrant community.

The authors wish to thank the anonymous reviewer and the editors for constructive comments that improved the quality of the article as well as Ms Hilary Warner-Evans for

proofreading the article.

References

Aalto, E. M. A., Yoshida, M., Ménard, L., Cardoso, W., & Laporte, C. (2024). Effects of an ultrasound biofeedback session on maximal tongue movements. In *Proceedings of ISSP 2024*, (pp. 63–66)., Autrans, France.

Adler, B. M., Bernhardt, B. M., Gick, B., & Bacsfalvi, P. (2007). The Use of Ultrasound in Remediation of North American English /r/ in 2 Adolescents. *American Journal of Speech-Language Pathology*, *16*(2), 128–139.

Articulate Instruments Ltd (2008). Ultrasound Stabilisation Headset Users Manual: Revision 1.4. Manual.

Articulate Instruments Ltd (2024). Articulate Assistant Advanced User Guide: Version 221.4.2. Software.

Csapó, T. G. & Xu, K. (2020). Quantification of Transducer Misalignment in Ultrasound Tongue Imaging. In *Interspeech 2020*, (pp. 3735–3739). ISCA.

Csapó, T. G., Xu, K., Deme, A., Gráczi, T. E., & Markó, A. (2020). Transducer Misalignment in Ultrasound Tongue Imaging. In *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)*, (pp. 166–169)., Online / New Haven, CT.

Dawson, K. M., Tiede, M. K., & Whalen, D. H. (2016). Methods for quantifying tongue shape and complexity using ultrasound imaging. *Clinical linguistics & phonetics*, *30*(3-5), 328–344.

Derrick, D., Carignan, C., Chen, W.-R., Shujau, M., & Best, C. T. (2018). Three-dimensional printable ultrasound transducer stabilization system. *The Journal of the Acoustical Society of America*, *144*(5), EL392.

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, *32*(2), 113–121.

Hueber, T. (2013). Ultraspeech-tools Acquisition, processing and visualization of ultrasound speech data for phonetics and speech therapy.

Kirkham, S., Strycharczuk, P., Gorman, E., Nagamine, T., & Wrench, A. (2023). Co-registration of simultaneous high-speed ultrasound and electromagnetic articulography for speech production research. In *Proceedings of the 20th International Congress of Phonetic Sciences*, (pp. 942–946).

Lulich, S. M., Berkson, K. H., & de Jong, K. (2018). Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion. *Journal of Phonetics*, *71*, 410–424.

Ménard, L., Aubin, J., Thibeault, M., & Richard, G. (2012). Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model. *Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics (IALP)*, *64*(2), 64–72.

Mielke, J., Baker, A., Archangeli, D., & Racy, S. (2005). Palatron: A technique for aligning ultrasound images of the tongue and palate. *Coyote Papers*, *14*, 96–107.

Palo, P. (2024). Randomise AAA stimulus list [R and Matlab software package]. Available in a public software repository, accessed 31 October 2024. https://github.com/giuthas-speech-research-tools/randomise_AAA_stimulus_list/.

Palo, P., Moisik, S. R., & Faytak, M. (2025). PATKIT: Phonetic Analysis ToolKIT [Python software package]. Available in a public software repository, accessed 8 February 2025. https://github.com/giuthas/patkit.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Spreafico, L., Pucher, M., & Matosova, A. (2018). UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science. In *Proc. Interspeech 2018*, (pp. 1517–1520).

Stolar, S. & Gick, B. (2013). An Index for Quantifying Tongue Curvature. *Canadian Acoustics*, *41*(1), 11–15.

Stone, M. (2005). A Guide to Analyzing Tongue Motion from Ultrasound Images. *Clinical Linguistics and Phonetics*, *19*(6–7), 455–502.

Stone, M. & Davis, E. P. (1995). A head and transducer support system for making ultrasound images of tongue/jaw movement. *The Journal of the Acoustical Society of America*, *98*(6), 3107–3112.

Suomi, K., Toivanen, J., & Ylitalo, R. (2008). *Finnish Sound Structure – Phonetics, Phonology, Phonotactics and Prosody.* STUDIA HUMANIORA OULUENSIA. University of Oulu.

Whalen, D., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-Lehouillier, H., Vatikiotis-Bateson, E., & Hailey, D. S. (2005). The Haskins Optically Corrected Ultrasound System (HOCUS). *Journal of Speech, Language, and Hearing Research*, *48*, 543–553.

Wrench, A. & Balch-Tomes, J. (2022). Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut. *Sensors*, *22*(3), 1133.

Zharkova, N., Gibbon, F. E., & Hardcastle, W. J. (2015). Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. *Clinical Linguistics & Phonetics*, *29*(4), 249–265.

Zharkova, N., Gibbon, F. E., & Lee, A. (2017). Using ultrasound tongue imaging to identify covert contrasts in children's speech. *Clinical Linguistics & Phonetics*, *31*(1), 21–34.

Artikulációs beszédszintézis megvalósítása dinamikus ultrahangfelvételek alapján

Trencsényi Réka[1], Czap László[2]

[1] *Debreceni Egyetem, Villamosmérnöki Tanszék*
[2] *Miskolci Egyetem, Automatizálási és Infokommunikációs Intézet*

Abstract

Starting from 2D dynamic ultrasound sources recording the movement of the vocal organs and the speech signal of the speaker in a simultaneous and synchronised manner, we produce machine speech by means of artificial intelligence. As visual objects, we use tongue and palate contours fitted automatically to the anatomic boundaries of the ultrasound images, and for training, we extract geometric information from these contours, as the change of their shape fundamentally describes the movement of the vocal organs during articulation. The geometric data consist of radial distances between the tongue and palate contours and coefficients of the discrete cosine transform of the curves, respectively. Relying on this dataset, parameters connected to the acoustic content of the speech signal are trained by the network. These parameters can be interpreted in the framework of the acoustic tube model of the vocal tract, and according to this, reflection coefficients and areas of the articulation channel are to be trained. In this study, sentences are synthesised using linear predictive coding and the acoustic tube model.

1. Bevezetés

A beszédkutatás egyik legfontosabb tématerülete a beszédszintézis, ami elemi alkotóját képezheti az ember-gép kapcsolatnak. Ez esetben a gép kommunikációs szerepe abban nyilvánul meg, hogy kódoló adóvá válik, azaz beszédet produkál. Napjainkban a beszédszintézis legelterjedtebb irányzata a szövegfelolvasók készítése, melyek írott szövegeket szólaltatnak meg. A beszédszintetizátorok megalkotásának célja a természetes emberi beszéd közben kialakuló akusztikai produktum élethű utánzása. Ebben a megközelítésben a beszéd hullámformája adja a kiindulópontot, amit kétfajta megoldásban alkalmaznak gépi beszéd előállítására. Az egyik csoportba az úgynevezett forráskódolású

Email addresses: trencsenyi.reka@science.unideb.hu (Trencsényi Réka), czap@uni-miskolc.hu (Czap László)

Beszédtudomány – Speech Science 2025/1. 90–116. DOI: 10.15775/Besztud.2025.1.90-116

technikák tartoznak, melyek segítségével a beszédjelből kivonják a lényegi információkat és ezeket bemeneti adatsorozatként kezelik a szintézis során (Kaur & Singh, 2022). A másik megoldás az emberi hangot közvetlenül használja fel a beszédépítéshez olyan módon, hogy a beszédjelből különböző hosszúságú hullámforma-részleteket vágnak ki és tárolnak el, majd az így kapott elemek megfelelő kiválasztásával és összefűzésével megkonstruálják a kívánt beszédhullámot (Panda & Nayak, 2017). Ezeken túlmenően, tágabb módszertani szempontok alapján megkülönböztetünk még szabályalapú, illetve statisztikai elven működő beszéd-előállítási eljárásokat. Az előbbi esetében megfigyelések és tapasztalatok szerint felállított szabályokkal koordinálják a szintézis egyes lépéseit (Carlson & Granström, 2008), az utóbbi esetében pedig valószínűségeken alapuló belső rendszerállapotok révén jutnak el a beszédprodukcióhoz. A statisztikai elvű módszerek egyik tipikus válfaja a gépi tanulóalgoritmusok szerkesztése és alkalmazása, ami a jelenlegi tudományos kutatások egyik legaktívabban prosperáló irányzataként tartható számon (Mahum et al., 2023). A magyarországi viszonyokat tekintve, a hazai kutatók az 1970-es években kapcsolódtak be a témakör tanulmányozásába. Az első magyar szövegfelolvasó rendszer, a HungaroVox 1980-ban készült el, majd folyamatos fejlesztések nyomán sorban követték egymást a ScriptoVox (Olaszy & Gordos, 1987), a Brailab (Kiss et al., 1987), a PC TALKER (Király, 1989), a PCROBOT, a MultiVox (Olaszy, 1989), a ProfiVox (Olaszy et al., 2000), illetve a FlexVoice (Balogh et al., 2000) rendszerek, melyek szépen kirajzolják a fejlődés ívét a nagyon robotos hangzású beszédtől a teljesen emberi hangzású, jól érthető beszédig.

A szövegfelolvasó rendszerek a beszédszintézis klasszikus ágát képviselik, ahol hosszú évtizedek során rendkívül sok tapasztalat és gazdag tudásanyag halmozódott fel, amit a szakirodalom számos közleménye is igazol (Arik et al., 2017; Mullah, 2015; Shiga et al., 2020). Emellett azonban olyan területek is kezdenek egyre élénkebben előtérbe kerülni, melyek kevésbé kidolgozottak, és rengeteg nyitott probléma vár még megoldásra. Ide sorolható például az artikulációs beszédszintézis (Csapó et al., 2017; Tóth et al., 2018; Juanpere & Csapó, 2019; Csapó, 2020; Csapó et al., 2020; Arthur & Csapó, 2021; Csapó

et al., 2022; Denby et al., 2023), ami az akusztikai produktum utánzását emberi hangminták helyett a hangképzés és artikuláció gépi leképezése révén próbálja megvalósítani. Ennek egyik technológiai vonulata a robotok beszédének előállításához szükséges artikulációs elektromechanikus beszédkeltőkre irányuló kísérletezés (Ashok et al., 2022; James et al., 2021; Pang et al., 2023). Szintén a jövő tendenciáinak kedvez a gégétől a száj-, illetve orrnyílásig terjedő artikulációs csatorna, más néven vokális traktus modellezésére épülő beszédszintézis, ami főként vizuális információkra támaszkodik. Számos tanulmány hitelesen alátámasztja, hogy az emberi beszéd fiziológiai folyamatairól nyert vizuális információk nagymértékben elősegítik a beszédképzés komplex mechanizmusának megértését, és ezen keresztül a beszédszintézis módszereinek hatékony fejlesztését (Birkholz et al., 2020; Leppävuori et al., 2021; Skordilis et al., 2017). A napjainkban rendelkezésünkre álló radiológiai és monitorozó eljárások – úgymint mágneses rezonanciás képalkotás (MRI), komputertomográfia (CT), ultrahang (UH), elektropalatográfia (EPG), elektromágneses artikulográfia (EMA) vagy elektroglottográfia (EGG) – nélkülözhetetlen szerepet játszanak az akusztikai-artikulációs konverzió problémájának kezelésében. A fentebb említett képalkotó és monitorozó technikák segítségével generált morfológiai és geometriai adatok felhasználásával maradéktalanul feltérképezhetők az adott beszédjelhez tartozó artikulációs mozgások. Nem triviális feladat azonban az artikuláció akusztikummal való összekapcsolása, azaz a vokális traktus morfológiai és geometriai adataira alapozott beszédprodukció megvalósítása. Ilyen jellegű kutatási eredmények már felfedezhetők a szakirodalomban (Cao et al., 2023; Gonzalez-Lopez et al., 2020; Jin et al., 2022), de ez a terület sok szempontból még a mai napig is nyitott. Az eddig publikált tanulmányok főképpen a vokális traktus geometriai modelljének megalkotására fókuszálnak, aminek alapját az esetek többségében MRI-, UH- vagy EMA-felvételek képezik (Denby & Stone, 2024; Otani et al., 2023; Toutios & Narayanan, 2013), vagy pedig a beszédjelből származó lényegi információkkal manipulálnak (Kaburagi, 2014, 2015), a tökéletes minőségű gépi beszéd tényleges megvalósításához vezető út azonban bőven tartogat még kihívásokat. Ezenkívül az artikuláció geometriai és akusztikai jellemzői között

fennálló fizikai kapcsolatok természetét és hátterét illetően is számos nyitott kérdés vár még megválaszolásra. A problémafelvetés aktualitását mutatja, hogy az artikulációs-akusztikai kapcsolatrendszer feltárása, illetve gyakorlati leképezése alapvető fontosságú lehet például a klinikai célú beszédterápiában, a nem anyanyelvi nyelvtanulási tréningek kialakításában vagy a néma beszéd megszólaltatásához szükséges szintetizátorok konstrukciójában és fejlesztésében, ami szolgálhatja többek között a gégeeltávolításon átesett emberek rehabilitációját is.

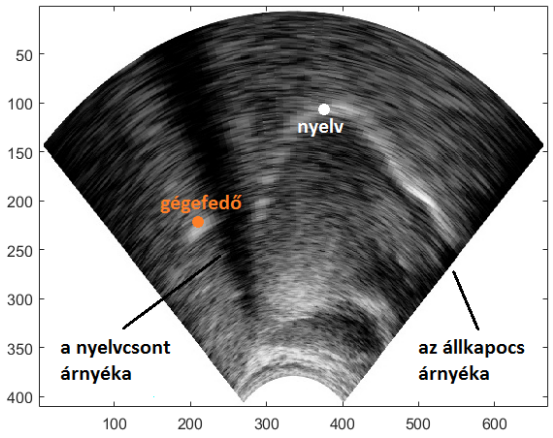
2. Az UH-felvételek jellemzése

Vizsgálataink alapvető eszközeit olyan audiovizuális források képezték, melyek ultrahangos (UH) eljárással készültek. A beszéd közben rögzített dinamikus felvételek képi formában megjelenítik a beszélő hangképző szerveit, miközben hallható a beszélő által kibocsátott akusztikus jel. A képi keretek sorozatán megfigyelhető néhány aktív hangképző szerv (nyelv, gégefedő) folytonos mozgása. A hang mint beszédjel időbeli eltolódás nélkül igazodik a felvételek képkockáihoz, így pontosan követhetők és egymáshoz rendelhetők a beszéd artikulációs és akusztikai mozzanatai. Végeredményben tehát létrejön az adott bemondáshoz tartozó, időben szinkronizált hang- és képcsomag. Az UH-felvételek kétdimenziós mozgóképek formájában álltak rendelkezésünkre. A felvételek síkját az emberi testet bal és jobb oldali részekre osztó függőleges szimmetriasík, az ún. középszagittális sík definiálja, amely lehetővé teszi a hangképző szervek kétdimenziós vetületi mozgásának, relatív elhelyezkedésének, illetve anatómiai szerkezetének részletes tanulmányozását. Az UH technikával többnyire csak a száj- és garatüreg egy része monitorozható kívülről, egy a beszélő álla alatt elhelyezett UH-fej alkalmazásával, melynek speciális elhelyezkedéséből adódóan az UH-felvételeken csak a nyelv és a gégefedő mozgása jeleníthető meg. A többi hangképző szerv ellenben nem látható, hiszen az ajkak és a gége kívül esik az eszköz letapogatási zónáján, a kemény és lágy szájpad pedig nem detektálható közvetlenül az UH-hullámok sajátos szájüregi visszaverődései miatt. Mindemel-

lett az UH-nyalábok nem képesek áthatolni a nyelvcsonton és az állcsonton, így a nyelv hátsó és elülső részeire árnyék vetül, aminek következtében a nyelv csak részlegesen mutatkozik meg. Itt jegyezzük meg, hogy a fogak nem láthatók a kereteken, mivel az UH-os eljárással kizárólag lágy szövetek tekinthetők át az emberi szervezetben.

Az 1. ábra egy statikus UH-keretet mutat be, ahol a vokális traktus látótérbe eső elemei különböző színekkel vannak felcímkézve. Látható, hogy a nyelvhát és a gégefedő csúcspontja világos sávokként rajzolódnak ki, a nyelvcsont és az állkapocs árnyékai pedig sötét zónák formájában észlelhetők.

Az általunk használt UH-csomag az MTA-ELTE Lendület Lingvális Artikuláció Kutatócsoportjának Micro rendszerével készült. Az UH-felvételeken 40 különböző mondat hangzik el egy magyar női beszélő bemondásaiban.

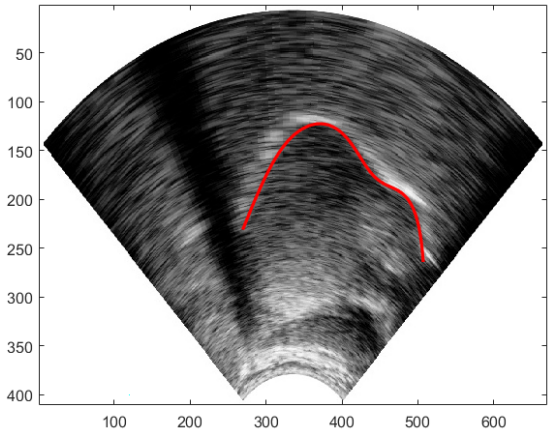


1. ábra. A vokális traktus látótérbe eső elemeinek elhelyezkedése egy statikus UH-kereten.

3. Az UH-felvételek anatómiai kontúrvonalainak megállapítása

A beszédhangok artikulációja során a nyelv alakja és pozíciója rendkívül fontos szerepet játszik, amit a nyelv felszíni formáinak tanulmányozásával lehet a legjobban leírni. Ebben nagy segítséget nyújthatnak az előző fejezetben bemutatott UH-felvételek, melyeken a nyelvszövet kétdimenziós vetületének határvonala nagyfokú pontossággal meghatározható a középszagittális síkban, ahol

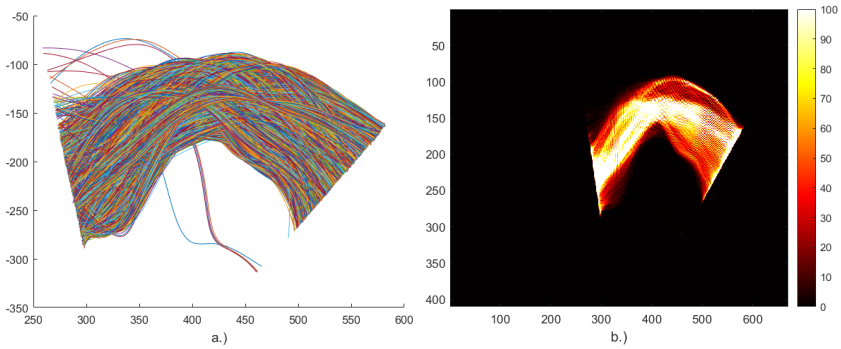
figyelemmel kísérhető a nyelv fel-le, illetve előre-hátra irányú mozgása. Az UH-keretek esetében a nyelvfelszín elmosódott, világos sávként azonosítható tartománya a nyelv és a felette lévő levegő határán kialakuló UH-hullámvisszaverődés eredményeként jön létre, így a nyelv felszíni vonalát a világos sáv alsó pereme jelöli ki (1. ábra). Mindezeket figyelembe véve, a nyelvkontúr követése a nyelvhát vonalát meghatározó képpontok sorozatának megkeresését jelenti a releváns képtartományon belül. A nyelvkontúrkövetés elsődleges célja a különböző beszédhangokhoz tartozó nyelvállások és nyelvalakok statikus vagy dinamikus leírása, illetve a koartikuláció során létrejövő hangátmeneteket jellemző nyelvmozgások vizsgálata. A kvalitatív analízis mellett a nyelvkontúr a beszéd kvantitatív jellegű tanulmányozásának is jó kiindulópontja lehet, hiszen a nyelvkontúrból származtatható számszerű értékek elősegíthetik az artikulációs modellek mélyebb megértését és fejlesztését. A vizsgálataink során egy olyan kontúrkövető algoritmust dolgoztunk ki és fejlesztettünk MATLAB-környezetben, amely a dinamikus programozás technikáját alkalmazza (Zhao & Czap, 2019). A módszer segítségével illesztett nyelvkontúrt a 2. ábra UH-keretén figyelhetjük meg az /o/ hang esetében.



2. ábra. A radiális geometriában megjelenő nyelvkontúr görbéje egy /o/ hanghoz tartozó UH-kereten.

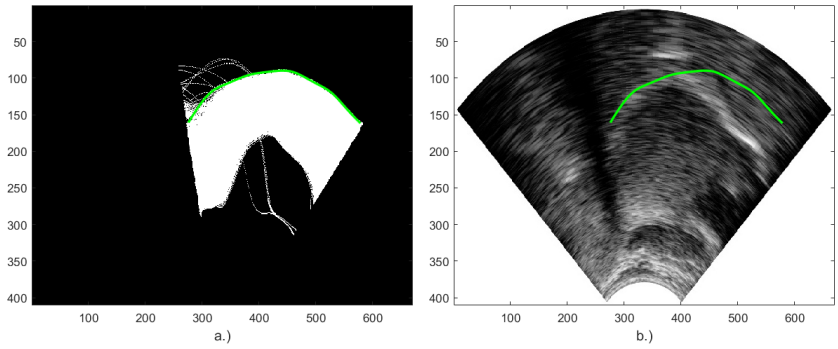
Az említett UH-felvételeken a nyelvkontúr mellett sokszor hasznos lehet a szájpadkontúr kijelölése is. Láttuk, hogy az UH-képeken egyáltalán nem mutatkozik meg sem a kemény, sem a lágy szájpad. Ezen oknál fogva különösen gondos megfontolásokat igényel a szájpadkontúr meghatározását célzó technika módszeres kidolgozása. Az UH-szájpadkontúr becslésének az volt az alapvető koncepciója, hogy a felvételeken megtaláljuk az artikuláció során a nyelvfelszín által érintett, legmagasabb helyzetben lévő szájüregi pontok halmazát, aminek nyomán valószínűsíthetjük a nyelv és a szájpad határvonalát. Ehhez természetesen olyan mássalhangzók vizsgálatára kell szorítkoznunk, melyek képzése során a nyelv biztosan érintkezik a szájpad palatális (kemény) és veláris (lágy) zónájával. Ez a feltétel a rendelkezésünkre álló, különböző bemondásokat tartalmazó UH-csomag esetében automatikusan teljesül, hiszen a rögzített mondatokban szereplő mássalhangzók képzésekor (pl. /k l ʃ t/) a nyelv más-más helyeken kerül kontaktusba a szájpad ívével. Így a szájpad kontúrjának kirajzolását lényegében egy szélsőérték-keresési probléma megoldásaként valósítottuk meg. Ehhez elsőként egy közös koordináta-rendszerben összegyűjtöttük a 40 UH-bemondás összes keretéhez tartozó 11111 darab nyelvkontúr görbéit, és ezzel párhuzamosan megalkottuk az összes nyelvkontúr lenyomata által kialakuló sűrűségtérképet is, melynek minden pixelje egy olyan valószínűségi mérték szerint kap értéket, hogy az adott képpont milyen gyakorisággal fordul elő lehetséges nyelvkontúrpontként. Az összes nyelvkontúr halmazát, illetve a görbeseregnek megfelelő sűrűségtérképet a 3. ábra demonstrálja.

A következő lépésben az összes nyelvkontúr által adott görbesereget egy olyan bináris sűrűségtérképen ábrázoltuk, amely csak 0 és 1 értékeket vehet fel aszerint, hogy a képmátrix adott pontját érinti-e bármely nyelvkontúrnak bármely pontja. Ennek értelmében a nyelvkontúrok által lefedett pixeleket 1 értékek jellemzik, a fennmaradó képpontokat pedig 0 értékekkel látjuk el, aminek folytán a kép fekete hátteréből fehér tartományként emelkedik ki az összes nyelvkontúr halmaza. A bináris sűrűségtérkép fekete-fehér kontrasztja kiváló terepet biztosít a kontúrkereső algoritmusunk futtatására, hiszen az eljárás segítségével detektálható a fekete és fehér domének maximális világosságú felső



3. ábra. A 40 UH-bemondás összes keretéhez tartozó 11111 darab nyelvkontúr halmaza (balra), illetve az összes nyelvkontúr által alkotott görbeseregnek megfelelő valószínűségi sűrűségtérkép (jobbra).

határvonala, ami éppen a szájpadkontúrt közelíti. A kontúrkereső algoritmussal kirajzolt szájpadkontúrt a 4. ábra zöld görbéje rögzíti a bináris sűrűségtérképen, illetve egy /k/ hanghoz tartozó UH-kereten.



4. ábra. A kontúrkereső algoritmussal detektált szájpadkontúr a bináris sűrűségtérképen (balra), illetve egy /k/ hanghoz tartozó UH-kereten (jobbra).

4. Artikulációs beszédszintézis

A gépi beszéd előállításának kiindulópontját az UH-felvételek képezték. A beszédszintézist olyan vizuális információkra támaszkodva valósítottuk meg, amik az említett kétdimenziós képi forrásokból közvetlenül vagy közvetve kinyerhetők. A szükséges geometriai adatok egy részét a felvételekre illesztett nyelv- és

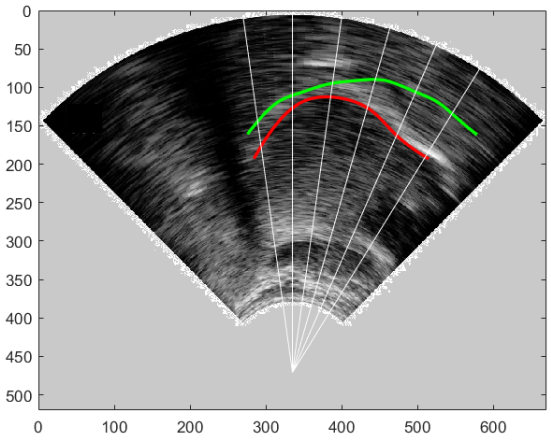
szájpadkontúrok segítségével származtattuk úgy, hogy MATLAB-környezetben kidolgoztunk egy algoritmust, melynek alkalmazásával a vokális traktusban dinamikus módon megmérhetők a szájpad és a nyelvfelszín közötti szagittális radiális távolságok. Az adathalmaz másik részét a nyelvkontúrokból kivont DCT-együtthatók formájában határoztuk meg. A kapott geometriai adatokat a gépi tanulás eszközei révén próbáltuk meg összekapcsolni a beszédet jellemző különböző artikulációs paraméterekkel, melyeket az akusztikus csőmodell vagy a lineáris predikció elvének keretében értelmeztük. Ennek során folyamatos beszédet kívántunk produkálni.

4.1. Dinamikus távolságmérés az UH-felvételeken

A vizuális adatok UH-felvételekből történő kinyeréséhez olyan anatómiai kontúrvonalakra volt szükségünk, melyek a lehető legteljesebb mértékben képesek lefedni a vokális traktus látótérbe eső tartományait. Ez az elvárás az UH-felvételek esetében már eddig is maximálisan teljesült, hiszen az előző fejezet szerint megkonstruált nyelv- és szájpadkontúrokon kívül nem detektálhatók további görbék a vokális traktusban. Ennek technikailag az az akadálya, hogy a részlegesen megjelenő szájüregen kívül az artikulációs csatorna többi része nem hozzáférhető az UH-letapogatás számára. A távolságméréshez két, különböző elveken alapuló mérési mechanizmust dolgoztunk ki az adatok dinamikus kinyerésére.

Az első megközelítésben a mérést radiális geometriában valósítottuk meg. Ehhez igazodva, elsőként egy rögzített középpontból indulva, sugárirányú metszeteket képeztünk az UH-felvételek releváns körcikkei által definiált tartományokban, amint azt az 5. ábra fehér vonalai is érzékeltetik a k hanghoz rendelt képkocka segítségével. A radiális metszetek a 0°-nál elhelyezkedő vertikális egyenes adott középpont körüli elforgatásával hozhatók létre a választott szögtartományon belül. Az 5. ábrán megrajzolt radiális metszetek a $[-8°, 32°]$ intervallum által megszabott szögtartományt fedik le úgy, hogy az egyes metszetek 8°-onként követik egymást. A releváns szögtartományt úgy állítottuk be, hogy az összes tanulmányozott bemondás képkockáira megfelelő legyen, azaz a

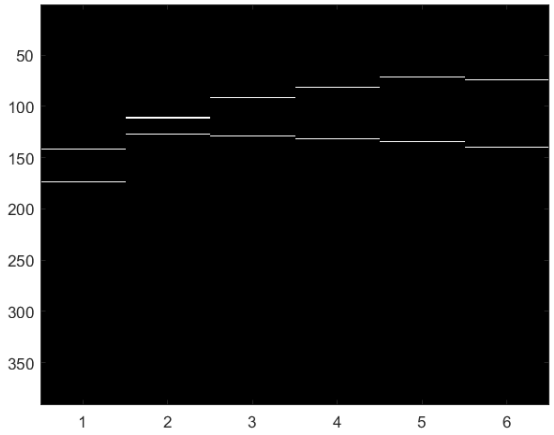
nyelvkontúr a nyelv mozgása során ne lépjen ki a felvett szögintervallumból. Az algoritmus második lépésében az 'intersect' függvény segítségével megkerestük a radiális metszetek és a nyelv-, illetve a szájpadkontúr által képzett metszéspontok koordinátáit, ami az 5. ábrával egyetértésben azt jelenti, hogy a fehér-piros, illetve a fehér-zöld görbepárok közös pontjait kell megtalálni. A görbepár metszéspontjainak birtokában kiszámolhatók a nyelv- és szájpadkontúr között mérhető radiális távolságok.



5. ábra. A szagittális radiális távolságok méréséhez felvett sugárirányú metszetek egy /k/ hanghoz tartozó UH-kereten.

A második megközelítésben a módszer kifejlesztése a radiális geometriának négyszöges elrendezésbe történő transzformációján alapult. Az eljárás fő lépése ugyanis az volt, hogy az UH-felvételek 5. ábrán rögzített középpontjából kiindulva és a kijelölt szögtartományt és szögbeosztást megőrizve, a létrejövő radiális irányok mentén mintavételeztük a nyelv- és szájpadkontúrokat, majd a kapott minták sorozatát mátrixos struktúrába rendeztük, ami a minták láncolatának négyszöges síkba való kifeszítését jelenti. A művelet eredményeként létrejövő, négyszöges geometriába alakított mintasorozatok láthatók a 6. ábrán, ahol a blokk felső, illetve alsó pixelláncai az 5. ábra szájpad-, illetve nyelvkontúrjaiból származnak. Ahogyan azt a 6. ábra vízszintes tengelye is mutatja, 6 oszlopot tartalmaznak a mátrixos struktúrák, ami azzal van összefüggésben, hogy

8°-onként beosztva a $[-8°, 32°]$ intervallumot, éppen 6 különböző szögérték áll elő. A mintasorozatok kényelmes alapot biztosítottak a távolságméréshez, amit úgy hajtottunk végre, hogy az összetartozó nyelv- és szájpadkontúrt leképező mátrix minden oszlopában kiszámítottuk a két vízszintes fehér vonal függőleges koordinátáinak különbségét.



6. ábra. Egy UH-keret szájpad- és nyelvkontúrjainak négyszöges geometriába transzformált mintasorozatai.

Kontroll céljából összehasonlítottuk a két távolságmérő módszer által produkált eredményeket, és arra jutottunk, hogy igen jó egyezés tapasztalható a két eljárással kapott kimenetelek között, hiszen a két adatsor elemeiben legfeljebb néhány pixelnyi eltérés mutatkozik. Ez a tendencia a felhasznált UH-keretek túlnyomó részében helytálló.

4.2. Diszkrét koszinusztranszformáció

A nyelvkontúrok mennyiségi jellemzésének egy lehetséges módja a diszkrét koszinusztranszformáció (Discrete Cosine Transform – DCT) alkalmazása, amely a matematikai transzformációknak egy speciális válfaja (Rao & Yip, 2014). A DCT segítségével egy N elemű valós $x_1, x_2, \ldots, x_N$ adathalmazt egy ugyanolyan elemszámú, szintén valós $X_1, X_2, \ldots, X_N$ értékhalmazra konvertálhatunk az

$$X_k = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} x_n \frac{1}{\sqrt{1+\delta_{k1}}} \cos\Big(\frac{\pi}{2N}(2n-1)(k-1)\Big) \qquad (1)$$

formula szerint, ahol $k = 1, 2, \ldots, N$. (1) alapján az inverz operáció

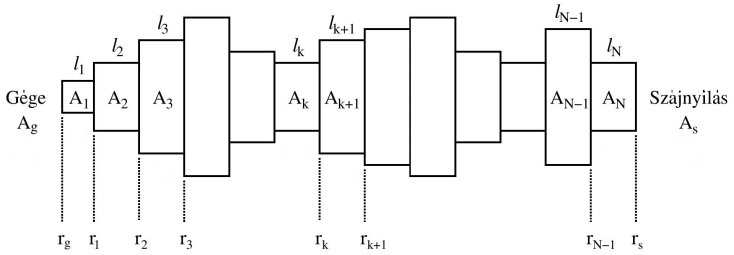
$$x_n = \sqrt{\frac{2}{N}} \sum_{k=1}^{N} X_k \frac{1}{\sqrt{1+\delta_{k1}}} \cos\Big(\frac{\pi}{2N}(2n-1)(k-1)\Big) \qquad (2)$$

alakban írható fel, mellyel az eredeti adattömb kapható vissza. Az (1) és (2) összefüggésekben δ_{k1} a Kronecker-deltát jelöli a $\delta_{k1} = 0$ (ha $k \neq 1$) és $\delta_{k1} = 1$ (ha $k = 1$) definíciónak megfelelően. Esetünkben az $x_1, x_2, \ldots, x_N$ paraméterek a nyelvkontúrok pontjainak vízszintes vagy függőleges koordinátáit testesíthetik meg, az $X_1, X_2, \ldots, X_N$ értékek pedig az adott görbéhez rendelt DCT-együtthatókat adják meg. A DCT inverz műveletével együtt kiválóan alkalmas görbék simítására, aminek igen nagy jelentősége van az UH-felvételekre illesztett nyelvkontúrok egyenetlenségeinek kiküszöbölésében is. Ennélfogva a DCT-együtthatók fontos információkat foglalnak magukba a nyelvkontúr alaki tulajdonságainak vonatkozásában.

4.3. Az akusztikus csőmodell

Az emberi beszédkeltés leképezésének egyik leggyakrabban alkalmazott és leghatékonyabb eszköze az akusztikus csőmodell (Fant, 1960). Ennek keretében a gégétől a száj- és orrnyílásig terjedő vokális traktust az egyik végén zárt, a másik végén nyílt rezonátorrendszerként kezeljük, mely akusztikai szűrőként működve egy több frekvenciakomponenst is tartalmazó hanghullámból csak adott frekvenciasávokba eső komponenseket enged át és sugároz ki a szabad térbe. A folytonos vokális traktust egy térben szabályosan kvantált csővel közelítjük, azaz felosztjuk N számú egymás után csatlakozó, egyenként állandó keresztmetszetű, rövid csőszakaszra, ahogyan azt a 7. ábra is szemlélteti. Az egyes csőszakaszok hosszúságát és keresztmetszeti paramétereit rendre az $l_1, l_2, l_3, \ldots, l_k, l_{k+1}, \ldots, l_{N-1}, l_N$, illetve az $A_1, A_2, A_3, \ldots, A_k, A_{k+1}, \ldots, A_{N-1}, A_N$ szimbólumok jelölik. A gégéhez és a szájnyíláshoz két extra keresztmetszetet

rendelünk A_g és A_s címkékkel. A 7. ábrán vázolt kép szerint gondolkodva azonnal felismerhető, hogy a szomszédos csőszakaszok csatlakozásánál történő ugrásszerű keresztmetszetváltásoknál a beszédhanghullámok visszaverődést szenvednek, amit fizikailag az $r_1, r_2, r_3, \ldots, r_k, r_{k+1}, \ldots, r_{N-1}$ reflexiós tényezőkkel írhatunk le. Ezt a halmazt kiegészítik a gégénél és a szájnyílásnál értelmezett r_g és r_s reflexiós tényezők. Az akusztikus csőmodell szerint $r_g = 1$ és $A_g = 1$. A modell tehát N csőszakasz alkalmazásakor $N + 1$ reflexiós tényezőt és $N + 2$ keresztmetszetet definiál.



7. ábra. A vokális traktus modellezése egyenként állandó keresztmetszetű, egyforma hosszúságú csőszakaszokkal.

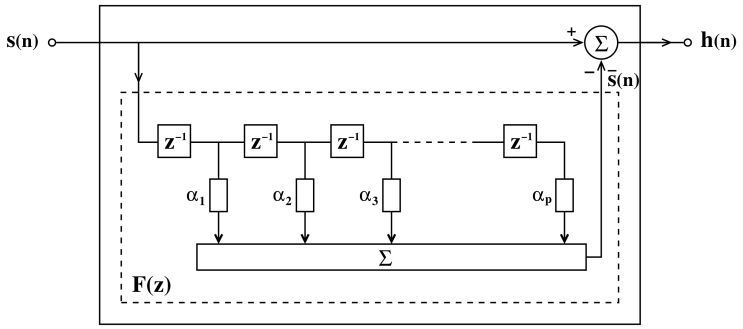
4.4. A lineáris predikció elve

A lineáris predikció (LPC – Linear Predictive Coding) a digitális jelek feldolgozásának és becslésének igen széles körben elterjedt és sikeres eszköze (Fant, 1960). Az eljárás azon alapul, hogy a jel adott pillanatbeli értékét az azt megelőző időpillanatokhoz tartozó jelértékek segítségével megpróbáljuk előrejelezni, idegen szóval predikálni. A predikció abban az esetben lineáris, ha a becsült jel a becslés során alkalmazott értékek lineáris függvénye. Amennyiben az s jel mint idősor n-edik elemét az azt megelőző p darab minta alapján állapítjuk meg, akkor a jelzett lineáris viszony az

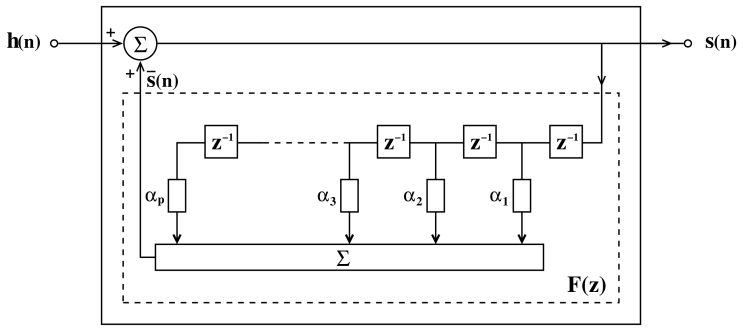
$$\overline{s}(n) = \sum_{i=1}^{p} \alpha_i \, s(n-i) \tag{3}$$

lineáris kombináció formájában jön létre, ahol az α_i koefficienseket predikciós együtthatóknak vagy LPC-együtthatóknak nevezzük. A p paraméter a predikció fokszámát adja meg.

A predikció 8. ábrán vázolt analízismodelljében a rendszer kimenetén az $s(n)$ eredeti és az $\overline{s}(n)$ becsült jel különbségeként előállítjuk a $h(n)$ hibajelet, a predikció 9. ábrán vizualizált szintézismodelljével pedig rekonstruálhatjuk az eredeti jelet a hibajel és a becsült jel összegeként. A 8. és 9. ábrák szaggatott vonallal határolt moduljaiban a z^{-1} faktorok egy ütemnyi késleltetést valósítanak meg, és a késleltetők láncolata mentén szorzásokat végzünk. Az i-edik lépésben kialakuló produktumot az aktuális leágazásnál található α_i koefficiens szorozza, végül a létrejövő tagokat a $\sum$ tömb összesíti.



8. ábra. A lineáris predikció analízismodellje.



9. ábra. A lineáris predikció szintézismodellje.

4.5. A neurális hálózatok felépítése

A beszédszintézisre irányuló törekvéseink során a gépi tanulás eszköztárához folyamodtunk, mivel a szintézishez szükséges artikulációs paramétereket geometriai adatokra alapozva, neurális hálózatok beiktatásával konstruáltuk meg. Folyamatos beszéd szintézisét irányoztuk elő, amihez az UH-felvételeken rögzített mondatokat használtuk fel. A hálózat bemeneti és kimeneti adathalmazát ugyanabból a forrásból származtattuk.

A rendszer bemenetén a nyelv- és szájpadkontúrok között mért szagittális radiális távolságokat vagy a nyelvkontúrokból kivont DCT-együtthatókat értelmeztünk. A távolságmérés során a korábban felvett 6 radiális metszettel dolgoztunk (5. ábra), a DCT-együtthatók számát pedig 9-re állítottuk be. A bemeneti paramétereket a korábbi fejezetekben ismertetett távolságmérő módszerek és transzformációs összefüggések szerint generáltuk. A rendszer kimenetén reflexiós tényezők vagy keresztmetszetek betanítását irányoztuk elő úgy, hogy 18 reflexiós tényezőt és 19 keresztmetszetet definiáltunk. A kimeneti paraméterek előállításához az adott bemondás akusztikus jeléből az 'lpc' függvény felhasználásával kiszámítottuk az LPC-együtthatókat, melyekből az 'lpcar2rf', majd az 'lpcrf2aa' függvények alkalmazásával meghatároztuk a reflexiós tényezők, illetve a vokális traktus keresztmetszeteinek halmazát. A neurális hálózat $(m \times n)$-es mátrixok formájában kezelt bemeneti és kimeneti blokkjainak dimenzióit a különböző rendszerbeállítások esetében az 1. táblázat foglalja össze. Az m sorindex a vizsgálatban részt vevő összes bemondáshoz tartozó képkeretek számát jelzi, míg az n oszlopindex a felvett radiális távolságok, DCT-együtthatók, reflexiós tényezők vagy keresztmetszetek számát fejezi ki.

Az 1. táblázatban feltüntetett paramétereket speciális technikával vontuk ki a beszédjelből. Folyamatos beszédről lévén szó, az egymáshoz kapcsolódó beszédhangok láncolata mentén haladva dinamikusan változik a vokális traktus alakja, és ezt a dinamikát a paraméterhalmaznak is tükröznie kell. Ezért a bemeneti adatok számításakor az adott bemondás összes keretét figyelembe kellett vennünk úgy, hogy minden egyes kerethez hozzárendeltük az aktuális nyelv- és szájpadkontúr alkalmazásával kapott radiális távolságokat és DCT-

1. táblázat. A neurális hálózat bemeneti és kimeneti adathalmazainak dimenziói az összes rendszerkonfiguráció esetében.

a bemenet mérete		**a kimenet mérete**	
radiális táv	**DCT-egyh.**	**reflexiós tény.**	**keresztmetszet**
9×6	–	9×18	9×19
4354×13	4354×9	4354×18	4354×19
6278×6	6278×9	6278×18	6278×19

együtthatókat. A kimeneti adatok kivonásakor pedig az adott bemondás teljes mintaszámát elosztva az összes keret számával, meghatároztuk a keretenkénti minták számát, így – a bemeneti oldalon érvényes módszerhez hasonlóan – ez esetben is minden egyes kerethez társítottuk az aktuális hangminták csoportjából eredeztetett reflexiós tényezőket és keresztmetszeteket.

A tanulóalgoritmus konstrukciójakor az skálázott konjugált gradiens (SCG – Scaled Conjugate Gradient) módszert alkalmaztuk a kimeneti paraméterek betanítására (Moller, 1993). A rendszerben egy rejtett réteget helyeztünk el, melybe 100 neuront ültettünk be. A hálózat kimeneti rétegének aktivációs függvényét a kimeneti paraméter típusához igazodva választottuk meg. Mivel a reflexiós tényezők értéke lehet pozitív és negatív is, esetükben megtartottuk a standard lineáris átvitelt, a keresztmetszetek tanításakor azonban a hiperbolikus tangens szigmoid függvény mellett döntöttünk, kizárva ezzel a negatív keresztmetszetek kialakulásának lehetőségét.

5. A beszédszintézis módszerei

5.1. Az akusztikus csőmodellen alapuló szintézis

Az akusztikus csőmodell keretében a vokális traktusban terjedő hanghullámot speciális fizikai mennyiségekkel jellemezhetjük, melyek közül programozástechnikai szempontból az $u(x,t)$ térfogatáram a legalkalmasabb a folyamatok modellezésére. A kétváltozós térfogatáram a vokális traktus tengelye mentén

mért x pozíciótól és a t időtől függ, és megmutatja az artikulációs csatorna adott keresztmetszetén egységnyi idő alatt átáramló levegő térfogatának mértékét. A térfogatáramot vektorfizikai mennyiségként kezeljük, mivel a nagyságán túl az iránya is mérvadó. A pozitív, illetve negatív x irányba terjedő térfogatáramot az $u^+(x,t)$, illetve az $u^-(x,t)$ szimbólumokkal láthatjuk el. A modell hipotéziseinek betartásával levezethető egy olyan komplett egyenletrendszer, mely egyértelműen összekapcsolja a szomszédos csőszakaszok határánál létrejövő térfogatáramokat, méghozzá oly módon, hogy a vokális traktus három sajátos tartományát külön-külön kezelve, eltérő szerkezetű egyenletcsoport alakul ki a gégénél található gerjesztési oldalon, a közbenső csőszakaszok régiójában, valamint a szájnyílásnál lévő sugárzási oldalon. A keletkező egyenletrendszer az

$$\begin{aligned}
u_1^+(t) &= \frac{1+r_g}{2}\, u_g(t) + r_g\, u_1^-(t) && \text{(a)} \\
u_{k+1}^+(t) &= (1+r_k)\, u_k^-(t-\tau_k) + r_k\, u_{k+1}^-(t) \\
u_k^-(t+\tau_k) &= -r_k\, u_k^+(t-\tau_k) + (1-r_k)\, u_{k+1}^-(t) && \text{(b)} \\
u_N^-(t+\tau_N) &= -r_s\, u_N^+(t-\tau_N) \\
u_s(t) &= -(1+r_s)\, u_N^+(t-\tau_N) && \text{(c)}
\end{aligned} \tag{4}$$

alakot ölti, ahol az (a), (b), (c) blokkok rendre a gerjesztési oldal, a közbenső csőszakaszok, illetve a sugárzási oldal térfogatáram-viszonyait jellemzik. A (4) egyenletrendszer tömörített jelölésformái szerint a k-adik csőszakasz elején pozitív vagy negatív irányba haladó térfogatáram az

$$u^{\pm}(x_k,t) = u_k^{\pm}(t), \tag{5}$$

míg a k-adik csőszakasz végénél pozitív vagy negatív irányba haladó térfogatáram az

$$u^{\pm}(x_k+l_k,t) = u_k^{\pm}(t\pm\tau_k) \tag{6}$$

szimbólumoknak megfelelően egyszerűsödik, $k = 1,2,\ldots,N$ indexelés mellett. Az (5) és (6) egyenlőségek bal oldali argumentumában szereplő helykoordinátát tehát a jobb oldali alsó index váltja fel, a (6)-ban megjelenő τ_k paraméter pedig azt az időeltolódást határozza meg, ami ahhoz szükséges, hogy

a $v = 343,14$ m/s sebességgel terjedő hanghullám áthaladjon az l_k hosszúságú csőszakaszon. A (4.a) és (4.c) egyenletekben szereplő $u_g(t)$ és $u_s(t)$ függvények speciálisan a gége által indukált gerjesztőjelet, illetve a szájnyíláson keresztül a szabad térbe sugárzott beszédjelet testesítik meg, az r_g, r_k, r_s faktorok pedig a 7. ábra szerint bevezetett reflexiós tényezőkkel azonosíthatók. A gépi beszéd produkciója a (4) egyenletrendszer programszintű ciklikus megvalósításával lehetséges, melynek során döntő szerepe van az $u_g(t)$ gerjesztőjelnek, hiszen a jel alakja alapvetően befolyásolja a szintézis eredményét. A gerjesztőjel jellegét természetesen az előállítandó hang típusa határozza meg attól függően, hogy zöngés vagy zöngétlen módon képződik. Zöngés gerjesztés esetén a gerjesztőjel periodikusan ismétlődő impulzusok formájában jön létre, zöngétlen gerjesztés esetén pedig véletlenszerűen változik. Az egymást követő beszédhangok típusának véletlenszerűségéhez igazodó zöngés-zöngétlen fluktuáció legegyszerűbben úgy indukálható, hogy gerjesztőjelként a lineáris predikció keretében értelmezett hibajelet alkalmazzuk. A csőmodell megvalósításához szükséges reflexiós tényezőket részben közvetlenül a neurális hálózat kimeneti csatornájából nyertük eleve betanított paraméterek formájában, másrészt pedig a betanított keresztmetszeteket átalakítottuk reflexiós tényezőkké az 'lpcaa2rf' függvény segítségével.

5.2. A lineáris predikció elvén alapuló szintézis

A predikciós elvű beszédszintézis végrehajtása a 8. és 9. ábrákon vázolt analízis- és szintézismodellek programszintű megalkotása révén lehetséges. Ehhez mindkét modell esetében ismernünk kell a teljes rendszer átviteli függvényét, ami a 8. és 9. ábrákon szaggatott vonallal határolt modul

$$F(z) = \sum_{i=1}^{p} \alpha_i \, z^{-i} \tag{7}$$

átviteli függvényére vezethető vissza, ahol z a komplex körfrekvenciát jelöli. Ennek megfelelően az analízismodell szerint megszerkesztett rendszer átviteli

függvénye

$$A(z) = 1 - F(z) = 1 - \sum_{i=1}^{p} \alpha_i \, z^{-i} = \sum_{i=0}^{p} \beta_i \, z^{-i} \tag{8}$$

alakban, a szintézismodell keretében felépített rendszer átviteli függvénye pedig

$$B(z) = \frac{1}{1 - F(z)} = \frac{1}{1 - \sum_{i=1}^{p} \alpha_i \, z^{-i}} = \frac{1}{\sum_{i=0}^{p} \beta_i \, z^{-i}} \tag{9}$$

formában kapható meg. A (8)-(9) egyenletekben megtörtént a (7) kifejezés helyettesítése, az utolsó egyenlőségek felírásakor pedig az α_i együtthatókat a β_i faktorokká transzformáltuk. A (8)-(9) átviteli függvényeket alapul véve, az analízis- és szintézismodell kimeneti jelei a 'filter' függvény segítségével generálhatók 'filter(a, b, x)' struktúrával, ahol az x-szel jelölt bemeneti adatokat a szűrésnek alávetett szintetizálandó hangminták vagy a hibajel alkotják, az a és b paraméterek pedig a vokális traktus átviteli függvényének számlálójában és nevezőjében megjelenő koefficienseket adják meg. Ennek megfelelően, a (8)-(9) kifejezésekhez igazodva, az analízismodell hibajelét az $a = \beta_i$ és $b = 1$, a szintézismodellel produkált beszédjelet pedig az $a = 1$ és $b = \beta_i$ együtthatók beállításával biztosíthatjuk. A hibajel előállításához szükséges β_i LPC-együtthatókat az 'lpc' függvény alkalmazásával vontuk ki az eredeti beszédjelből. A hangminták szintetizálásában részt vevő β_i LPC-együtthatókat pedig a betanított reflexiós tényezőkből, illetve keresztmetszetekből származtattuk az 'lpcrf2ar', illetve az 'lpcaa2rf' függvények felhasználásával.

6. Eredmények

Az eredeti és a szintetizált beszédjelek elérhetők és meghallgathatók az alábbi linken: https://drive.google.com/drive/folders/1LIz6y6wKZfMRIyE4EP1YW62SqXiZ-W8K?usp=drive_link. A 'MONDATOK' főmappában található almappák elnevezése minden esetben 'X_Y_Z' szerkezetű, ahol 'X' a szintézis során alkalmazott módszert rejti, 'Y' és 'Z' pedig a neurális hálózat bemeneti és kimeneti paramétereit egyértelműsíti, tehát az 'X_Y_Z' címke 'Y' adatokkal táplált és 'Z' adatokat betanuló neurális hálózat kimeneti eredményeit felhasználó, 'X' módszerrel realizált szintézist takar. Ezzel összhangban az 'X' helyére 'cso' vagy

'lpc' kerül, ami az akusztikus csőmodellt vagy a lineáris predikció elvét jelzi. Az 'Y' pozícióban 'tav' vagy 'dct' feliratok lehetségesek, melyek a radiális távolságokat vagy a DCT-együtthatókat jelölik. A 'Z' elem 'ref' vagy 'ker' alakú, utalva a reflexiós tényezőkre vagy a keresztmetszetekre. Ezek mellett a főmappa tartalmazza az eredeti bemondások audiofelvételeit is. Az eredményeink elemzése során arra a következtetésre jutottunk, hogy a beszédszintézis minden esetben sikerrel zárult, mivel kvalitatív és kvantitatív szinten is relatíve jó egyezést tapasztaltunk az eredeti és a szintetizált bemondások között.

A kvalitatív értékelés során azt állapítottuk meg, hogy a szintetizált bemondások mindegyike elég jól érthető és a beszéd felismerhető az eredeti hanginformáció nélkül is, bár megjegyezzük, hogy a szintetizált jelekben zajkomponensek is tapasztalhatók, ami valamelyest rontja az akusztikai élményt, de természetesen a kutatómunkánk későbbi fázisaiban szeretnénk majd a torzításokat a lehető legjobb mértékben kiküszöbölni és javítani a beszéd minőségén.

Az eredményeink értékeléséhez néhány szintetizált bemondást szubjektív audioteszt formájában véleményezésre bocsátottunk egy erre a feladatra felkért, sok résztvevős célcsoport körében, akik független minősítőként semmilyen módon nem kapcsolódtak a kutatáshoz. A felmérés célja az volt, hogy a szubjektív audioteszt kimenetelének ismeretében egyértelműen állást lehessen foglalni, hogy az 1. táblázatban rendszerezett konfigurációk közül melyik bizonyul a leghitelesebbnek a beszédszintézisben, azaz melyik neurális hálózat beállításával lehet a legjobb minőségű beszédet előállítani. Ezzel a vizsgálattal nem a modellszintű hatékonyságot szerettük volna tesztelni, ezért az akusztikus csőmodell és a lineáris predikció révén kapott hangminták összehasonlítására irányuló kísérlet nem történt. Számunkra inkább az volt a legfőbb eldöntendő kérdés, hogy az 1. táblázat mely paraméterkombinációjának alkalmazásával érhető el a legtisztábban érthető gépi beszéd. A szubjektív minősítés megvalósításához kiválasztottunk egy lineáris predikció szerint szintetizált mondatot, és azt négy különböző változatban tártuk a célcsoport elé. A kiválasztott mondat A, B, C, D címkékkel ellátott négy verzióját a 2. táblázat sűríti, ahol azonosítható a szintézis alapjául szolgáló neurális hálózat bemeneti és kimeneti paramétertípu-

sainak összekapcsolása. A minősítésben részt vevő személyeket nem szerettük volna semmilyen módon befolyásolni, így nem hoztuk a tudomásukra, hogy az A, B, C, D felvételekhez milyen paraméterek vannak hozzárendelve, mindössze annyit közöltünk velük, hogy a négy mondatot négy különböző metódus szerint generáltuk, és pusztán az auditív percepció alapján kellett eldönteniük, hogy melyik esetben érthető legjobban a beszédjel.

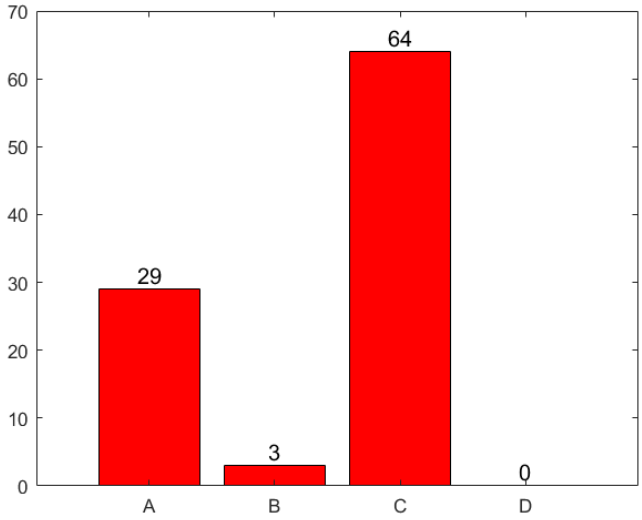
2. táblázat. A szubjektív tesztelésre bocsátott szintetizált audiofelvétel négy különböző változata a neurális hálózat paramétereinek függvényében.

	bemeneti paraméter	**kimeneti paraméter**
A	radiális távolság	reflexiós tényező
B	radiális távolság	keresztmetszet
C	DCT-koefficiens	reflexiós tényező
D	DCT-koefficiens	keresztmetszet

A szubjektív értékeléshez összesen 96 alany csatlakozott. Véleményük megoszlását a 10. ábra szemlélteti, ahol megfigyelhető, hogy 3 résztvevő kivételével senki nem voksolt a B és D variánsokra, a szavazatok lényegében az A és C verziók szintjén differenciálódnak többségében C szerinti állásfoglalással a 29:64 aránynak megfelelően, vagyis az A-hoz viszonyítva kb. kétszer annyian jelölték be a C-t. Ez az eredmény tehát arra enged következtetni, hogy a keresztmetszetekkel szemben a reflexiós tényezők betanításával jobb minőségű gépi beszéd állítható elő, és a tanítóalakzatok szintjén a radiális távolságokkal szemben előnyt élveznek a DCT-koefficiensek.

7. Összegzés

Jelen tanulmányban az artikulációs beszédszintézis különböző aspektusait vizsgáltuk MATLAB-környezetben. Elemzéseink során olyan ultrahangos (UH) technikával készült audiovizuális forrásokra támaszkodtunk, melyek a kétdimenziós szagittális síkban vizualizálják a vokális traktus hangképző szerveinek rela-



10. ábra. Az auditív percepción alapuló szubjektív tesztben részt vevők szavazatainak megoszlása négy különböző módon szintetizált mondat (A, B, C, D) próbájakor.

tív helyzetét és mozgását, miközben rögzítik a beszélő által kibocsátott audiojelet. Így a kép- és hangtartalom a szinkronizációnak köszönhetően egyértelmű módon összekapcsolódik.

Az UH-felvételek jó alapot biztosítottak ahhoz, hogy geometriai és akusztikus paramétereket nyerhessünk ki a kép- és hangforrásokból. A geometriai adatokhoz való hozzáférést nagyban megkönnyítette az automatikus kontúrkövető algoritmusunk alkalmazása, melyekkel elvégeztük az UH-keretek nyelvkontúrjainak dinamikus letapogatását. Az így kapott görbecsoportot kiegészítettük az UH-felvételekre rajzolt szájpadkontúrral, amit egy általunk kidolgozott eljárással konstruáltunk meg. A nyelvkontúrok ismeretében származtattuk a görbék diszkrét koszinusztranszformációjában részt vevő együtthatókat (DCT-együtthatók). Emellett a nyelv- és szájpadkontúrokból kiindulva kifejlesztettünk két különböző módszert a nyelv és szájpad anatómiai felszínei között mérhető szagittális radiális távolságok dinamikus meghatározására. Az akusztikus paramétereket a beszédjelek időfüggvényeiből eredeztettük a lineáris predikció elvéhez kapcsolódó LPC-együtthatók formájában, melyeket az akusztikus cső-

modellben értelmezett reflexiós tényezőkké, illetve a vokális traktus keresztmetszeti adataivá konvertáltunk.

Az artikulációs beszédszintézis során folyamatos beszédet állítottunk elő. A szintézis előkészítéseként olyan neurális hálózatokat szerkesztettünk, amik bemeneti adatként fogadják a vokális traktus szagittális radiális távolságaival és a nyelvkontúrokból kivont DCT-együtthatókkal felépített mátrixokat, a kimeneten pedig a beszédjelből származtatott reflexiós tényezők, valamint keresztmetszetek által alkotott struktúrákat produkálnak. A betanított paraméterek felhasználásával beprogramoztuk az akusztikus csőmodellt, illetve a lineáris predikció analízis- és szintézismodelljeit, melyek segítségével véghez vittük a gépi beszédprodukciót.

Köszönetnyilvánítás

Őszintén hálásak vagyunk és kegyeletteljes köszönetet mondunk Csapó Tamás Gábornak, hogy – amíg közöttünk volt – végtelenül segítőkész és önzetlen módon a rendelkezésünkre bocsátotta az MTA-ELTE Lendület Lingvális Artikuláció Kutatócsoportjának Micro rendszerével készült UH-felvételeket.

Hivatkozások

Arik, S., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning, PMLR* (p. 195–204). volume 70.

Arthur, F., & Csapó, T. (2021). Towards a practical lip-to-speech conversion system using deep neural networks and mobile application frontend. In *The International Conference on Artificial Intelligence and Computer Vision* (p. 441–450).

Ashok, K., Ashraf, M., Thimmia Raja, J., Hussain, M., Singh, D., & Haldorai, A. (2022). Collaborative analysis of audio-visual speech synthesis with sensor

measurements for regulating human–robot interaction. *International Journal of System Assurance Engineering and*, *Management*, 1–8.

Balogh, G., Dobler, E., Grobler, T., Smodies, B., & Szepesvári, C. (2000). Flexvoice: A parametric approach to high-quality speech synthesis. In *IEE Seminar on State of the Art in Speech Synthesis* (p. 189–194).

Birkholz, P., Kürbis, S., Stone, S., Häsner, P., Blandin, R., & Fleischer, M. (2020). Printable 3d vocal tract shapes from mri data and their acoustic and aerodynamic properties. *Scientific Data*, *7*, 255.

Cao, B., Ravi, S., Sebkhi, N., Bhavsar, A., Inan, O., Xu, W., & Wang, J. (2023). Magtrack: A wearable tongue motion tracking system for silent speech interfaces. *Journal of Speech, Language, and Hearing Research*, *66*, 3206–3221.

Carlson, R., & Granström, B. (2008). *Rule-based speech synthesis*. Springer Handbook of Speech Processing.

Csapó, T. (2020). Speaker dependent articulatory-to-acoustic mapping using real-time mri of the vocal tract. In *INTERSPEECH 2020* (p. 2722–2726).

Csapó, T., Gosztolya, G., Tóth, L., Shandiz, A., & Markó, A. (2022). Optimizing the ultrasound tongue image representation for residual network-based articulatory-to-acoustic mapping. *Sensors*, *22*, 8601.

Csapó, T., Grósz, T., Gosztolya, G., Tóth, L., & Markó, A. (2017). Dnn-based ultrasound-to-speech conversion for a silent speech interface. In *INTERSPEECH* (p. 3672–3676).

Csapó, T., Zainkó, C., Tóth, L., Gosztolya, G., & Markó, A. (2020). Ultrasound-based articulatory-to-acoustic mapping with waveglow speech synthesis. In *INTERSPEECH 2020* (p. 2727–2731).

Denby, B., Csapó, T., & Wand, M. (2023). Future speech interfaces with sensors and machine intelligence. *Sensors*, *23*.

Denby, B., & Stone, M. (2024). Speech synthesis from real time ultrasound images of the tongue. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing* (p. –685). volume 1.

Fant, G. (1960). Acoustic theory of speech production.

Gonzalez-Lopez, J., Gomez-Alanis, A., Donas, J., Pérez-Córdoba, J., & Gomez, A. (2020). Silent speech interfaces for speech restoration: A review. *IEEE access*, *8*, 177995–178021.

James, J., Balamurali, B., Watson, C., & MacDonald, B. (2021). Empathetic speech synthesis and testing for healthcare robots. *International Journal of Social Robotics*, *13*, 2119–2137.

Jin, Y., Gao, Y., Xu, X., Choi, S., Li, J., Liu, F., Li, Z., & Jin, Z. (2022). Earcommand: "hearing" your silent speech commands in ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *6*, 1–28.

Juanpere, E., & Csapó, T. (2019). Ultrasound-based silent speech interface using convolutional and recurrent neural networks. *Acta Acustica united with Acustica*, *105*, 587–590.

Kaburagi, T. (2014). Determining the length and cross-sectional area of the vocal tract jointly from formants using acoustic sensitivity function. *Acoustical Science and Technology*, *35*, 290–299.

Kaburagi, T. (2015). A method for estimating vocal-tract shape from a target speech spectrum. *Acoustical Science and Technology*, *36*, 428–437.

Kaur, N., & Singh, P. (2022). Speech waveform reconstruction from speech parameters for an effective text to speech synthesis system using minimum phase harmonic sinusoidal model for punjabi. *Multimedia Tools and Applications*, *81*, 26101–26120.

Király, J. (1989). A pc talker beszédszintetizátor és digitális hangrögzítő-visszajátszó rendszer. O.

Kiss, G., Arató, A., Lukács, J., Sulyán, J., & Vaspöri, T. (1987). Brailab, a full hungarian text-to-speech microcomputer for the blind. In *Proceedings of World Conference on Phonetics* (p. 1–4).

Leppävuori, M., Lammentausta, E., Peuna, A., Bode, M., Jokelainen, J., Ojala, J., & Nieminen, M. (2021). Characterizing vocal tract dimensions in the vocal modes using magnetic resonance imaging. *Journal of Voice*, *35*, 804– 27.

Mahum, R., Irtaza, A., & Javed, A. (2023). Text to speech synthesis using deep learning. In *Intelligent Multimedia Signal Processing for Smart Ecosystems* (p. 289–305). Cham: Springer International Publishing.

Moller, M. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, *6*, 525–533.

Mullah, H. (2015). A comparative study of different text-to-speech synthesis techniques. *International Journal of Scientific Engineering and Research*, *6*, 287–292.

Olaszy, G. (1989). Multivox – a flexible text-to-speech system for hungarian, finnish, german, esperanto, italian, and other languages for ibm-pc. In *First European Conference on Speech Communication and Technology (EUROSPEECH '89* (p. 2525–2528). Paris: France.

Olaszy, G., G., O., P., K., G., Z., C., & Gordos, G. (2000). Profivox — a hungarian text-to-speech system for telecommunications applications. *International Journal of Speech Technology*, *3*, 201–215.

Olaszy, G., & Gordos, G. (1987). Automatic text-to-speech system applied in a reading machine. In *Proceedings of European Conference on Speech Technology (EUROSPEECH '87* (p. 25–29). Edinburgh, UK.

Otani, Y., Sawada, S., Ohmura, H., & Katsurada, K. (2023). Speech synthesis from articulatory movements recorded by real-time mri. In *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech* (p. 127–131).

Panda, S., & Nayak, A. (2017). A waveform concatenation technique for text-to-speech synthesis. *International Journal of Speech Technology*, *20*, 959–976.

Pang, B., Teng, J., Xu, Q., Song, Y., Yuan, X., & Li, Y. (2023). Chinese personalised text-to-speech synthesis for robot human–machine interaction. *IET Cyber-Systems and Robotics*, *5*, 12098.

Rao, K., & Yip, P. (2014). *Discrete cosine transform: algorithms, advantages, applications*. Academic Press.

Shiga, Y., Ni, J., Tachibana, K., & Okamoto, T. (2020). Text-to-speech synthesis. In Y. Kidawara, E. Sumita, & H. Kawai (Eds.), *Speech-to-Speech Translation*. Singapore: Springer Briefs in Computer Science. Springer.

Skordilis, Z., Toutios, A., Töger, J., & Narayanan, S. (2017). Estimation of vocal tract area function from volumetric magnetic resonance imaging. In *2017 IEEE International Conference on Acoustics, Speech and Speech and Signal Processing (ICASSP* (p. 924–928).

Toutios, A., & Narayanan, S. (2013). Articulatory synthesis of french connected speech from ema data. In *Interspeech* (p. 2738–2742).

Tóth, L., Gosztolya, G., Grósz, T., Markó, A., & Csapó, T. (2018). Multi-task learning of speech recognition and speech synthesis parameters for ultrasound-based silent speech interfaces. In *INTERSPEECH* (p. 3172–3176).

Zhao, L., & Czap, L. (2019). A nyelvkontúr automatikus követése ultrahangos felvételeken. *Beszédkutatás*, *27*, 331–343.

Estimation of second subglottal resonance based on F2 measurements and its application in consonant-vowel classification in Hungarian[a)]

Tamás Gábor Csapó[b)]
Laboratory of Speech Technology, Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary

(Dated: June 21, 2009)

It has been shown for several languages that subglottal resonances (SGRs) play a dividing role in the frequency space of consonants and vowels (e.g. vowels are separated into the back-front categories by the second subglottal resonance). Consonant-vowel transitions are characterized by a regression line (locus equation), and can be classified into distinct categories in the locus equation space, according to their place of articulation. Several attempts have shown that the dividing lines between these categories may be the SGRs. In this paper, the relation between CV transitions in the locus equation space and the separating role of the subglottal resonances are further investigated. Locus equation space of one native speaker of Hungarian is examined. Consonant-vowel transitions are classified based on SGRs estimated from the locus equations of a subset of CV sequences. The hit rates and false alarm rates of the classification are comparable to a baseline experiment where the subglottal resonances were measured from accelerometer signal.

PACS numbers:
Keywords: subglottal resonance, locus equation, Hungarian

I. INTRODUCTION

It has been shown for several languages that subglottal resonances (SGRs) play a dividing role in the frequency space of consonants and vowels (e.g. American English: Lulich (2009) and Hungarian: Csapó *et al.* (2009)). It seems that speakers try to avoid putting formants in the regions of SGRs, consequently vowels are divided into distinct categories (e.g. back vs. front by $Sg2$ and low vs. non-low by $Sg1$, see Csapó *et al.* (2009)).

Consonant-vowel transitions are characterized by a regression line (locus equation), which shows the correlation between the second formant at the onset of voicing in the consonant ($F2_{burst}$) and that of at the steady state of the vowel ($F2_{vowel}$), as described in Lulich (2008). When illustrating $F2_{burst}$-$F2_{vowel}$ together in the locus equation space, groups of CV transitions are distinguishable according to their place of articulation. Chen and Lulich (2009) showed that the boundaries between these distinct groups are related to subglottal resonances.

In this experiment, the relation between CV transitions in the locus equation space and the separating role of the subglottal resonances is further investigated based on the acoustic data of a native speaker of Hungarian. After that, an $Sg2$ estimation algorithm (Chen and Lulich, 2009) is applied and further improved for use in Hungarian. Subglottal resonance frequencies of six other speakers are measured and examined, in order to find correlation between $Sg1$, $Sg2$ and $Sg3$. For the first speaker, estimated SGR values are used in the clustering of CV transitions. Hit rates and false alarm rates of the classification of consonants and vowels are investigated.

II. METHODS

A. Accelerometer recordings

Acoustic data were collected from seven native speakers of Hungarian (referred as TB and S1-S6) in an anechoic chamber. While the speakers uttered several sentences read from a paper, voice and accelerometer recordings were done. The speech utterances were recorded using an EMC 100 condenser microphone at a distance of approximately 15 cm from the lips. The subglottal data were recorded using a K&K HotSpot accelerometer attached to the skin of the neck below the thyroid cartilage. The two signals were digitized at 8 kHz with a Terratec DMX 6 Fire USB external sound card and were recorded to separate channels using Wavesurfer (Sjölander and Beskow, 2009). These voice recordings were unrelated for this experiment.

B. Subglottal resonance measurements

The first three subglottal resonances were measured manually 25 times from the accelerometer signal of speaker TB, using Wavesurfer. A sample FFT taken from the accelerometer signal can be seen in Fig. 1, which shows that this measurement is similar to reading off formants. A detailed description about measuring SGRs can be found in Lulich (2009) and Chi and Sonderegger (2007). For speakers S1-S6, the subglottal resonances were measured 10 times from their accelerometer signal.

The mean, median and standard deviation values of

[a)] Final project for "Speech Acoustics" course
[b)] Electronic address: csapot@tmit.bme.hu; URL: http://speechlab.tmit.bme.hu/csapo

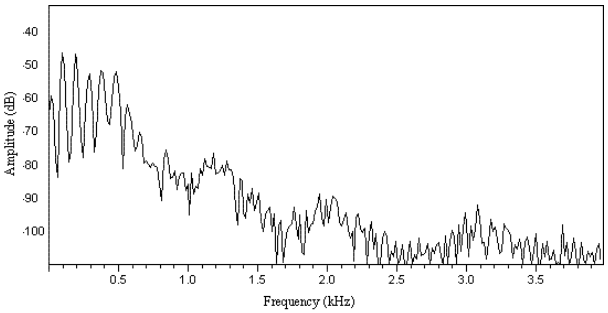


FIG. 1. Sample FFT from the accelerometer data of speaker TB. Spectral peaks (454 Hz, 1211 Hz, 2023 Hz and 3067 Hz) may correspond to subglottal resonances. However, $Sg1$ seems to be very low due to the coupling to the supraglottal tract.

subglottal resonances for speaker TB are shown in Table I. Table II shows the medians of SGRs, for speakers S1-S6. The measured resonance frequencies correspond to the values reported in the literature (Stevens, 1998).

C. Speech recordings

In a separate recording session, speaker TB uttered "ɔCVbɔ"-like nonsense words in an anechoic room. At the place of the first consonant, all 8 Hungarian stop consonants were included (labials: [b,p], alveolars: [d,t], velars: [g,k] and palatals: [ɟ,c]). All of the 14 Hungarian vowels were included in the second non-stressed syllable ([ɔ,aː,o,oː,u,uː,ɛ,eː,i,iː,ø,øː,y,yː]). The nonsense words were uttered in a randomized order, 10 times each, for a total of 1120 utterances. Voice data were recorded using an EMC 100 condenser microphone and digitized at 48kHz with a Terratec DMX 6 Fire USB external sound card using Wavesurfer.

TABLE I. *Mean, median and standard deviation of measured SGRs for speaker TB.*

	$Sg1$	$Sg2$	$Sg3$
Mean	545 Hz	1241 Hz	2027 Hz
Median	554 Hz	1244 Hz	2022 Hz
Std. dev.	60	42	145

TABLE II. *Median of measured SGRs for speakers S1-S6.*

	$Sg1$	$Sg2$	$Sg3$
S1	612 Hz	1498 Hz	2283 Hz
S2	572 Hz	1417 Hz	2265 Hz
S3	662 Hz	1561 Hz	2420 Hz
S4	577 Hz	1235 Hz	1994 Hz
S5	617 Hz	1412 Hz	2237 Hz
S6	582 Hz	1306 Hz	2070 Hz

TABLE III. *Medians of the measured $F2_{burst}$ values for speaker TB (all numbers in Hz). $F2_{burst}$ values were measured at the burst of the stop consonants.*

		Labial		Alveolar		Velar		Palatal	
		b	p	d	t	g	k	ɟ	c
Back	ɔ	1045	1435	1074	2022	1066	1001	1560	2058
	o	830	1304	843	1651	878	841	1514	1714
	oː	817	1374	853	1632	782	793	1499	1860
	u	805	1486	852	1703	807	789	1587	1899
	uː	825	1435	805	1690	784	798	1526	2035
Front	aː	1236	1655	1714	2001	1752	1266	1638	2106
	ɛ	1518	1726	2021	2181	2101	1542	1753	2179
	ø	1348	1661	1374	2076	1524	1390	1695	2018
	øː	1518	1726	1635	2055	1688	1525	1729	2007
	y	1594	1841	1730	2149	1809	1569	1860	2116
	yː	1708	1899	1796	2198	1961	1774	1934	2149
	eː	1769	1894	2112	2242	2299	1997	1880	2264
	i	1939	2022	2242	2225	2292	1956	1947	2244
	iː	2014	2025	2235	2217	2266	2308	1945	2309

TABLE IV. *Medians of the measured $F2_{vowel}$ values for speaker TB (all numbers in Hz). $F2_{vowel}$ values were measured at the midpoint of the vowels.*

		Labial		Alveolar		Velar		Palatal	
		b	p	d	t	g	k	ɟ	c
Back	ɔ	1056	1251	1114	1295	1095	1037	1197	1322
	o	797	978	875	1003	845	786	958	1036
	oː	651	691	675	720	661	633	674	703
	u	691	878	749	919	712	686	849	976
	uː	619	712	640	691	644	552	678	728
Front	aː	1478	1506	1593	1560	1564	1504	1527	1541
	ɛ	1678	1716	1798	1846	1795	1706	1678	1812
	ø	1433	1500	1475	1583	1477	1446	1500	1613
	øː	1659	1680	1600	1702	1621	1602	1703	1663
	y	1803	1904	1740	1909	1782	1824	1975	1881
	yː	1953	2002	1824	1927	1849	1878	1911	1848
	eː	2278	2302	2288	2306	2300	2308	2287	2296
	i	2209	2281	2300	2255	2258	2240	2274	2190
	iː	2317	2380	2409	2334	2357	2312	2358	2358

D. Formant measurements

Sound boundaries of the "aCVba" utterances were labelled automatically using a Hungarian speech recognition engine (Mihajlik *et al.*, 2002) in forced aligned mode. Second formant frequencies were measured automatically in Praat (Boersma and Weenink, 2008) at the burst of the stop consonant and at the midpoint of the second vowel. Doubtful F2 values were hand-corrected.

Tables III and IV show the medians of the measured formant frequencies for the burst of the stop consonants and the midpoint of the vowels, respectively.

III. RESULTS

A. Locus equation space

From speaker TB's measured $F2$ and SGR frequencies a locus equation space was drawn (Fig. 2). As the figure shows, the locus equation space is separated into dis-

tinct regions by the subglottal resonances. The $F2_{burst}$ - $F2_{vowel}$ pairs can be clustered according to the place of articulation of the consonants and vowels.

In the figure, six regions are indicated by numbers, each rectangle (surrounded by SGRs) corresponds to a different group of CV transitions:

1. Labial and velar consonants with back vowels
2. Alveolar and palatal consonants with back vowels
3. Alveolar, labial and velar consonants with front vowels, except [i, i:, e:]
4. Alveolar and labial consonants with unrounded non-low front vowels ([i, i:, e:])
5. Palatal consonants with front vowels, except [i, i:, e:]
6. Palatal and velar consonants with unrounded non-low front vowels ([i, i:, e:])

These regions differ partly from the ones reported for American English (Chen and Lulich, 2009). It seems that in English, velars before all front vowels have $F2_{burst}$ higher than $Sg3$. In this Hungarian experiment, velars before front vowels, except [i, i:, e:] are below $Sg3$ and only velars before [i, i:, e:] have $F2_{burst}$ higher than $Sg3$. Palatal consonants were not reported in Chen and Lulich (2009).

It seems that SGRs classify the locus equation space well into distinct regions. However, some smaller CV groups spread over these boundaries. Palatal consonants followed by back vowels occupy a large space in $F2_{burst}$ direction, some of them have $F2_{burst}$ values higher than $Sg3$. A well-defineable group of palatals lies in the crossing of vertical $Sg2$ and horizontal $Sg3$ (with $F2_{vowel}$ higher than $Sg2$). Most of the palatal consonants followed by front vowels have $F2_{burst}$ values higher than $Sg3$, whereas approximately a third of the "palatal-front, except [i, i:, e:]" CV transitions are below $Sg3$. Labial consonants followed by vowels [i, i:, e:] are mainly in region 4, but some parts extend to region 6. A few extreme cases of these are visible in the figure, with the highest $F2_{burst}$ values over all CV transitions. Alveolar, labial and velar consonants followed by unrounded non-low front vowels are distributed across regions 4 and 6.

B. Locus equations of various consonants

Linear regressions were estimated for the different groups of CV transitions (according to the place of articulation of the consonant). The equations and the Pearson's coefficient of regression are shown in Table V. The linear regressions show that the slopes (m) and y-intercepts (b) differ for the groups. Alveolars and palatals have slopes of about 0.3, while for labials and velars the steepness is closer to 1. Labials and velars have higher Pearcon's coefficient values, explaining their shape in Fig. 2. These CVs have rather linear relation of $F2_{burst}$ and $F2_{vowel}$.

TABLE V. *Linear regression coefficients and Pearson's coefficient of regression for the different clusters of the locus equation space.* $F2_{burst} = m \cdot F2_{vowel} + b$

	m	b	R^2
Alveolar	0.333	1184.35	0.768
Labial	0.732	301.22	0.915
Palatal	0.307	1552.82	0.628
Velar	0.912	179.195	0.936

C. Relation of SGRs

From the subglottal resonance measurements shown in Table II (speakers S1-S6), dependence of $Sg1$ and $Sg3$ on $Sg2$ were calculated, separetely. The relation between SGRs seems to be linear. Linear regression calculations resulted in the following equations:

$Sg1 = 0.226 \cdot Sg2 + 285.743$ ($R^2 = 0.631$)
$Sg3 = 1.266 \cdot Sg2 + 433.354$ ($R^2 = 0.964$).

D. Subglottal resonance estimation

The algorithm described in Chen and Lulich (2009) was used for calculating $Sg2$ from the CV dataset of speaker TB. The algorithm selects iteratively m, $m+1$, $m+2$, ... data points and calculates their first order locus equations (FOLE). After that, a second order locus equation (SOLE) is calculated, from which $Sg2$ can be derived. As m grows, the calculation approximates a value ($\widetilde{Sg2}$), and the deviation is smaller and smaller. The goal is to calibrate the algortihm so that the estimated $\widetilde{Sg2}$ value is close to the real $Sg2$.

Therefore, in our experiment the original method was modified. For the estimation of $\widetilde{Sg2}$ from the locus eqation space, only CV transitions with a labial or a velar consonant were used. Fig. 3 shows the calculated $\widetilde{Sg2}$ as the size of the subset (m, number of data points) grows. When m is small (2, 3 or 4), $\widetilde{Sg2}$ is much higher (at about 2000 Hz) than the real value. As m reaches 7, $\widetilde{Sg2}$ is very close to the measured $Sg2$ (approximately 1250 Hz). As m grows further, $\widetilde{Sg2}$ goes below the real $Sg2$ value and alternates in the region of 1050–1150 Hz.

For the calculation of $Sg1$ and $Sg3$ based on $Sg2$, the previously shown equations were used. When the size of the subset is greater than 10, this results in the estimated subglottal resonance frequencies of about 530 Hz, 1100 Hz and 1830 Hz (approximately 10% lower than the measured values).

E. Classification of CV sequences based on estimated SGRs

A modified version of the classifiction algorithm introduced in Chen and Lulich (2009) was used, in order to classify the CV transitions into different groups corresponding to their place of articulation. As discussed earlier, the locus equation space of the Hungarian speaker TB differs partly from the one reported for American English in Chen and Lulich (2009). Regions 1–6 shown in Fig. 2 were taken into consideration in the classification. The boundaries of these regions can be described in the form of inequalities between SGRs, $F2_{burst}$ and $F2_{vowel}$

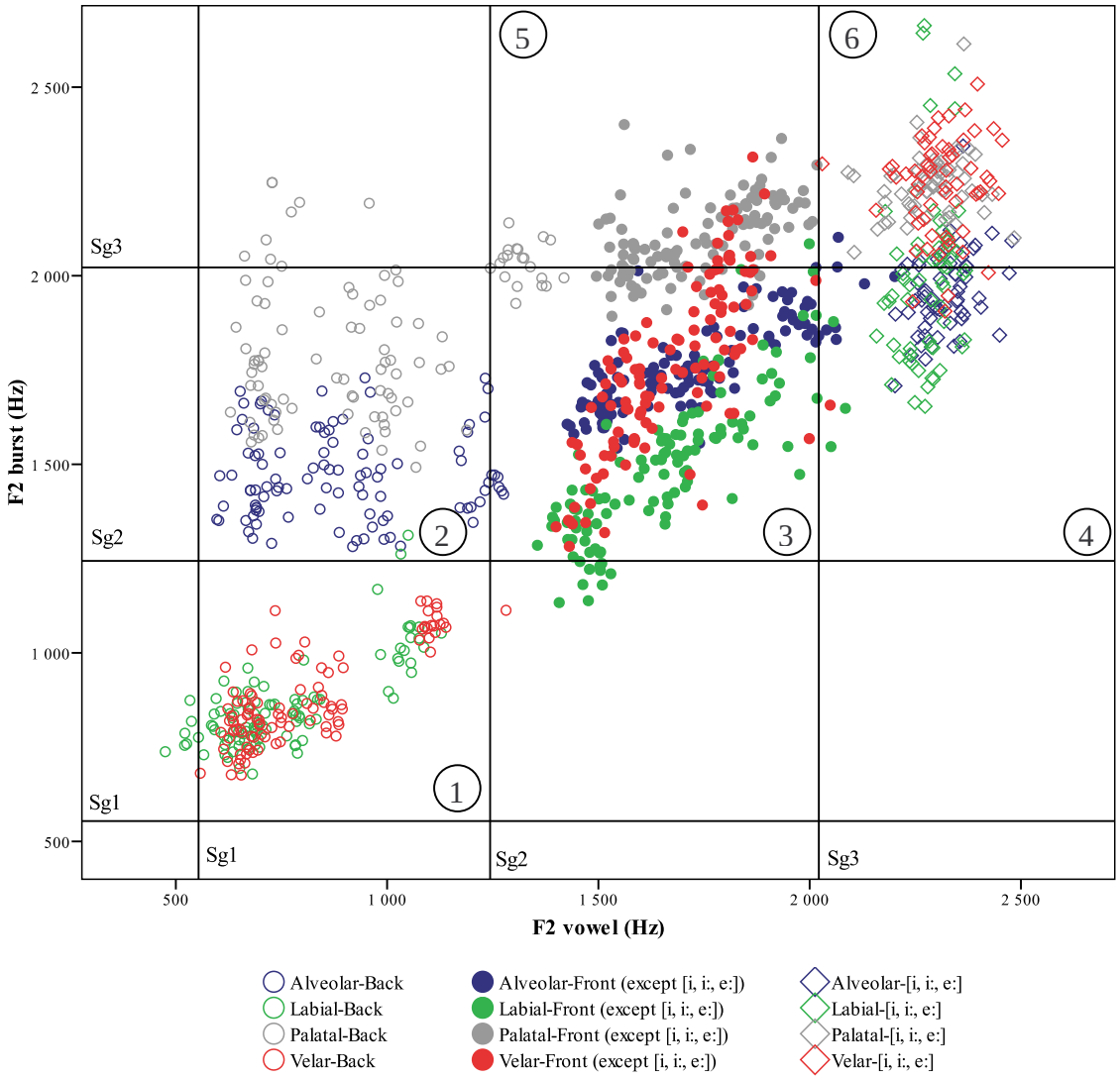


FIG. 2. Locus equation space of speaker TB. 1120 data points are shown, measured in CV transitions of nonsense words. Horizontal and vertical lines indicate subglottal resonances. Stop consonants with various place of articulation are denoted with different colors. Back, front (except [i, iː, eː]) and unrounded non-low front vowels ([i, iː, eː]) are denoted with different forms. The $F2_{onset}$-$F2_{vowel}$ pairs can be clustered according to the place of articulation of the consonants and vowels.

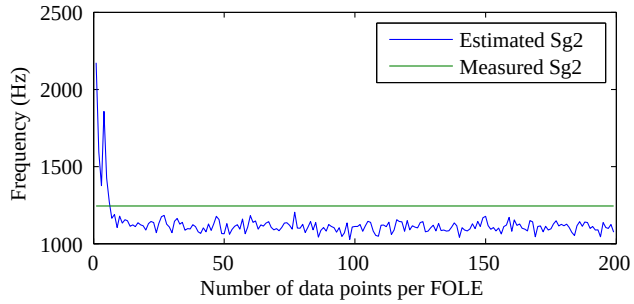


FIG. 3. Estimated value of the second subglottal resonance ($\widetilde{Sg2}$) as a function of the number of data points used in the calculation. $\widetilde{Sg2}$ reaches the measured $Sg2$, when the number of data points per FOLE is 7.

TABLE VI. *Inequalities describing the CV regions. (A stands for alveolar, L for labial, P for palatal and V for velar consonants. B stands for back, F for front except [i, iː, eː] and F* for front unrounded non-low vowels.)*

Region	CVs	Inequality 1	Inequality 2
1	LV-B	$Sg1 < F2_{burst} < Sg2$	$Sg1 < F2_{vowel} < Sg2$
2	AP-B	$Sg2 < F2_{burst} < Sg3$	$Sg1 < F2_{vowel} < Sg2$
3	ALV-F	$Sg2 < F2_{burst} < Sg3$	$Sg2 < F2_{vowel} < Sg3$
4	AL-F*	$Sg2 < F2_{burst} < Sg3$	$Sg3 < F2_{vowel}$
5	P-F	$Sg3 < F2_{burst}$	$Sg2 < F2_{vowel} < Sg3$
6	PV-F*	$Sg3 < F2_{burst}$	$Sg3 < F2_{vowel}$

values. Table VI shows the inequalities corresponding to each region.

TABLE VII. *Consonant hit rates and false alarm rates of a baseline classification based on measured SGRs. (C stands for all consonants summarized, A for alveolar, L for labial, P for palatal and V for velar consonants. B stands for back, F for front except [i, iː, eː] and F* for front unrounded non-low vowels.)*

	C	LV-B	AP-B	ALV-F	AL-F*	P-F	PV-F*
Hit rate	84	95	83	89	65	74	97
False alarm rate	2.4	0	0.2	5.5	1.3	2.9	4.3

TABLE VIII. *Vowel hit rates and false alarm rates of a baseline classification based on measured SGRs. (V stands for all vowels summarized.)*

	V	B	F	F*
Hit rate	97	93	97	100
False alarm rate	1.8	0	4.1	1.3

TABLE IX. *Consonant hit rates and false alarm rates of the classification based on estimated SGRs.*

	C	LV-B	AP-B	ALV-F	AL-F*	P-F	PV-F*
Hit rate	67	88	64	68	18	64	100
False alarm rate	5.3	0	0.5	3.8	2.8	5.7	19.3

TABLE X. *Vowel hit rates and false alarm rates of the classification based on estimated SGRs.*

	V	B	F	F*
Hit rate	87	85	75	100
False alarm rate	7.7	0	9.2	13.9

First, a baseline classification was carried out with speaker TB's measured subglottal resonances for the above-mentioned regions. The hit rates and false alarm rates are shown in Tables VII and VIII for consonants and vowels, respectively. For the consonants altogether, the hit rate is 84%, while the false alarm rate is 2.4%. For all of the vowels, the hit rate is 97% and the false alarm rate is 1.8%.

Our goal was to come close to the hit rates and false alarm rates of the baseline classification. The modified algorithm was run on the CV dataset of speaker TB. The algorithm runs with growing subset size, as described earlier. First it estimates the second subglottal resonance ($\widetilde{Sg}2$), after that it calculates the $\widetilde{Sg}1$ and $\widetilde{Sg}3$ on the basis of the linear relation between them. Using these estimated SGRs, a classification is carried out.

Fig. 4 shows the resulting hit rates and false alarm rates of the classification, showing the tendency while the subset size is growing. Hit rates are highest when the number of data points in each subset is small (m is about 7–8). This is the consequence of the $Sg2$ estimation method, Fig. 3 showed that $\widetilde{Sg}2$ was closest to the measured $Sg2$, when $m = 7$.

The overall hit rates and false alarm rates of the classification algorithm are shown in Tables IX and X (for subset size of 200). For the consonants, the overall hit rate is 67%. High hit rates were reached for most consonant types, except "AL-F*". This was caused by the overlapping dense data points in Regions 4 and 6. As the estimated SGRs are approximately 10% lower than the real values, Region 4 was moved lower in $F2_{burst}$ direction. The high false alarm rates of group "PV-F*" are due to the same shift. For the classification of the vowels, an overall hit rate of 87% was obtained. False alarm rates of unrounded non-low front vowels (F* group, [i, iː, eː]) are as high as 14%, because estimated SGRs shifted the boundaries of the classes lower in the $F2_{vowel}$ region.

IV. DISCUSSION AND CONCLUSIONS

In this paper, a series of experiments investigating locus equation space and the separating role of subglottal resonances were described. First, the locus equation space of the native speaker of Hungarian TB was examined, and six regions were defined according to the place of articulation of the consonant-vowel transitions. The locus equations were calculated for the various consonant types, and velar and labial consonants were chosen for use in an $Sg2$ estimation algorithm. The relation of $Sg1$, $Sg2$ and $Sg3$ was investigated in order to calculate the SGRs from the estimated $\widetilde{Sg}2$. The algorithm described in Chen and Lulich (2009) was modified, and applied for the CV dataset. From the estimated SGRs consonant and vowel classifications were run. These resulted in 10-15% lower hit rates than a baseline classification, in which the measured subglottal resonance frequencies of the same speaker were used. The results of the classification are similar to the one reported in Chen and Lulich (2009) for American English.

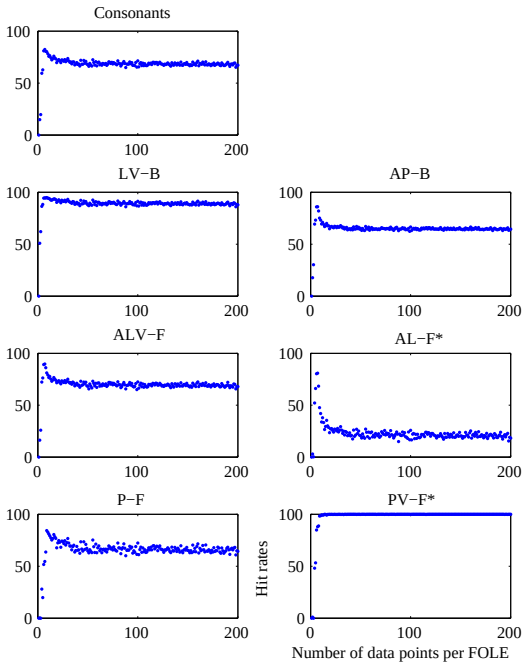
Future work includes the improvement of the $Sg2$ estimation algorithm. In this experiment, the estimated SGRs were approximately 10% lower than the measured SGRs, while in Chen and Lulich (2009) higher estimations were reported than the real subglottal resonance frequencies. If estimated SGRs were closer to the real ones, higher hit rates and lower false alarm rates could be reached in the classification of the CV sequences.

In this work only the data of one speaker was analyzed. A further experiment in this topic should investigate the classification of CV transitions with more speakers. It should be examined, if subglottal resonances play a similar separating role in the locus equation space of other speakers.

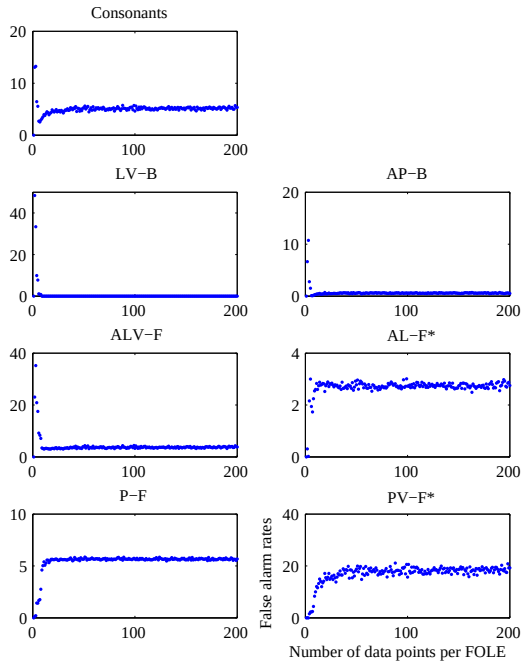
Acknowledgements

We would like to thank Steven M. Lulich for handing out this task. His lectures and explanations related to subglottal resonances helped a lot in understanding the role of SGRs, and contributed to the preparation of this final project.

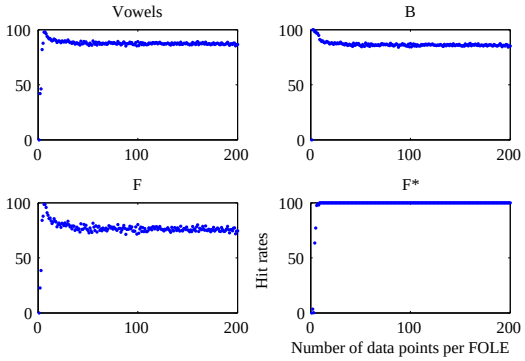
We also thank Tamás Bőhm for recording the CV dataset, and speakers S1-S6 for their accelerometer recordings.



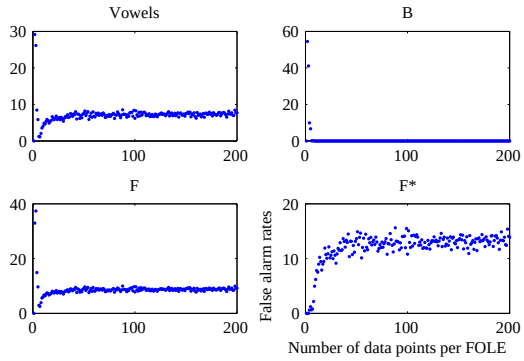
(a) Consonant hit rates.



(b) Consonant false alarm rates.



(c) Vowel hit rates.



(d) Vowel false alarm rates.

FIG. 4. Hit rates and false alarm rates for consonants and vowels. (A stands for alveolar, L for labial, P for palatal and V for velar consonants. B stands for back, F for front except [i, iː, eː] and F* for front unrounded non-low vowels.)

References

Boersma, P. and Weenink, D. (**2008**). *Praat [Computer program] (Version 5.0.20)*, URL `http://www.praat.org`, accessed on 11 Dec 2008.

Chen, N. F. and Lulich, S. M. (**2009**). "Context-free classification of consonant-vowel transitions based on subglottal resonances and the second formant", JASA **125**.

Chi, X. and Sonderegger, M. (**2007**). "Subglottal coupling and its influence on vowel formants", JASA **122**, 1735–1745.

Csapó, T. G., Bárkányi, Z., Gráczi, T. E., Bőhm, T., and Lulich, S. M. (**2009**). "Relation of formants and subglottal resonances in Hungarian vowels", in *Proceedings of Interspeech*, (in press).

Lulich, S. M. (**2008**). "On the relation between locus equations and subglottal resonances", JASA **124**, 2558.

Lulich, S. M. (**2009**). "Subglottal resonances and distinctive features", Journal of Phonetics Doi:10.1016/j.wocn.2008.10.006.

Mihajlik, P., Révész, T., and Tatai, P. (**2002**). "Phonetic transcription in automatic speech recognition", Acta Linguistica Hungarica **49**, 407–425.

Sjőlander, K. and Beskow, J. (**2009**). *Wavesurfer [Computer program] (Version 1.8.5)*, URL `http://www.speech.kth.se/wavesurfer`, accessed on 3 Apr 2009.

Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA).

Búcsú Csapó Tamás Gábortól

(Róma 12,15 „Örüljetek az örülőkkel, és sírjatok a sírókkal.")

Németh Géza

BME TMIT

Smartlabs

Csapó Tamás Gáborral 2006 tavaszán ismerkedtem meg, amikor az akkor még 5 éves mérnök informatikus képzés 6. félévében a teljes évfolyam számára tartottuk a Beszédinformációs rendszerek tárgyat. Jeles eredménnyel vizsgázott és megkeresett azzal, hogy érdekli a téma és szívesen foglalkozna vele. Ősszel IV. évesen már TDK dolgozatot adott elő a prozódiai változatosság gépi megvalósításának témakörében, amivel rögtön I. díjat nyert a BME VIK-en. A 2007-es tavaszi informatikai OTDK-n pedig szintén I. helyezett lett ezzel a dolgozattal. Ez az előadás ma is elérhető a sajnos már befagyasztott honlapján (https://speechlab.tmit.bme.hu/csapo/downloads/CsapoTG-otdk2007-eloadas.pdf). Személyesen nem tudtam ott lenni az előadásán, de ma is ott cseng a fülemben az öröme, mikor az eredményről telefonon beszámolt. Szerényen fogadta, hogy első tudományos munkája alapján rögtön elismerték tehetségét. A témáról már 2007 nyarán megjelent az első közös közleményünk az Interspeech konferencián, ami a szakmánk egyik kiemelkedő rendezvénye.

18 év alatt 183 publikáció (374 független hivatkozás) mellé került a neve az mtmt-ben, ami elképesztő teljesítmény. 39 éves korára (túl)teljesítette az MTA doktora cím minimum-követelményeit (http://www.hit.bme.hu/ ghorvath/tudometer/mtoscoring/10041912).

2008-ban Fék Márkkal közösen voltunk sikeresen megvédett diplomaterve konzulensei. A beszéd prozódiai változatosságának kihívása a beszédtechnológiában azóta sem megoldott, napjainkban is kutatás tárgya. 2008 őszén kezdte meg PhD hallgatóként tanulmányait a BME TMIT-en. Közben bekapcsolódott az oktatásba és az alapkutatástól az alkalmazásokig terjedő projektjeinkbe is. Hamarosan csapatunk oszlopos tagjává vált.

11 BSc és MSc szakdolgozat (3 külföldi hallgató), és az egyik első BME

Stipendium Hungaricum-os PhD védés konzulense volt. Mohammed Al-Radhi ma már egyik stabil munkatársunkká vált. Tamás 2024-ben három PhD hallgató témavezetője volt.

Magyar és angol nyelvű oktatási anyagokat dolgozott ki és adott elő (Beszédinformációs rendszerek, Ember-gép interfész/Human-Computer Interaction, Infocommunication, Szoftver laboratórium, Deep learning a gyakorlatban Python és LUA alapon a BME-n, Beszédtudományok az ELTE-n) amellett, hogy jellemzően a kutatás érdekelte, többszöri javaslatom ellenére sem indult el az oktatói ranglétrán.

Az első kutatási projekt, amibe már elsőéves doktoranduszként (2008-ban) bekapcsolódott, a TELEAUTO Jedlik kiemelt országos program volt, autós információs rendszerek létrehozása céljából. Azóta mintegy 20 kutatási kihívásban volt társunk. Ezek közül 10-et kompetitív EU-s pályázatok keretében nyertünk el. A pályázatok megírásában és a feladatok megoldásában is kulcsszerepe volt.

Már hallgatóként is érdekelték a nemzetközi kapcsolatok. Doktorandusz hallgatóként Párizsba és Finnországba utazott nyári egyetemekre, ahonnan mindig lelkesen, új kapcsolatokról beszámolva tért haza. Érdekes folyamat volt, ahogy Gósy Mária az 1990-es évek elején Fulbright ösztöndíjat kapott Ken Stevens professzorhoz az MIT-re, és ő ajánlotta Bőhm Tamás kollégánkat, aki 2004-2005-ben lehetett Ken Stevens egyik utolsó PhD hallgatója. Az ő hallgatótársa volt Steven Lulich, aki az Acoustic Phonetics tárgy egyik oktatója is volt. Steven Lulich az MIT PhD fokozat után egy másik amerikai egyetemre került és felajánlotta, hogy távoktatásban leadja a mi hallgatóinknak az Acoustic Phonetics kurzust. Ennek volt egyik hallgatója Csapó Tamás, aki ezen az úton ismerte meg Steven Lulich-t, aki hamarosan új labor létrehozására kapott lehetőséget az Indiana University-n és meghívta oda Tamást. 2014-ben Tamás Fulbright ösztöndíjasként egy fél évet töltött családjával ott, és motivációt kapott az artikuláció ultrahangos vizsgálatára.

Hazatérve megvédte PhD disszertációját, majd egyik kezdeményezője volt az ELTE-n a BME közreműködésével, Markó Alexandra vezetésével megvalósuló Lingvális Artikuláció Kutatócsoportnak (MTA-Lendület 2016-21). 2017-ben nyerte el első OTKA pályázatát Artikulációs mozgás alapú beszédgenerálás témakörben (2017-22). 2022-ben újabb OTKA támogatást nyert el Az artikuláció és az agyi jelek elemzése beszéd alapú agy-gép interfészhez témában (2022-26). Szoros kapcsolatot épített ki az SZTE munkatársaival is. Meghatározó szerepet játszott hazai (pl. Mesterséges Intelligencia Nemzeti Labor és Infokommunikációs Nemzeti Labor) és nemzetközi (pl. H2020, AAL, Horizon Europe) pályáza-

taink elnyerésében és megvalósításában. Kreatív módon működött közre ipari alkalmazásaink kidolgozásában is. Tavaly karácsonykor még arra biztattam, hogy kezdje el előkészíteni a habilitációs és az MTA doktori dolgozatát.

Tamás nemcsak tudós informatikusként, de közösség és kapcsolati hálózat építőként is kiváló ember volt. Nyitott, nyugodt és barátságos természete, mélyen istenhívő gondolkodása megkönnyítette a kapcsolódást hozzá. Örömmel vettünk részt 2010-ben esküvőjükön, majd érdeklődéssel követtük a család gyarapodását a négy gyermekkel. A COVID ideje alatt családjával vidéken kezdett új életet. Jó volt hallani lelkes beszámolóit a ház felújításáról. Rendkívüli módon odafigyelt a társaira. Például az az Ő javaslatára rendeztük meg 2023-ban a meglepetés ünnepséget Olaszy Gábor 80. születésnapjára a pályatársak részvételével, mintegy 40 fővel. Azon a nyáron pedig Moonshine néven kisebb nemzetközi konferenciát is szervezett a falujába. Részt vett az 2023. szeptemberében indult ENFIELD kiválósági hálózat munkájában is. 2024.január 25-én még egy kiváló kutatási tervet juttatott el hozzám.

Derült égből villámcsapásként ért a hír, hogy 2024. január 31-én véget ért földi útja és a mennyei hazába költözött. Sem a közelebbi, sem a távolabbi munkatársainknak nem volt tudomása arról, hogy milyen lelki terheket hordoz Tamás. Örök titok marad, hogy mi vezetett idáig. Tanulságként velünk marad, hogy igyekezzünk figyelni egymásra, támogatni a körülöttünk levőket, merjünk segítséget kérni a nehézségek idején . Felesége és gyermekei számíthatnak szolidaritásunkra és támogatásunkra.

2024. november 22.

A BME TMIT és a Smartlabs nevében Németh Géza egyetemi tanár,
laborvezető

Személyes szösszenet, avagy emlékdarabkák Csapó Tamás Gáborról

Juhász Kornélia

HUN–REN Nyelvtudományi Kutatóközpont
Eötvös Loránd Tudományegyetem

Tamást 2019. június 5-én ismertem meg a doktori felvételi elbeszélgetésemen, ahol mindössze csak annyit tudtam meg róla, hogy a Lendület LingArt Kutatócsoport tagja. Pár hónappal később e kutatócsoport munkájába bekapcsolódva volt lehetőségem jobban megismerni őt, és egyik első konkrét emlékem vele kapcsolatban az volt, hogy egy megbeszélésen milyen nagy örömmel és lelkesen ropogtatta az uzsonnára szánt, gondosan darabokba vágott sárgarépaszeleteket. Már akkoriban feltűnt, hogy milyen apró dolgokkal fel lehet dobni a napját és milyen pozitívan viszonyul az élet minden kis történéséhez és megtalálja a dolgok szórakoztató oldalát. Már az ismeretségünk legelején nyilvánvaló volt számomra, hogy Tamás egy rendkívül segítőkész és végtelenül kedves és alapos ember, akitől soha nem hallottam egy ideges vagy csalódott szót sem. Még a Lendület pályázatához Painttel készített ábráimat is nagy örömmel szemlélte, pedig azok a kezdetleges rajzok egyáltalán nem voltak sem profik, sem esztétikusak. Éppen az ábrákhoz kapcsolódó levelezésünkben írt egyszer nekem arról – ami számomra meghatározó motívum maradt vele kapcsolatban – hogy egyre fontosabbnak tartja a pozitív visszacsatolást a diákokkal/munkatársakkal való kommunikációban és ezután mindenkit igyekszik majd még többször és jobban megdicsérni a munkájáért. Úgy hiszem, ehhez az elhatározásához szigorúan tartotta is magát mindvégig: visszaolvasva az említett levelezést, majdnem minden üzenetében pozitív, dicsérő sorokkal üdvözölt, függetlenül attól, hogy hogyan és milyen mértékben tudtam hozzájárulni a kutatási terv adott részéhez. Tamás különleges egyénisége szinte minden tevékenységében megmutatkozott: még a konferenciák szürkeségébe is tudott színkavalkádot vinni, mint például amikor a Beszédkutatás konferencia szünetében a konferencia résztvevői kipróbálhatták a

hangszórókra csatlakoztatott elektro-akusztikus kalimbáját. Kreativitása és az apró részletekig menő gondossága nem csak a protokollok készítésében, hanem előadásaiban is megmutatkozott: számára a Covid-19 járvány és a veszélyhelyzet nem gátló vagy demotiváló tényezőként jelent meg az életében, inkább lehetőséget nyújtott arra, hogy az online előadásformátumból kihozza a legtöbbet: on-the-spot szólalt meg rajzfilmhősök hangján, illusztrálva azt, hogy napjainkban mennyi lehetőségünk van az interneten megtalálható ingyenes szoftverekkel pillanatok alatt eltorzítani a valódi beszédünket. Ismeretségünk öt éve alatt mindig csodálattal figyeltem a szakmai lelkesedését és kitartását, ami abban mutatkozott meg, hogy szinte lehetetlen volt olyan ötlettel előállni, ami iránt nem érdeklődött volna és nem lehetett olyan konferenciaszervezést vagy kísérletet kitalálni, amit nem támogatott volna. A kutatócsoportban elképedve hallgattuk a történeteit arról, hogy hogyan tud helytállni az élet minden frontján a mindennapokban, hogyan gondoskodik a családjáról, játszik a gyerkőcökkel, zenél, nyelveket tanul, felújít, kertészkedik és nem utolsó sorban rengeteget dolgozik. Tamás kedvessége, nagylelkű segítőkészsége és vidám természete sok hallgatóban és kollégában marad emlékül, remélve, hogy ha eljön az ideje, egyszer majd újra találkozunk.

Words are as important as action
In the memory of late Csapó Tamás Gábor

Ibrahim Ibrahimov

BME Department of Telecommunications and Artificial Intelligence

During the decision-making process of my master's thesis topic in 2022 at the University of Szeged, I had a great chance to get to know Tamás after becoming interested in the Silent Speech Interface topic during my talk with Gosztolya Gábor. (It is the right time to thank Gábor for connecting me with Tamás back then.) For a semester, we had only online meetings, yet even then, he was very enthusiastic about educating me on the topic and ensuring that I had a great thesis topic. His passion and enthusiasm for research initially caught my attention, leading me to continue my academic path in the topic we worked on. Eventually, I started my PhD journey under his supervision at BME. I am very grateful and lucky to have him as my supervisor, as I had the greatest chance to learn how to learn, implement practically, and connect with other colleagues in the field. I remember our meetings in the very early morning (around 7 a.m.) to discuss and push forward toward deadlines together. This is the kind of support every student wishes to have from their professors or mentors, so I am thankful that our life paths crossed in this way. He would always listen and make his points clear to me, never growing tired of explaining. I was inspired by his passion for the topic and research in general. You could easily tell how 'hungry' he was for new information and new experiments. Basically, he showed me and made sure I understood that nothing is impossible in this field if you give it a try. His eyes were always bright and wide open with full attention during our conversations. His ability to be at ease with himself was obvious and gave me peace whenever I was around him or wanted to discuss any idea I had. He was always open to discussions and never showed disinterest in anything I proposed.

He was the professor you could talk to not only about research projects but also about life struggles. His support in all dimensions of life was with me

throughout the time I knew him. He never showed any personal or research-related struggles. He was always optimistic and encouraging, even when things were heading in an unpromising direction. Knowing that I could call him and talk about my struggles made me feel calm about my research path, as I could feel his hands on my shoulders. His effort to understand the person in front of him was a spark for me to become someone like him. At the core of his being was humanity ahead of everything else. He would first talk to me as a fellow human being and then as a professor or colleague. Our calls in his garden, with his kids around, are unforgettable memories that continue to inspire me to push forward, even in difficult moments.

Everything aside, he was a friend, a father I never had, a brother everyone wishes to have, and most importantly, a great person with a truly warm heart and mind. I always tried to express my gratitude to him in our conversations, but I will make sure that his legacy stays with me and is reflected in any work I do. He is the greatest reason why I have the passion and enthusiasm I do for this field today.

Lastly, I want any reader of this article to know that they are never alone. Please make sure to surround yourself with people you can talk to and have open communication with, without any judgment. Tamás was that person for me, and I will never forget this fact.

May his soul rest in peace!