The 5th International Symposium on Applied Phonetics
ISAPh 2024

Book of Abstracts

University of Tartu, Estonia
30 September – 2 October 2024

The 5th International Symposium on Applied Phonetics (30 September – 2 October 2024) is organized at the University of Tartu, Estonia. The conference series is dedicated to all kinds of applications of phonetics for human activities and focuses on a wide range of areas of applied phonetics. The conference follows the four previous ISAPh conferences held in Nagoya (2016), Aizuwakamatsu (2018), Tarragona (2021), and Lund (2022).

Organising Committee: Pärtel Lippus (chair), Eva Liina Asu, Katrin Leppik, Anton Malmi, Pire Teras, Helen Wilbur

Scientific Committee:
Adrian Leemann (University of Bern)
Benjamin V. Tucker (Northern Arizona University)
Bettina Braun (University of Konstanz)
Christian Jensen (University of Copenhagen)
Cristian Tejedor García (Radboud Universiteit)
David House (KTH Royal Institute of Technology)
Debbie Loakes (University of Melbourne)
Einar Meister (Tallinn University of Technology)
Elisabeth Zetterholm (Linköping University)
Francesco Cangemi (University of Cologne)
Francis Nolan (University of Cambridge)
Gilbert Ambrazaitis (Linnaeus University)
Heini Kallio (Tampere University)
Ian Wilson (Center for Language Research, University of Aizu)
Jaques Koreman (NTU Trondheim)
Joaquin Romero (Universitat Rovira i Virgili)
Juraj Šimko (University of Helsinki)
Katalin Mády (HUN-REN Hungarian Research Centre for Linguistics)
Lagle Lehes (University of Tartu)
Liisi Piits (Institute of Estonian Language)
Marilyn M. Vihman (University of York)
Marja-Liisa Mailend (University of Tartu)
Mechtild Tronnier (Lund University)
Michael Ashby (University College London)
Mikael Roll (Lund University)
Murray Munro (Simon Fraser University)
Oliver Niebuhr (University of Southern Denmark)
Patricia Ashby (University of Westminster)
Paul Foulkes (University of York)
Riika Ullakonoja (University of Jyväskylä)

Website: www.isaph2024.ut.ee
Contact: isaph2024@ut.ee
Cover picture by Ragnar Vutt

Table of contents

Invited speakers

Day 1 - oral presentations

Day 1 - posters

Day 2 - oral presentations

Day 2 - posters

Forever young:
No “critical period” for speech learning ability

Ocke-Schwen Bohn
Aarhus University, Denmark

***Keywords:* — nonnative speech learning, age factor in language learning, Critical Period Hypothesis**

I. Introduction

At first sight, it may appear as if the ability to produce and perceive the sounds of a nonnative language inevitably declines with age. This superficial impression has often been interpreted as providing support for the Critical Period Hypothesis (CPH), which claims that language learning ability is severely compromised after puberty because “the brain behaves as if it has become set in its ways” [1].

This presentation addresses three problems which any maturational account of assumed age difference in language learning (like the CPH) encounters: First, that “age” is a variable that may (or may not) be associated with biological and psycholinguistically relevant maturation or, more plausibly, with differences in the quality and quantity of linguistic input across the life span. Secondly, the presentation will review empirical findings which address the question of whether successful speech learning is restricted to a prepubescent window of opportunity. The conclusion from this review is that studies which seem to provide support for the CPH are often flawed either because of confounding variables (e.g., differences in language experience between young and older learners) or because of misleading data analyses and presentations. Third, this presentation will present recent evidence from our lab on the ability of seniors (60-78 years old) to restructure their sound systems in a perceptual training regime. Comparisons with younger trainees revealed no age differences in the efficacy of training (i.e., increase in perceptual accuracy) and no age differences in the training trajectory over 10 training sessions.

The presentation concludes that the existing evidence strongly supports one of the main tenets of current models of nonnative speech learning, the PAM-L2 [2] and the SLM-r [3], which claim that speech learning ability remains intact across the life span. As far as speech learning ability is concerned, time is ripe to recognize Critical Period Hypothesis for what it is: a neuromyth [4].

References

[1] E. H. Lenneberg, Biological foundations of language. New York: Wiley 1967.

[2] C. T. Best and M. Tyler. "Nonnative and second-language speech perception," in Language experience in second language speech learning, O.-S. Bohn and M. J. Munro, Eds. Amsterdam: J. Benjamins, 2007, pp. 13-34.

[3] J. E. Flege and O.-S. Bohn. "The revised speech learning model (SLM-r)," in Second language speech learning: Theoretical and empirical progress, R. Wayland, Ed. Cambridge: Cambridge Universitry Press, K Macdonald et al. 2021, pp. 3-83.

[4] K. Macdonald et al. “Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths,” Frontiers in Psychology 8, pp. 1314, 2017.

Navigating the search for ‘normal’ in children’s speech and language disorders

Sofia Strömbergsson
Karolinska Institutet, Sweden

Keywords — clinical phonetics, speech pathology, speech acquisition

I. Introduction

The number of correct consonants, or the number of correct grammatical inflections in children’s utterances, are examples of measures of speech and language competence in children. In speech-language pathology research, measures like these are central in quantifying speech and language difficulties, and in separating groups of children who have a speech/language disorder from those who follow a typical trajectory. And in clinical practice, measures like these form the basis of developmental milestones, to which the observed speech and language in a specific child are compared. Producing canonical babbling by 10 months [1], combining words by 24 months [2], and having acquired the majority of Swedish consonants by the age of 5 [3], are all examples of such milestones. In clinical practice, the comparison to norms is important when identifying risk for later difficulties, and deciding whether intervention is needed. In this talk, I will examine the potential conflict between measures used when identifying disorders, and what aspects of speech, language and communication actually matter in daily life for children with speech/language disorders.

II. Correct/incorrect speech

At a closer look, the boundary between ‘correct’ and ‘incorrect’ is rarely clear-cut. The assessment of whether a consonant is produced ‘correctly’ or not involves a reduction of phonetic detail that may convey different degrees of ‘correctness’ [4]. This reduction of detail may obscure important insights into a child’s phonological competence, as evidenced in observations of covert contrast [4], [5]. ‘Covert contrast’ refers to when a child expresses a measurable phonetic distinction between speech sounds, which goes unnoticed to the assessor – or which, at least, is not documented in transcription. In other words, the way we document articulation will affect how we characterize children’s speech.

Information concerning communication in daily life is an integral part of clinical assessment. For children with speech difficulties, clinicians routinely collect information from caregivers concerning how children make themselves understood in everyday contexts. A well-established instrument for such assessment is the Intelligibility in Context Scale (ICS) [6], which can be used to identify departures from expected development. However, closer inspection is needed for understanding when and why intelligibility is disrupted [7]. Assessments of intelligibility rely not only on information in the speech signal, but also on the assessor and their degree of training [8], their experience with the target language [9], and their familiarity with the speech material [10]. One may question, therefore, whether speech-language pathologists (SLPs) are indeed representative assessors of speech intelligibility. At least, one might acknowledge a need to calibrate SLP assessments against other listeners’ assessments.

When used as a descriptor of speech, ‘intelligibility’ presumably reflects how much an ‘average listener’ can decode from the speech signal. Given the variability between listeners, this requires input from a panel of listeners [11]. And to include not only expert listeners (such as SLPs) in assessments, task instructions need to be understood without prior training. Furthermore, evaluation methods that allow real-time collection of responses may be preferred over those that require, for example, listeners transcribing what they perceive. In our research, we have introduced Audience Response Systems (ARS)-based assessments as a window into listeners’ perception of disordered speech. The ARS-based method allows intuitive task instructions and real-time response collection, thus meeting at least some desirable methodological features. We have used the ARS-based method in assessments of voice [12], and in assessments of speech produced by children with speech sound disorders [13], [14], with listener panels including experts (SLPs), lay adults, and children. We found no systematic differences between experts and lay listeners in their evaluations of intelligibility; at least in this task, SLPs seem to be representative of other listeners [13].

The ARS-based method has also served as a window into listeners’ evaluation of acceptability, that is, the perceived “differentness” of children’s way of speaking. Just like reduced intelligibility, reduced acceptability is a threat to successful communication for children with speech disorders [14]. Compared to intelligibility, however, acceptability is considerably less studied. The ARS-based method allowed us to investigate acceptability and intelligibility in parallel, and to compare the two constructs with reference to the same continuous scale. As expected, listeners reacted more frequently to speech sounding ‘awkward’ than to not understanding the spoken message [14]. In terms of listener differences, children appeared less critical in their evaluations than the adult listeners [14].

To conclude, developmental norms of children’s articulation and intelligibility are important when identifying children in need of clinical intervention. However, documentation underlying norms often obscures detail. Also, norms depend on who gets to set the boundary between ‘correct’ and ‘incorrect’, and between ‘intelligible’ and ‘unintelligible’. As such, developmental norms should be handled with care.

III. Correct/incorrect language

Developmental Language Disorder (DLD) is common in children and can have long-term consequences for academic achievement and psychosocial well-being [15]. Identifying DLD during preschool years can therefore be important for mitigating an adverse developmental trajectory. For that purpose, comparing observed language in a child to developmental norms concerning for example vocabulary and grammar is important. But similarly to developmental norms concerning speech, norms concerning language depend on who sets the norms. Also, their insensitivity to detail makes existing norms ill-fit for tracking potential progress in children with limited verbal language, such as in children with DLD.

In our research group, we are exploring multimodal language-sample analysis as a way of tracking language development in children with severe DLD. For these children, intervention is often not limited to strengthening verbal language, but also aims to encourage communication via alternative means, like gestures, manual sign and/or pictures. I will present an insight into our ongoing work, as a case in point illustrating the value of sensitivity to more fine-grained aspects than ‘correct’/’incorrect’.

IV. Correct/incorrect – to whom?

Finally, I hope to encourage reflection concerning the seeming dissonance between the measures we use, and what actually matters for children with speech/language disorders and their families. When asked what their preferred outcomes of intervention are, children themselves rarely mention consonants or grammatical inflections, but rather aspects like having fun with friends, being listened to, and not being teased [16]. And parents raise aspects like social inclusion, friendship and independence [16]. To clinicians, these perspectives are probably very familiar, but as researchers, we might need to remind ourselves of the potential gap between the topics we study and what impact they have in the lives of children with speech/language disorders.

Re-evaluation of historical SLP practices and attitudes sheds light on cultural biases against marginalized groups [17], calling previously held truths about what is ‘normal’ and what is ‘abnormal’ into question. As active in the field of children’s speech/language disorders, relying on norms based on decisions of what is ‘correct’ and what is ‘incorrect’, we should at least expose ourselves to such perspectives, and ask ourselves whose perspective we represent and why.

References

[1] D. K. Oller, R. E. Eilers, A. R. Neal, and A. B. Cobo-Lewis, ‘Late onset canonical babbling: a possible early marker of abnormal development’, *Am J Ment Retard*, vol. 103, no. 3, pp. 249–263, Nov. 1998, doi: 10.1016/s0021-9924(99)00013-1.

[2] J. M. Rudolph and L. B. Leonard, ‘Early Language Milestones and Specific Language Impairment’, *Journal of Early Intervention*, vol. 38, no. 1, pp. 41–58, Mar. 2016, doi: 10.1177/1053815116633861.

[3] A. Lohmander, I. Lundeborg, and C. Persson, ‘SVANTE – The Swedish Articulation and Nasality Test – Normative data and a minimum standard set for cross-linguistic comparison’, *Clinical Linguistics & Phonetics*, vol. 31, no. 2, pp. 137–154, Feb. 2017, doi: 10.1080/02699206.2016.1205666.

[4] S. Strömbergsson, G. Salvi, and D. House, ‘Acoustic and perceptual evaluation of category goodness of /t/ and /k/ in typical and misarticulated children’s speech’, *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3422–3435, Jun. 2015, doi: 10.1121/1.4921033.

[5] B. Munson, J. Edwards, S. Schellinger, M. E. Beckman, and M. K. Meyer, ‘Deconstructing Phonetic Transcription: Covert Contrast, Perceptual Bias, and an Extraterrestrial View of Vox Humana’, *Clin Linguist Phon*, vol. 24, no. 4–5, pp. 245–260, Jan. 2010, doi: 10.3109/02699200903532524.

[6] S. McLeod, ‘Intelligibility in Context Scale: cross-linguistic use, validity, and reliability’, *Speech, Language and Hearing*, vol. 23, no. 1, pp. 9–16, Jan. 2020, doi: 10.1080/2050571X.2020.1718837.

[7] T. B. Lagerberg, E. Anrep-Nordin, H. Emanuelsson, and S. Strömbergsson, ‘Parent rating of intelligibility: A discussion of the construct validity of the Intelligibility in Context Scale (ICS) and normative data of the Swedish version of the ICS’, *International Journal of Language & Communication Disorders*, vol. 56, no. 4, pp. 873–886, Jul. 2021, doi: 10.1111/1460-6984.12634.

[8] D. O’Leary, A. Lee, C. O’Toole, and F. Gibbon, ‘Intelligibility in Down syndrome: Effect of measurement method and listener experience’, *International Journal of Language & Communication Disorders*, vol. 56, no. 3, pp. 501–511, 2021, doi: 10.1111/1460-6984.12602.

[9] T. B. Lagerberg, J. Lam, R. Olsson, Å. Abelin, and S. Strömbergsson, ‘Intelligibility of Children With Speech Sound Disorders Evaluated by Listeners With Swedish as a Second Language’, *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 10, pp. 3714–3727, Oct. 2019, doi: 10.1044/2019_JSLHR-S-18-0492.

[10] N. Miller, ‘Measuring up to speech intelligibility’, *International Journal of Language & Communication Disorders*, vol. 48, no. 6, pp. 601–612, 2013, doi: 10.1111/1460-6984.12061.

[11] T. B. Lagerberg, K. Holm, A. McAllister, and S. Strömbergsson, ‘Measuring intelligibility in spontaneous speech using syllables perceived as understood’, *Journal of Communication Disorders*, vol. 92, p. 106108, Jul. 2021, doi: 10.1016/j.jcomdis.2021.106108.

[12] K. Johansson, S. Strömbergsson, C. Robieux, and A. McAllister, ‘Perceptual Detection of Subtle Dysphonic Traits in Individuals with Cervical Spinal Cord Injury Using an Audience Response Systems Approach’, *Journal of Voice*, vol. 31, no. 1, p. 126.e7-126.e17, 2017, doi: 10.1016/j.jvoice.2015.12.015.

[13] S. Strömbergsson, K. Holm, J. Edlund, T. B. Lagerberg, and A. McAllister, ‘Audience Response System-Based Evaluation of Intelligibility of Children’s Connected Speech – Validity, Reliability and Listener Differences’, *Journal of Communication Disorders*, vol. 87, p. 106037, Sep. 2020, doi: 10.1016/j.jcomdis.2020.106037.

[14] S. Strömbergsson, J. Edlund, A. McAllister, and T. B. Lagerberg, ‘Understanding acceptability of disordered speech through Audience Response Systems-based evaluation’, *Speech Communication*, vol. 131, pp. 13–22, Jul. 2021, doi: 10.1016/j.specom.2021.05.005.

[15] D. V. M. Bishop, M. J. Snowling, P. A. Thompson, and T. Greenhalgh, ‘Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology’, *Journal of Child Psychology and Psychiatry*, vol. 58, no. 10, pp. 1068–1080, 2017, doi: 10.1111/jcpp.12721.

[16] S. Roulstone, J. Coad, A. Ayre, H. Hambly, and G. Lindsay, ‘The preferred outcomes of children with speech, language and communication needs and their parents’, Department for Education, London, UK, DFE-RR247-BCRP12, Dec. 2012. Accessed: Apr. 23, 2020. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/219625/DFE-RR247-BCRP12.pdf

[17] J. F. Duchan and L. E. Hewitt, ‘How the Charter Members of ASHA Responded to the Social and Political Circumstances of Their Time’, *American Journal of Speech-Language Pathology*, doi: 10.1044/2022_AJSLP-22-00273.

Analysis of disfluency behaviour for forensic phonetic applications

Kirsty McDougall
University of Cambridge, England

Keywords — disfluency, disfluency features, fluency, forensic speaker comparison, TOFFA

Perturbations of the flow of a speaker’s speech, such as filled and silent pauses, repetitions, self-interruptions and sound prolongations, occur with relatively high prevalence in the speech of people who stutter (‘***dys***fluency’). They are also present to differing extents in the spontaneous speech of normally-fluent speakers (‘***dis***fluency’) (e.g. [7]), yet the speaker-specificity of such features has received little attention in phonetic research. Aspects of speech such as filled and silent pausing may play a part in the planning of speech (e.g. [2]) and therefore may be influenced by psycho- or socio-linguistic demands, thus they have strong potential for individual variation. Other breaks in fluency such as repetition, prolongation or self-interruptions might also function as part of the speech planning process and be difficult to control consciously, thus offering a further source of individual variation.

Analysis of individuals’ use of disfluencies has great potential for application in forensic speaker comparison cases, in which voice recordings each of an unknown speaker committing a crime and of a suspect are compared with a view to assessing the likelihood that the same speaker is on both recordings. The bulk of the literature investigating speaker-distinguishing properties of phonetic features for forensic applications has focussed on variables which bear a direct relationship with a speaker’s anatomy, for example fundamental frequency which reflects the length and mass of a speaker’s vocal folds, or formant frequencies which reflect the dimensions and configuration of the vocal tract (see e.g. [1]). Investigating the speaker-distinguishing potential of disfluency features focusses on a very different aspect of a speaker’s performance: speech features which are behavioural rather than anatomical. As well as the phonetic theoretical reasons for investigating the speaker-specificity of disfluency features, these features are largely realised through the temporal domain and therefore generally well-preserved in the poor recording conditions of forensic cases where background noise and telephone transmission (with its reduced bandwidth) are typical. This is in contrast to the ‘anatomical’ features mentioned above which are conveyed through spectral (frequency) information for which adverse recording conditions are more problematic.

This talk will present findings from an ongoing programme of research by McDougall and Duckworth into individual variation in fluency behaviour and its application in forensic speaker comparison casework. The TOFFA framework ‘Taxonomy of Fluency features for Forensic Analysis’ devised by McDougall and Duckworth [3, 4] for quantifying individual differences in disfluency will be outlined and results from studies applying TOFFA to a number of forensically relevant datasets will be presented, considering the effects of important factors such as speaking style and (lack of) contemporaneity of recording session, as well as variation across different accents of a language (e.g. [5]).

The talk will also illustrate the application of TOFFA to forensic casework practice, using a number of example cases where analysis of disfluencies was of key importance. These cases come out of collaborative work conducted with J.P. French Associates, United Kingdom, a forensic phonetic consultancy where the TOFFA framework has been applied to characterize disfluency usage in forensic speaker comparison cases for a number of years [6]. Ongoing practical issues and directions for further research will be outlined.

The talk will conclude that when it can be implemented, systematic disfluency analysis is a valuable tool in the forensic phonetician’s toolkit, and one which complements other types of analysis well.

ACKNOWLEDGEMENTS

The author is grateful to her collaborators Peter French, Toby Hudson, Christin Kirchhübel, Debbie Loakes, Alice Paver, Richard Rhodes who have all contributed to this work in various ways, and especially to Martin Duckworth for his inspiration to start this line of research and his close involvement with this research from its inception many years ago to current developments.

REFERENCES

[1] P. Foulkes and P. French, “'Application of the 'TOFFA' framework to the analysis of disfluencies in forensic phonetic casework,” in The Oxford Handbook of Language and Law, P. M. Tiersma and L. M. Solan, Eds. Oxford: Oxford University Press, 2012, pp. 557-572.

[2] F. Goldman-Eisler, Psycholinguistics: Experiments in Spontaneous Speech. London: Academic Press, 1968.

[3] K. McDougall and M. Duckworth, “Profiling fluency: an analysis of individual variation in disfluencies in adult males,” Speech Communication, vol. 95, pp. 16-27, 2017.

[4] K. McDougall and M. Duckworth, “Individual patterns of disfluency across speaking styles: a forensic phonetic investigation of Standard Southern British English,” International Journal of Speech Language and the Law, vol. 25(2), pp. 205-30, 2018.

[5] K. McDougall, M. Duckworth and T. Hudson, “Individual and group variation in disfluency features: a cross-accent investigation,” Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, Australia, Paper number 0308, August 2015.

[6] K. McDougall, R. Rhodes, M. Duckworth, P. French, and C. Kirchhübel, “Application of the ‘TOFFA’ framework to the analysis of disfluencies in forensic phonetic casework,” Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia, pp. 731-735, August 2019.

[7] E. Shriberg, “To ‘errrr’ is human: ecology and acoustics of speech disfluencies,” Journal of the International Phonetic Association, vol. 31, pp. 153-69, 2001.

Accounting for phonetic patterns in the expression of sex and gender in prepubertal and adult voices

Adrian P. Simpson
Friedrich Schiller University Jena, Germany

***Keywords* — gender, gender role, gender identity, sociophonetics, prepubertal voice**

When asked to identify the gender of stimuli from a randomly chosen sample of adult speakers, listeners make judgments that agree almost 100% with the self-reported gender of the speakers. Many of the acoustic correlates responsible for these systematic and consistent judgments can initially be attributed to sex-specific biological differences that arise during puberty. Longer and thicker male vocal folds produce an average fundamental frequency that is typically half that of the female value [1]. Disproportionate lowering of the larynx produces a longer male vocal tract giving rise to formant frequencies that are lower than those produced in an average female vocal tract. Despite the acoustic consequences of these average biological differences, it is also clear that the magnitude and form of the acoustic differences are in part attributable to socio-culturally acquired patterns. This is apparent from intercultural differences in the magnitude and non-uniformity of gender-specific differences [2, 3, 4]. Likewise, long-term studies have found marked reductions in female fundamental frequency over an interval of several decades, indicating changes in voice accompanying changes in gender role [5]. However, the most intriguing example of learnt gender-specific patterns is the vocal expression of gender in prepubertal voices. Any anatomical differences prior to the onset of puberty are negligible [6]. Nevertheless, gender identification of prepubertal sentence-length stimuli is still above-chance, typically at around 70%. However, this figure belies a more complex picture, in which the gender identification of some speakers remains at chance level, while for others listeners' ratings approach those found for adult speakers [7]. This suggests that some children are producing a consistent and robust set of acoustic correlates that listeners use to decode a child's gender.

This talk will examine the difficulty of teasing apart nature and nurture when accounting for gender-specific differences in adult and prepubertal voices. We will consider the importance of providing a differentiated picture of a speaker's gender, gender identity and gender role [8, 9]. Finally, we will look at ways of identifying the acoustic correlates that produce systematic gender ratings in children's voices.

References

[1] Stevens, Kenneth N. 1998. Acoustic Phonetics. Massachusetts: M.I.T. Press.

[2] Fant, Gunnar. 1975. Non-uniform vowel normalization. STL-QPSR 2–3, 1–19.

[3] Henton, Caroline G. 1995. Cross-language variation in the vowels of female and male speakers. In Proc. XIIIth ICPhS, vol. 4. Stockholm, 420–423.

[4] Traunmüller, Hartmut & Anders Eriksson. 1993. The frequency range of the voice fundamental in the speech of male and female adults. Unpublished ms. Stockholm.

[5] Pemberton, Cecilia, Paul McCormack, & Alison Russell. 1998. Have women's voices lowered across time? A cross sectional study of Australian women's voices. Journal of Voice 12(2), 208–213.

[6] Fitch, W. Tecumseh & Jay Giedd. 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. Journal of the Acoustical Society of America 106(3), 1511–1522.

[7] Funk, Riccarda & Adrian P. Simpson. 2023. The acoustic and perceptual correlates of gender in children's voices. Journal of Speech, Language and Hearing Research, 1–18.

[8] Weirich, Melanie & Adrian P. Simpson. 2018. Gender identity is indexed and perceived in speech. PLoS ONE 13(12), e0209226.

[9] Weirich, Melanie & Adrian P. Simpson. 2019. Effects of gender, parental role and time on infant- and adult-directed read and spontaneous speech. Journal of Speech, Language and Hearing Research 62(11), 4001–4014.

Ultrasound tongue imaging for phonetic research

Sam Kirkham, Claire Nance

Lancaster University, England

Keywords — ultrasound, articulatory phonetics, tongues, imaging, data analysis

I. INTRODUCTION

In this workshop we will introduce participants to using ultrasound tongue imaging for phonetic research. Ultrasound provides a (relatively) easy method of viewing the tongue in two dimensions for articulatory research. Most commonly, video is recorded of the tongue in midsagittal position (Fig. 1). The best image is obtained from the surface of the tongue and typically researchers extract the coordinates of the midsagittal tongue surface for comparison within and across speakers. Typically, the image is oriented such that the tongue root is on the left and the tongue tip on the right.

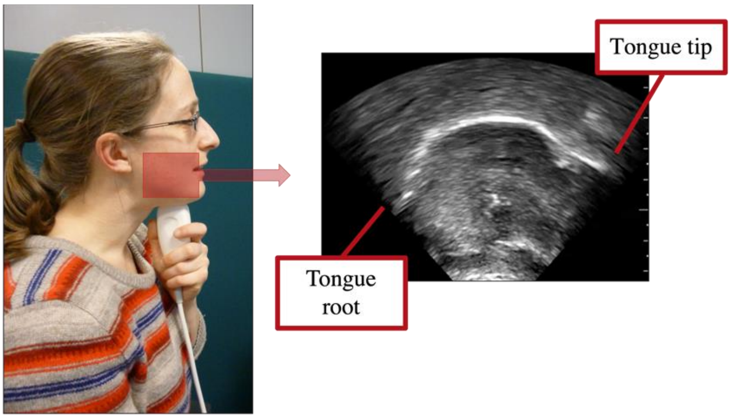


Fig. 1. Example of the view typically obtained from ultrasound tongue imaging in phonetic research.

We will work with workshop participants on two ultrasound machines to demonstrate the equipment and explain a basic workflow which could be used for a small research project, developing teaching resources, or in public engagement work. First, we will discuss what kind of research questions ultrasound analysis can answer and give some examples from when we have used ultrasound in teaching, public engagement, and research. Then, we will introduce participants to the software now most commonly used for recording ultrasound data, Articulate Assistant Advanced (AAA) [1]. We will explain how to obtain the best images for research purposes and workshop participants will make some recordings in AAA. To explain the initial stages of data analysis, we will demonstrate how to label audio data recorded in AAA, fit splines to ultrasound tongue images, and export spline coordinates for further analysis. Finally, we will show workshop participants how to extract videos from ultrasound data for teaching, demonstrations, and knowledge exchange.

II. WHY OR WHY NOT USE ULTRASOUND?

Ultrasound is a practical and relatively cheap option for collecting articulatory phonetic data. The probe is placed under a participant’s chin making it less invasive than, for example, EMA, and no calibration process is required. These advantages can make ultrasound an attractive option for research in this area. Typically, it is easier and more fruitful to be able to make within-speaker comparisons of sounds which are easily differentiated by different tongues shapes in the midsagittal dimension, for example, advanced and retracted tongue root vowels [2], liquid consonants across languages in the same speakers [3], or palatalised and non-palatalised consonants [4]. We will also give some tips on using ultrasound for teaching, demonstrations, and public engagement, with examples from our own experience e.g. [5]. Ultrasound tongue imaging can also be combined with lip camera images, although this won’t be covered in our workshop.

III. SOFTWARE AND HARDWARE

In this workshop we will use the hardware setup recommended for research by Articulate Instruments Ltd (see their webpage). This includes the Telemed MicrUS ultrasound machine, Convex 2-4MHz 20mm radius ultrasound probe, Pulse Stretch audio/ultrasound syncronisation unit, Ultrafit probe stabilisation headset [6], and microphone and sound cards options (Fig. 2). During the workshop we will record data in the software AAA [1], and conduct some initial analysis in Praat and AAA. Participants can work together on the laptops we will provide as the software requires a proprietary licence and only runs on Windows. If workshop participants have access to a AAA dongle and Windows laptop, they are welcome to download AAA and use their own devices.

IV. Recording Data

We will first discuss how to best fit the headset on a range of research participants with different sized heads and hairstyles. Workshop participants will be able to practise on each other and we will advise on how to obtain the best images from participants with different anatomies. We will then demonstrate how to set up a small research project in AAA and record some data. We first recommend recording the occlusal plane for each participant, for example by using a bite plate [7]. We then suggest recording each research participant swallowing some water to obtain an image of the hard palate for reference. We will discuss optimal time and numbers of repetitions for recording stimuli from different participants, as well as settings for the ultrasound and recording.

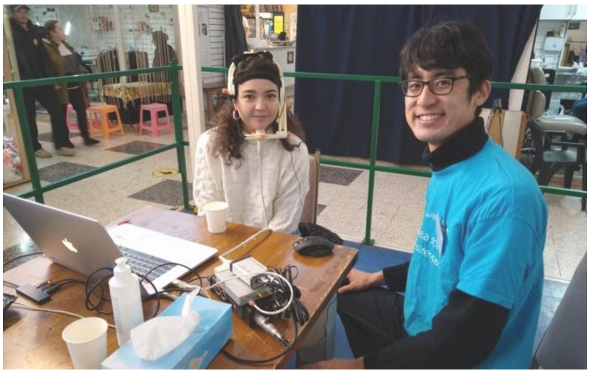


Fig. 2. Takayuki Nagamine recording ultrasound data with the Telemed MicrUS and Ultrafit headset.

V. Analysis First Steps

Once we have recorded some data, we will then show workshop participants a simple workflow for data extraction and analysis. Workshop participants will learn how to obtain tongue surface coordinates based on acoustic landmarks. In order to do this, we will export audio from AAA, and then label in Praat [8]. We will then reimport Praat TextGrids into AAA so that acoustic events are labelled synchronous with ultrasound video. The coordinates of the tongue surface are obtained by automatically fitting splines to the data using the DeepLabCut plugin in AAA [9], [10]. We will then show workshop participants how to export the coordinates of the tongue splines rotated to each research participant's occlusal plane.

VI. Exporting Videos

Finally, we will demonstrate how workshop participants can export videos from their data for sharing with participants, embedding in research presentations, or using for teaching and public engagement. Videos can be exported of fitted tongue splines and/or the ultrasound image and audio. For examples of how we have used this kind of resource, see this website (part of [11]).

References

[1] A. Wrench, *Articulate Assistant Advanced (Version 221.2)*. Edinburgh: Articulate Instruments, 2023.

[2] S. Kirkham and C. Nance, 'An acoustic-articulatory study of bilingual vowel production: Advanced tongue root vowels in Twi and tense/lax vowels in Ghanaian English', *J. Phon.*, vol. 62, pp. 65–81, 2017.

[3] T. Nagamine, 'Dynamic tongue movements in L1 Japanese and L2 English liquids', in *Proceedings of the 20th International Congress of the Phonetic Sciences*, R. Skarnitzl and J. Volín, Eds., Charles University, Prague: Guarant International, 2023, pp. 2442–2446.

[4] C. Nance and S. Kirkham, 'Phonetic typology and articulatory constraints: The realisation of secondary articulations in Scottish Gaelic rhotics', *Language*, vol. 98, no. 3, pp. 419–460, 2022.

[5] C. Nance *et al.*, 'Acoustic and articulatory characteristics of rhoticity in the North-West of England', in *Proceedings of the 20th International Congress of the Phonetic Sciences*, R. Skarnitzl and J. Volín, Eds., Charles University, Prague: Guarant International, 2023, pp. 3573–3577.

[6] L. Spreafico, M. Pucher, and A. Matosova, 'UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science', in *Interspeech 2018*, 2018. doi: 10.21437/interspeech.2018-995.

[7] J. M. Scobbie, E. Lawson, S. Cowen, J. Cleland, and A. Wrench, 'A common co-ordinate system for mid-sagittal articulatory measurement', *QMU CASL Work. Pap.*, vol. 20, 2011.

[8] P. Boersma and D. Weenik, 'Praat: doing phonetics by computer [Computer program]. Version 6.4.01'. Accessed: Nov. 30, 2023. [Online]. Available: http://www.praat.org/

[9] A. Mathis *et al.*, 'DeepLabCut: markerless pose estimation of user-defined body parts with deep learning', *Nat. Neurosci.*, vol. 21, no. 9, pp. 1281–1289, 2018, doi: 10.1038/s41593-018-0209-y.

[10] A. Wrench and J. Balch-Tomes, 'Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut', *Sensors*, vol. 22, no. 3, p. 1133, 2022, doi: 10.3390/s22031133.

[11] E. Lawson, J. Stuart-Smith, J. Scobbie, and S. Nakai, 'Seeing Speech: An articulatory web resource for the study of Phonetics.' Accessed: Aug. 23, 2023. [Online]. Available: https://seeingspeech.ac.uk

Phonetic experiments with *PsychoPy* and *PsychoJS*

Stefan Werner
University of Turku, Finland

Keywords — listening experiments, speech production experiments, PsychoPy, PsychoJS

I. AUDIENCE AND AIM OF THE WORKSHOP

This workshop is targeted at researchers with no previous experience in using *PsychoPy* [1] or *PsychoJS* [2]; familiarity with *Praat's* multiple forced choice listening experiments may be helpful but is also not required. Having attended the workshop, participants should be able to design, implement and run simple phonetic experiments both on a local computer and via the World Wide Web. The workshop's three-hour time limit is behind the "simple" experiment constraint; nevertheless, the acquired basic *PsychoPy/PsychoJS* competency should also help participants with working independently towards more sophisticated setups later.

II. WORKSHOP TOPICS

The first part of the workshop will introduce participants to the basic workings of *PsychoPy*. Since 2002, this open-source program has been widely used for behavioral experiments in, above all, psychology and psycholinguistics, but less so in phonetics, despite the fact that recent versions offer comprehensive audio processing functionality. *PsychoPy* is much more versatile and adaptable than Praat's ExperimentMFC or demo window and, at the same time, easier to use and more comprehensively documented.

PsychoPy provides two different environments for designing experiments, the visual BUILDER and the text-based CODER. Whilst the CODER makes possible an even larger variety of experimental setups than the BUILDER it also comes with a steeper learning curve, especially for users without prior exposure to Python programming. Thus, in our workshop we will concentrate on the BUILDER to construct a phonetic experiment.

In the second part of the workshop we will find out how to adapt this experiment to online use, gathering data from subjects via the internet. This is made possible by *PsychoJS*, the online variant of *PsychoPy*. JS stands for JavaScript, the programming language most often used in web applications executed in your browser – but again, workshop participants will not be required to learn how to write programming language code. Instead, *PsychoPy's* graphical BUILDER tool can also be used to produce the necessary JavaScript and HTML code.

LINKS

[1] https://psychopy.org/

[2] https://github.com/psychopy/psychojs

The perceived similarity between retroflexes by bi- and multilinguals

Anna Balas [a], Krzysztof Hwaszcz [a,b], Magdalena Wrembel [a,], Kamil Kaźmierski [a]
[a] *Adam Mickiewicz University, Poznań, Poland,*
[b] *University of Wrocław, Poland*

Keywords — non-native speech perception, similarity, retroflex, multilinguals,

I. Introduction

This paper investigates perceived cross-linguistic similarity of retroflexes by multilingual and bilingual learners. It has previously been shown for non-native speech that L2 learners rely on their L1, but no study has so far examined how L3 learners perceive differences between either L1 or L2 and L3 consonants. Also, our understanding of the factors that can affect judgements of phonetic dissimilarity for L3 (e.g., [1] and [2]) is very limited. Addressing this research gap, the present study takes the idea of cross-linguistic similarity further than it has been done so far in L2 to apply it in multilingual phonological acquisition. It examines crosslinguistic similarity of Norwegian retroflexes and similar retroflex and non-retroflex sounds by multilingual (L1 Polish, L2 English and L3 Norwegian), bilingual (no familiarity of Norwegian) and Norwegian control listeners. Previous research on the perception of retroflexes includes a study by [3] in which she found differences in perceptual difficulty related to phonemic status, experience and voicing and a study by [4] which showed that prior allophonic experience with dental and retroflex stops could actually be detrimental to learning a new contrast based on retroflexion.

II. Methodology

A subtractive language group design [5] was used. The experimental group featured 34 L1 Polish, L2 English and L3 Norwegian students majoring in Norwegian within a BA program (all classroom learners, 5 males, 28 females and one non-binary person, mean age 21.3). The bilingual group featured 35 L1 Polish L2 English subjects (undergraduate students of English language and literature program, 11 males, 24 females, mean age 20.9). The Norwegian native speaker control group consisted of 25 listeners with L1 Norwegian and L2 English (10 females, 14 males, 1 non-binary person, mean age 23.08).

During the experiment, the listeners were instructed to listen to pairs of nonce-words, pay attention to the consonant in the middle of each word and rate how similar or dissimilar the consonants they heard were on a seven-point scale (1= very dissimilar, 7 = very similar). Norwegian retroflexes /ʈ, ɖ, ʃ, ɭ, ɳ/ and their non-retroflex counterparts /t̪, d, s, l, n/ were paired with either Polish or English retroflex or non-retroflex counterparts. Even though Polish /ʂ/, /ʐ/, /t͡ʂ/ and /d͡ʐ/ do not involve backward bending of the tongue, they are classified as retroflexes on the basis of X-ray tracings in literature [6], [7], electromagnetic articulography [8], phonological evidence [9], [10], acoustic features [11], [12], [13] and sound change in Slavic languages [14], [15]. Also, allophonic retroflexion occurs in [ʈ] and [ɖ] [12], [13] when they are followed by one of the retroflex sibilants, as in *trzeba* [ʈʂɛba] 'it is necessary' and *drzewo* [ɖʐɛvɔ] 'tree.' The above method of stimuli selection resulted in four conditions in which the pairs of sounds were presented. The two stimuli in a pair either (1) matched with regard to retroflexion and had similar place and manner of articulation; (2) did not match with regard to retroflexion and had similar place and/or manner of articulation; (3) matched with regard to retroflexion but had different place and manner of articulation; (4) did not match with regard to retroflexion and had different place and/or manner of articulation.

III. Results and discussion

Comparing proportions of (dis-)similarity ratings, we can observe that in conditions 3 and 4 Norwegian controls evaluated stimuli as 'very different' less frequently than bilinguals or trilinguals. We fitted a mixed-effects ordinal regression model of similarity ratings as a function of three treatment-coded categorical predictors: condition (in which the sounds were presented, levels 1 through 4, see the paragraph above), language of the non-Norwegian phone (levels: English and Polish) and group (levels: Norwegian controls, bilinguals and trilinguals) and their three-way interaction. We found significant effects of condition ($\chi^2(3) = 6335.4$, $p < .001$). Based on this significant result and on the results of post-hoc pairwise comparisons across the different levels of condition on similarity ratings, it can be claimed that retroflexion turned out to be a weaker cue to similarity than place and/or manner of articulation. The results also revealed that experience with a given language did not influence similarity ratings in a wholesale manner but rather in a precise manner related to the presence or absence of retroflexion (cf. [16]'s finding about the lack of language-specific exposure effect for the perception of retroflexes by English-Mandarin bilinguals). The perceived cross-linguistic similarity by multilinguals has turned out to be gradient, dependent more on similar places/manners of articulation rather than on the presence or absence of retroflexion or familiarity with a given language combination.

The results provide evidence for the ability of adult multilingual learners to, at least partially, separate the three language systems. Trilinguals, familiar with Norwegian, appeared to be more sensitive to retroflexion than bilinguals with no knowledge of Norwegian. This result points to the importance of experience with a given contrast. Future studies could disentangle the role of multilingualism as opposed to experience with a feature. Gradience in perceptual salience also needs to be taken into account, as not all foreign sound features are equally well heard by learners and it is beneficial teachers or curricula designers to realize which features are more robust.

Acknowledgements

This research is supported by Norway funds/ NCN project grant GRIEG-1 (UMO-2019/34/H/HS2/00495) ADIM "Across-domain investigations in multilingualism: Modeling L3 acquisition in diverse settings". Special thanks should also go to Jarosław Weckwerth and Agnieszka Pludra for their idea of shading and help with the execution. We would also like to thank Hanna Kędzierska for her help with data collection in Wrocław and Sylwiusz Żychliński, Anna Skałba and Zuzanna Cal for cooperation during data collection in Poznań and Szczecin. Finally, we would like to thank anonymous reviewers for their helpful comments.

References

[1] A. Rato, "Perceptual categorization of English vowels by native European Portuguese speakers," *LinguíStica – Estudos Experimentais sobre o Português, 14*(2), pp. 61-80, 2018. https://doi.org/10.31513/linguistica.2018.v14n2a17542

[2] J. Cebrian, "Perception of English and Catalan vowels by English and Catalan listeners: Part II. Perceptual vs ecphoric similarity," *Journal of the Acoustical Society of America, 152*(5), 2781, 2022, doi:10.1121/10.0014902. PMID: 36456284

[3] L. Polka, "Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions." *The Journal of the Acoustical Society of America* 89(6), pp. 2961-2977, 1991.

[4] J. S. Pruitt., J. J. Jenkins, and W. Strange, "Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese," *The Journal of the Acoustical Society of America* 119(3), pp.1684-1696, 2006. doi:10.1121/1.2161427

[5] M. Westergaard, N. Mitrofanova, Y. Rodina, and R. Slabakova, "Full transfer potential in L3/Ln acquisition: Cross-linguistic influence as a property-by-property process," in *The Cambridge handbook in thrid language acquisition,* J. Cabrelli, A. Chauch-Orozco, J. González Alonso, S.M. Pereira Soares, E. Puig-Mayenco, and J. Rothman, Eds. Cambridge: Cambridge University Press, 2023, pp. 219-242.

[6] S. Hamann, *The Phonetics and Phonology of Retroflexes,* Utrecht: LOT, 2003.

[7] S. Hamann, 'Retroflex fricatives in Slavic languages', *Journal of the International Phonetic Association*, vol. 34, no. 1, pp. 53–67, Jan. 2004, doi: 10.1017/S0025100304001604.

[8] A. Lorenc, "Articulatory characteristics of Polish retroflex sibilants. *An analysis using electromagnetic articulography," Logopedia,* vol. 47, pp. 103–129, Dec. 2018.

[9] T. A. Hall, „The historical development of retroflex consonants in Indo-Aryan," *Lingua*, vol. 102, no. 4, pp. 203–221, Aug. 1997, doi: 10.1016/S0024-3841(96)00050-2.

[10] T.A. Hall, *The phonology of coronals*, Amsterdam: John Benjamins, 1997.

[11] M. Żygis, S. Hamann, "Perceptual and acoustic cues of Polish coronal fricatives," in: Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, August 3-9, M. J. Solé, D. Recasens, and J. Romero, Eds., ICPhS Archive, 2003, pp. 395-398.

[12] M. Żygis, D. Pape, and L. M. T. Jesus, "(Non-)retroflex Slavic affricates and their motivation: Evidence from Czech and Polish," *Journal of the International Phonetic Association*, vol. 42, no. 3, pp. 281–329, Dec. 2012, doi: 10.1017/S0025100312000205.

[13] M. Żygis, D. Pape, and L. M. T. Jesus, „(Non-)retroflex Slavic affricates and their motivation: Evidence from Czech and Polish", *Journal of the International Phonetic Association*, vol. 42, no. 3, pp. 281–329, Dec. 2012, doi: 10.1017/S0025100312000205.

[14] J. Padgett and M. Zygis, "The evolution of sibilants in Polish and Russian", *ZASPiLl*, vol. 32, pp. 155–174, Jan. 2003, doi: 10.21248/zaspil.32.2003.190.

[15] M. Żygis and J. Padgett, "A perceptual study of Polish fricatives, and its implications for historical sound change," *Journal of Phonetics*, vol. 38, no. 2, pp. 207–226, Apr. 2010, doi: 10.1016/j.wocn.2009.10.003.

[16] H. L. Goh, L. Onnis, and S. J. Styles, "Is retroflexion a stable cue for distributional learning for speech sounds across languages? Learning for some bilingual adults, but not generalisable to a wider population in a well powered pre-registered study," *PeerJ*, vol. 11, p. e15467, Jul. 2023, doi: 10.7717/peerj.15467.

Finetuning Large Pretrained Phonetic ASR Models for Reading Miscue Detection in Primary School

Lingyun Gao [a], Cristian Tejedor-Garcia[a], Louis ten Bosch[a], Helmer Strik [a], Catia Cucchiarini[a],

[a] *Centre for Language Studies, Radboud University Nijmegen, Netherlands*

Keywords — automatic speech recognition, phoneme recognition, child speech recognition, reading miscue detection, large pretrained models

I. Introduction

Recent advancements in Large Pretrained Automatic Speech Recognition (ASR) have unveiled many promising digitalized education applications. One such application involves utilizing ASR for reading diagnosis in primary schools [1,2], a practice that holds the potential to benefit both teachers, enabling more efficient evaluation of students' reading abilities, and students, facilitating easier access to reading exercises with feedback.

The task of phoneme-level reading miscue detection in reading diagnosis for children in primary school is particularly challenging due to the following reasons. Firstly, previous research [3,4] has shown that state-of-art (SOTA) Dutch ASR models pretrained on adult speech exhibit significantly reduced performance when applied to child speech recognition, especially in phoneme recognition and leads to challenge in downstream reading miscue detection tasks. The decrease in performance may be attributed to the larger variability in children speech (e.g. due to different vocal tract characteristics), and children's tendency to produce numerous disfluencies (hesitations, broken words, and [filled] pauses), which are rare in healthy adult reading speech [1]. Secondly, conventional automated reading miscue detection methodologies typically operate under the assumption of a single canonical phonemic transcription for each word [1,5]. However, in real-life situations, a single word is usually articulated in various ways, including reduced pronunciations. The oversight of reduced pronunciations can lead to many false detections of phoneme-level reading miscues.

To address the above challenges, we plan to investigate finetuning large pretrained models on child speech, followed by post-processing to address reduced pronunciation. [1] We anticipate making the following contributions: (1) Investigating diverse finetuning strategies for obtaining the best phoneme transcription with word boundaries, within low resource Dutch native children speech settings, sourced from the Jasmin-CGN corpus [6]; (2) offering SOTA open-source Dutch native children phoneme-level recognition models and finetuning test pipeline codes; and (3) developing a novel finetuned large pretrained ASR-based system capable of supporting phoneme-level reading miscue detection and mitigating the impact of reduced pronunciation for primary school students.

II. Method

A. Dataset

This study will utilize reading speech data from Dutch native primary school children obtained from the Jasmin-CGN Corpus [6]. This Corpus contains recordings of children reading aloud at their mastery reading level, which were aligned with phonemic annotations, from 71 primary school pupils (age range = 6-13 years old), consisting of 35 female and 36 male children. Additionally, we employ prompt text, the reading miscue and reading strategy annotations for the first read story available in [7].

The child speech in Jasmin-CGN is separated into the training and testing dataset. The training dataset comprises 6.55 hours of speech, while the testing dataset comprises the speech of the first story read by children, including 2.05 hours of speech with 14,251 reading attempts, with 615 reading attempts labeled incorrect with certain phoneme-level miscues.

B. Tasks

The first task consists of the automatic classification of four different phoneme-level miscues: insertion, deletion, substitution and reverse orders. In the next step, we will consider detecting more detailed phoneme-level miscues: restart insertion, insertion of vowels or consonants, deletion of vowels or consonants, substitution of vowels or consonants and reverse orders

C. Finetune ASR models and Metrics

In previous recent research [3], it has been tested the top-three publicly available large pretrained phoneme level ASR model in recognizing child speech data from the JASMIN-CGN corpus as shown in Table I. Hubert-Large demonstrates the best phoneme error rate at 27.1% for Jasmin-CGN native children in primary school. In the current study, we go beyond the SOTA by developing a new child speech recognition system based on this top-performing model, Hubert-Large. The Phoneme Error Rate (PER) metric will be used for phoneme-level recognition evaluation.

[1] This work was supported by the NWO research programme AiNed Fellowship Grants under the project Responsible AI for Voice Diagnostics (RAIVD) - NGF.1607.22.013.

TABLE I. PHONEME RECOGNITION PERFORMANCE OF LARGE PRETRAINED ASR ON CHILD SPEECH

Phoneme ASR Model	Wav2vec-base [8]	Wav2vec-xlsr [9]	Hubert [10]
PER	32.6%	36.1%	27.1%

In our reading miscue annotation, phoneme-level reading miscues are recorded within each reading prompt. Additionally, as reduced pronunciations are defined for each word, for example, /h E t/ or /h @ t/ or /E t/ for word [het], it is required in our task to have word or reading attempt boundary in phoneme output from the ASR. There are diverse finetune strategies in recent research to obtain such output. In this study, we would like to investigate what is the best finetuning strategy for phoneme-level child speech recognition in low-resource settings and we consider the following options shown in Table II.

TABLE II. PHONEME RECOGNITION PERFORMANCE OF LARGE PRETRAINED ASR ON CHILD SPEECH

Finetuning Strategy	Description
Phoneme ASR + finetuned with Jamin speech and phoneme transcription with word boundaries	Direct finetuning approach.
Phoneme ASR + pretrained and finetuned with Jamin speech and phoneme transcription with word boundaries	Adds a pretraining step to the direct finetuning method.
Phoneme ASR + finetuned with Jamin speech and phoneme transcription + finetuned with Jamin speech and phoneme transcription with word boundaries	Inspired by cumulative learning method, potentially improves performance over direct learning.
Phoneme ASR + finetuned with Jamin speech and phoneme transcription + Boundary detector trained on JASMIN phoneme transcriptions with word boundaries (or with more text data from other Adult Dutch dataset)	Utilizes additional text data for potentially better word boundaries detection performance.

D. Post-processing for Reduced Pronunciation

We explore two potential approaches to post-processing in ASR output. Firstly, standardization involves utilizing lexicons to convert each reduced pronunciation identified in ASR phoneme output into a single canonical pronunciation. This method offers computational efficiency as its key advantage. Alternatively, dynamic alignment with lexicons enables providing reduced pronunciations for each word in the reading prompt (reference) and search for the best match for hypothesis ASR phoneme output. This approach preserves the way children pronounce a word, allowing for deeper analysis, albeit at the cost of increased algorithmic complexity. Specifically, in dynamic alignment, Levenshtein distance is computed by measuring minimum edit distance from hypothesis phoneme segment ***a*** to reference ***bi*** belong to ***lexicon(b). bi*** is any acceptable pronunciation of word ***b.***

E. Phoneme-level Reading Miscue Detection

We will develop our reading miscue detection system based on previous finetuned ASR, post-processing and a miscue classifier. Specifically, we align reading prompt with ASR output to detect phoneme-level reading miscues and compare them with human annotated phoneme-level miscues. Precision, Recall and F1/F2 scores are employed for evaluating phoneme-level reading miscue detection performance.

EXPECTED RESULTS

Phoneme recognition: We expect to provide SOTA open-source phoneme-level recognition models for Dutch native primary school children's data in Jasmin-CGN corpus, which have better performance than the results reported in [2,3].
Analysis: We expect to conduct an analysis of four fine-tuning methods, specifically focusing on comparing the performance across different aspects with the baseline pretrained model and providing insights on what has been improved. This will include assessing the top 10 phonemes with the lowest accuracy, identifying the top 10 confusion pairs, insertion and deletion errors.
Detection: We intend to create a novel large pretrained ASR-based system, finetuned to facilitate phoneme-level reading miscue detection while mitigating the impact of reduced pronunciation. We expect that through post-processing, there will be enhancements in the performance of phoneme-level reading miscue detection compared to systems lacking such post-processing.

References

[1] Y. Bai, C. Tejedor-Garcia, F. Hubers, C. Cucchiarini, and H. Strik, "An asr-based tutor for learning to read: How to optimize feedback to first graders," in *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*, 2021, pp. 58–69.

[2] J. Mostow, J. Nelson-Taylor, and J. E. Beck, "Computer-guided oral reading versus independent practice: Comparison of sustained silent reading to an automated reading tutor that listens," *Journal of Educational Computing Research*, vol. 49, no. 2, pp. 249–276, 2013.

[3] Anonymous, "In review: Reading Miscue Detection in Primary School through Automatic Speech Recognition," in *Interspeech*, 2024.

[4] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[5] J. Proença, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigao, "Mispronunciation Detection in Children's Reading of Sentences," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 26, no. 7, pp. 1207–1219, 2018.

[6] C. Cucchiarini, J. Driesen, H. Van hamme, and E. Sanders, "Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus.," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[7] S. Limonard, C. Cucchiarini, R. van Hout, and H. Strik, "Analyzing read aloud speech by primary school pupils. Insights for research and development," 2020.

[8] C. Apavou, "Fine-tuned phoneme recognition Wav2Vec2 base Model." 2022. https://huggingface.co/Clementapa/wav2vec2-base-960h-phoneme-reco-dutch/tree/main.

[9] Q. Xu, A. Baevski, and M. Auli, "Simple and Effective Zero-shot Cross-lingual Phoneme Recognition." arXiv preprint arXiv:2109.11680, 2021.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." 2021.

An Open Source Tool for Practicing the Pronunciation of Words in Lesser Used Languages

Wilbert Heeringa [a], Hans Van de Velde [a,b]
[a] *Fryske Akademy, The Netherlands,*
[b] *Utrecht University, The Netherlands*

***Keywords* — pronunciation training, lesser used languges, minority languages, open source tool, language learning**

I. Introduction

When moving to another area, it may be both useful and enriching to learn the regional language of that area. At the same time, learning the pronunciation of a regional language can be sensitive. Minority languages often hold significant cultural and historical importance to the communities that speak them, and correct pronunciation is often closely tied to the identity of the native speakers. Learning correct pronunciation is therefore important, but may become a challenge when language learning tools are lacking due to little or even no (financial) possibilities for the language community to develop them. This in contrast to larger languages such as English for which there is a wealth of language learning tools available.

We present an open source web app for practicing the pronunciation of words in low-resource and commercially less-interesting languages. Since it is web-based it works on all platforms: MS Windows, Mac OS X, Linux, Android, iOS, etc. The web app is device independent and works on desktop computers, laptops, tablets and smartphones. The app will become available as open source software and in the public domain as a generic application. i.e. with minimal effort the app can be adapted so that it can be used for any language when recordings of a series of words pronounced by at least one reference speaker are provided. The app has initially been developed for Latgalian and West Frisian. Latgalian is an Eastern Baltic language and mostly spoken in Latgale, the eastern part of Latvia. West Frisian is a Germanic minority language spoken primarily in the Dutch province of Fryslân.

II. The Interface

The app will work as follows. In the first screen the user chooses a reference speaker, a set of words, the assessment level and the gender (see Fig. 1). As reference speaker a speaker is chosen whose pronunciation the user wants to imitate. As to the words to be pronounced, the user can choose a set of words that have a particular vowel or the full set of words. There are two assessment levels: lenient and strict. Providing the gender is important when the reference speaker that was chosen and the user do not have the same gender.

After a few screens with instructions on how to use the app, the actual training starts. The user hears a word, clicks on the recording button and pronounces the same word (see Fig. 2). Then a rating is given of maximally five stars when the user's pronunciation is a perfect imitation of the pronunciation of the reference speaker. The user can also play the user's own pronuncation after it has been recorded and compare this to the pronunciation of the reference speaker by playing the reference speaker's pronunciation again. By clicking on the arrow at the bottom of the screen the program moves to the next word.

In order to compare the user's pronunciation with the pronunciation of the reference speaker we use the methodology of [1], who developed a distance measure which they used for assessing foreign accent strength in American-English. The speech of non-native American-English speakers was compared to a collection of native American-English speakers. The authors found a strong correlation between the acoustic distances and human judgments of native-likeness provided by more than 1,100 native American-English raters ($r = -0.71$, $p < 0.0001$).

The procedure that we used for rating the user's pronunciation is as follows. For both the recording of the word pronounced by the reference speaker and the recording of the same word pronounced by the user a representation based on Mel-frequency cepstral coefficients (MFCCs) is calculated. A MFCC representation consists of a series of frames where each frame includes 12 MFCC coefficients. The 12 parameters are related to the amplitude of the frequencies. MFCCs are popular due to their greater invariance to physical differences between speakers [2]. Prior to calculating the MFCC representations silence preceding and following the word pronunciation is cut off using a Praat script. Then with the same script the representations are calculated with a window length of 0.015 seconds and a time step of 0.005 seconds.

The quality of the MFCC feature representation is highly influenced by the presence of noise in the speech samples [3][4]. The effect of noise can be reduced by standardizing the MFCC coefficients. Individually for each of the 12 parameters the mean and standard deviation are calculated over the MFCC coefficients in the course of the time. Subsequently, the mean is removed from the coefficients, and the resulting values are divided by the standard deviation. We understand that [1] normalize per speaker. However, this requires all the recordings to be available in advance, which is naturally not the case in our application. Therefore, we standardize per word sample.

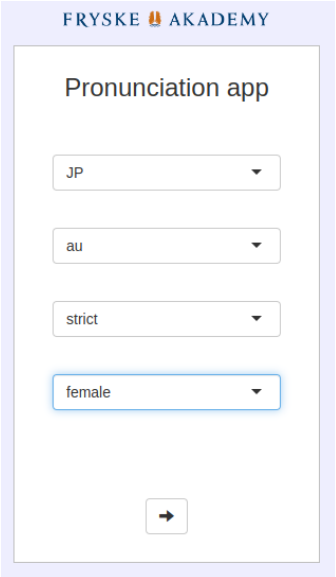


Fig. 1. Opening screen of the app. The speaker ‘JP’ is chosen, words containing *au* in the orthographic from will be trained at a strict level. The user is a female speaker.

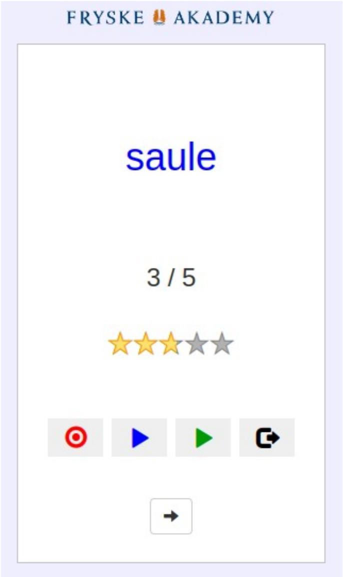


Fig. 2. The Latgalian word *saule* ‘sun’ is played. The user clicks on the red button and pronounces the same word. The pronunciation is rated with three stars.By clicking on the blue button the word *saule* can be played again. When clicking on the green button the user’s own pronunciation is played.

III. Rating the User’s Pronunciation

The acoustic word distance between the pronunciation of the user and the pronunciation of the reference speaker is computed using the dynamic time warping (DTW) algorithm [5]. The frames of the two respective MFCC representations are aligned to each other. Every frame in the one representation must be matched with one or more frames from the other representation, and vice versa. In order to find a logical match of the frames in the one representation with the frames in the other representation, frames are compared to each other. Reference [1] uses the Euclidean distance. We calculate 1 minus Pearson’s correlation as distance between two frames, which gives easy to interpret distances between 0 and 1, while we found it functioning well. The DTW algorithm matches the frames so that the overall distance between the two sequences of frames is minimized. Subsequently, [1] normalizes that distance by dividing it by the sum of the lengths of the two representations. Instead we normalize by dividing by the length of the alignment, which we judge to be more precise. Since the frame distance varies between 0 and 1, the normalized distance will vary between 0 and 1 as well. In the app the distances are mapped on five stars, where five stars are obtained when the distance is 0 and the assessment level is set to ‘strict’. When the assessment level is set to lenient, five stars are obtained already when the distance is slightly higher than 0, for example 0.2. The lenient level is meant to prevent discouragement when background noise or the quality of the user’s microphone make it impossible to achieve a fair rating.

IV. Recording Reference Speakers

The app needs to be provided with recordings of reference speakers that have pronounced a series of different words. The words should be chosen so that all vowels and consonants of the language are represented in different contexts. We provide an open source recording app that can be used to make the recordings. In the app the words are pronounced one by one in a randomized order. Since the words are recorded one by one, no subsequent cutting or labeling of words is necessary, the word samples can be used immediately in the training app. In the recording app you can choose how many times each word type should be pronounced. Then afterwards, for each word type the most representative version can be found with a script and be used in the training app.

V. Research Tool

With the training app it is possible to monitor the training process, if the user has given permission for this. When the user's number of attempts per word is recorded, as well as the ratings per attempt, we get an idea of the user's achievements and how much effort it took to arrive at the (most) correct pronunciation. These data become especially meaningful if it is also known what the native language of the speaker was.

The app is still in development, therefore, benchmarks are not available yet.

References

[1] M. Bartelds, C. Richter, M. Liberman, and M. Wieling, "A New Acoustic-Based Pronunciation Distance Measure., Front. Artif. Intell. vol. 3:39, doi: 10.3389/frai.2020.00039, 2020.

[2] S. Davis, and P. Mermelstein, " Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,", IEEE Trans. Acoust. Speech Signal Process, vol. 28, pp. 357–366, doi: 10.1109/TASSP.1980.1163420, 1980.

[3] S. Ganapathy, J. Pelecanos, and M.K. Omar, "Feature normalization for speaker verification in room reverberation," , IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Prague), pp. 4836–4839, doi: 10.1109/ICASSP.2011.5947438, 2011.

[4] A. Shafik, S.M. Elhalafawy, S. Diab, B.M. Sallam, and F.A. El-Samie, "A wavelet based approach for speaker identification from degraded speech," Int. J. Commun. Netw. Inform. Secur., vol. 1:52, 2009.

[5] J. Galbally, and D. Galbally, "A pattern recognition approach based on DTW for automatic transient identification in nuclear power plants," Ann. Nucl.Energy 81, 287–300, doi: 10.1016/j.anucene.2015.03.003, 2015.

Practicing Estonian pronunciation with SayEst

Anton Malmi [a], Katrin Leppik [a]

[a] *University of Tartu, Institute of Estonian and General Linguistics, Estonia*

Keywords — speech training, mobile app, vowel perception, speech production, pronunciation

I. Introduction

SayEst is an Android mobile app designed to help users improve their Estonian pronunciation. Previous studies have shown that such tools and phonetic speech training can help to improve pronunciation [1], [2] [3], [4]. The app is available in English and Russian and focuses on training the perception and production of vowels and consonants. It has three different exercises: exposure, identification, and pronunciation, and includes theoretical videos explaining Estonian pronunciation.

One-fourth of the Estonian population speaks Russian as their L1, and the number of Russian L1 learners of Estonian has increased in recent years. This has also increased the need for different tools for speech training. Previous studies involving Russian L1 learners of Estonian have focused mainly on vowels and quantity degrees (e.g., [5], [6], [7]). Less is known about the production and perception of consonants. [8] found that Russian-accented speech in Estonian consists of a combination of different acoustic features, the most relevant of which are the deviations in temporal structure, stress, and quality of phonemes. Well-known speech acquisition theories, like SLM(-r) [9] and PAM [10], show that speech accent arises from the fact that the phonetic and phonological systems of learners' L1 directly affect the way their L2 is produced and perceived. From that, patterns of acquisition emerge. Russian has a smaller vowel inventory than Estonian, and Russian L1 learners of Estonian tend to assimilate the Estonian vocalic categories to their native categories. For example, [6] found that vowels /ø/ and /ɤ/ produced by Russian L1 learners deviated from L1 vowels significantly, and the vowels /i/, /y/, and /o/ were shifted backward along the front-back dimension. The current study aims to find out to what extent SayEst improves the learners' perception and production of Estonian and which vowels and consonants are difficult for Russian L1 speakers. The data collected in this study will be used to develop and train an L2 speech verifier that will give the user qualitative feedback on each segment in a word.

II. Methods

To assess the efficacy of the app, we asked 30 Russian L1 learners of Estonian (24 females, 6 males, mean age 31, SD = 9.3) with different proficiency levels (A - beginner, B - intermediate, C - advanced, 10 speakers in each group) to participate in a pre-and post-test design study with an unsupervised training between them. First, the participants completed two perception tasks (vowel and minimal pair identification) and a reading task at the phonetics lab. The same procedure was repeated after the participant completed the exercises in the app at home, which, on average, took them about 4–10 hours. They were instructed to return to the post-test no later than a week, but on average, they returned in 21 days (min = 9, max = 42, SD = 7.8). The participants filled out a self-assessment questionnaire before and after using the app and were interviewed about their experience.

In the vowel identification task, the listener heard an isolated vowel in Estonian and had to choose to click on the correct corresponding letter. In the minimal pair identification task, the listener heard a word and had to choose the correct word from a minimal pair. The reading task consisted of a list of 180 short words similar to the words that they practiced in the app. The words consisted of all the nine vowels /ɑ, e, i, o, u, ɤ, æ, ø, y/ that are present in Estonian and consonants /h, j, k, l, m, n, p, r, s, t, v/. After the data collection, the data was randomly divided between four expert phoneticians. They annotated the sound files and assessed the auditory quality of each of the phonemes in the word on a 3-point scale. 1 was given when the phoneme sounded very native-like, 2 when the phoneme was close but deviated towards another phoneme, and 3 when a different phoneme was produced than expected. T-test was used in R to analyze the differences and similarities between groups.

III. Results and discussion

The results of the questionnaire showed that, on average, the self-rating of the participants' pronunciation (2,83 vs. 3,26) and their confidence rating in speaking Estonian (2,93 vs. 3,27) were higher after using the app, but these changes were not statistically significant. The questionnaire showed that the participants paid more attention to how they spoke (3,47 vs. 4,1, t = 2.1) after using the app.

The results of the vowel identification task showed that there were no statistically significant differences between the results of the pre-and post-test. However, the learners of group B improved slightly (correct responses in pre-test 92%, post-test 95%) while the percentage of correct responses of group A (pre - 74%, post - 72%) and C (pre - 94%, post - 90%) decreased. The comparison by proficiency levels showed that group A had a lower percentage of correct answers than groups B and C, and they made more mistakes. In general, the learners misidentified the vowels /ɤ/, /ø/ and /y/ most often (percentage of correct responses of /ɤ/ by group A - 30%, B - 85%, C - 90%; /ø/ A - 60%, B - 86%, C - 75%, and /y/ A - 65%, B - 80%, C - 81%). These vowels were confused with each other and with /u/ or /i/. This aligns with previous studies, e.g., [6] also found that vowels /ø/ and /ɤ/ produced by Russian L1 learners deviated the most from L1 production. Somewhat unexpectedly, all groups often misidentified the vowel /e/ as /i/.

The results for the perception test of minimal pairs showed that there were no differences between the pre- and post-tests (4,5% vs. 4,8% misidentification). When the answers were pooled together, the results showed that the overall percentage of misidentification decreased with proficiency level. Group A misidentified the highest percentage of pairs - 10%, group B - 3,1% and C - 1%. The minimal pairs task was much easier for the participants. This was evident from the small number of errors that they made compared to the vowel identification task, where they had to choose between 9 vowels.

The analysis of the scores given by the annotators showed that in the case of accented pronunciation, the vowels produced by all groups were more often rated as deviant (rating 2, 13%) rather than incorrect phonemes (rating 3, 4,4%). On average, the pre-test ratings (ratings 2 and 3 pooled together) were higher than the post-test ratings (pre - 9,6% vs. post – 7,8%), but the difference was not statistically significant. On closer inspection, most of the pronunciation errors occurred in group A (ratings 2 and 3 pooled together) with vowels /ø/ - 46,3%, /ɤ/ - 40,7%, /æ/ - 23,8% and /y/ - 24,8%. Noticeably, group B also had most of the errors with the same vowels /ø/ - 13,5%, /ɤ/ - 21,5%, /æ/ - 6,7%, and /y/ - 8,5%. The errors with /ɤ/ - 11,4% stand out in the C group. The percentages of pronunciation errors for /y/ in group A and for /ɤ/ in group B were higher in the post-test. In general, the number of errors reduced with the proficiency level of the speaker in the post-test.

The annotators noted a smaller number of pronunciation errors with consonants than vowels. The consonants in all the groups were rated with 2 (9,5%) rather than 3 (1,4%). On average, the pre-test ratings were higher (pre - 6,6% vs. post - 4,4%), but the difference was not statistically significant. Most pronunciation errors occurred again in group A with consonants /j/ - 22,5%, /l/ - 21,9%, /t/ - 10,2%, /k/ - 8,6%, and /h/ - 8,9%. Group B had problems with the same consonants. /l/ - 13,9%, /j/ - 10,2%, /k/ 10,8%, /t/ - 8,2%. Group C had the most errors with /l/ - 7,4% and /p/ - 5,8%. The annotators noted that the Russian L1 learners often produce plosives as voiced, but Estonian does not make the phonological distinction between voiced and voiceless plosives as Russian does. The annotators also noticed that the learners produced Estonian /l/ with a darker quality and often confused /i/ as /j/ in the production.

The results showed a small difference between the pre- and post-test results. In some cases, the post-test results were worse than the pre-test results. The inconsistency between the pre- and post-test results might be because the learners were overwhelmed with all the new information they acquired, and while they were more conscious of their pronunciation, they were more uncertain. It would benefit the learners to have a teacher from whom they can ask questions regarding the material. Also, it is important to mention that when grouped together, the values do not capture the nuances of individual differences. On the positive side, the time spent with the app made the users notice aspects of their pronunciation that they did not notice before.

IV. Conclusions

The current study concluded that the mobile app SayEst has the potential to improve the learners' perception and production of Estonian vowels and consonants. Using the app helped the users to notice the nuances of Estonian pronunciation. It might be best to use it in a classroom as supplementary material, as unsupervised training did not seem to give the best results. The vowels /ɤ/, /ø/, /æ/, and /y/ were misidentified most often and received the lowest ratings from annotators. The learners had the most problems with consonants /j/, /l/, /t/, /k/, and /h/. In order to get a better understanding of which aspects of Estonian are difficult for Russian L1 learners, an acoustic analysis will be carried out as the next step of the study. The data gathered in this study will be used to develop and train an Estonian L2 speech verifier and to improve the app further.

V. References

[1] S. Savo and M. S. Peltola, 'Arabic-speakers Learning Finnish Vowels: Short-term Phonetic Training Supports Second Language Vowel Production', *Journal of Language Teaching and Research*, vol. 10, no. 1, Art. no. 1, Jan. 2019, doi: 10.17507/jltr.1001.05.

[2] K. Leppik, C. Tejedor-García, E. L. Asu, and P. Lippus, 'Improving Spanish L1 learners' perception and production of Estonian vowels', *Proc. ISAPh 2022, 4th International Symposium on Applied Phonetics*, pp. 34–39, 2022, doi: 10.21437/ISAPh.2022-7.

[3] L. Taimi, K. Jähi, P. Alku, and M. S. Peltola, 'Children learning a non-native vowel - The effect of a two-day production training', *Journal of Language Teaching and Research*, vol. 5, no. 6, pp. 1229–1235, Nov. 2014, doi: 10.4304/jltr.5.6.1229-1235.

[4] J. F. Hacking, B. L. Smith, and E. M. Johnson, 'Utilizing electropalatography to train palatalized versus unpalatalized consonant productions by native speakers of American English learning Russian', *Journal of Second Language Pronunciation*, vol. 3, no. 1, pp. 9–33, Apr. 2017, doi: 10.1075/jslp.3.1.01hac.

[5] L. Meister, *Eesti vokaali- ja kestuskategooriad vene emakeelega keelejuhtide tajus ja häälduses. Eksperimentaalfoneetiline uurimus. (Doktoritöö, Tartu Ülikool, eesti ja üldkeeleteaduse instituut).* in Dissertationes philologiae estonicae Universitatis Tartuensis, no. 30. Tartu: Tartu Ülikooli Kirjastus, 2011.

[6] L. Meister and E. Meister, 'Production and perception of Estonian vowels by native and non-native speakers', *Interspeech 2011 : Conference Program and Abstract Book, 27-31 August 2011, Florence, Italy*, pp. 1145–1148, 2011.

[7] E. Meister and L. Meister, 'The production and perception of Estonian quantity degrees by native and non-native speakers', in *Interspeech 2012 : Spoken Language Processing and Biomedicine, 13th Annual Conference of the International Speech Communication Association, September 9-13, 2012, Portland, Oregon*, Portland: International Speech Communication Association, 2012, pp. 886–889.

[8] L. Meister, 'Vene aktsent eesti keeles: akustilise analüüsi tulemusi', *Eesti Rakenduslingvistika Ühingu aastaraamat*, vol. 2, no. 0, Art. no. 0, 2006, doi: 10.5128/ERYa2.10.

[9] J. E. Flege and O.-S. Bohn, 'The revised Speech Learning Model (SLM-r)', in *Second Language Speech Learning: Theoretical and Empirical Progress*, R. Wayland, Ed., Cambridge: Cambridge University Press, 2021, pp. 3–83. doi: doi:10.1017/9781108886901.002.

[10] C. T. Best and M. D. Tyler, 'Nonnative and second-language speech perception', in *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*, vol. 17, O.-S. Bohn and M. J. Munro, Eds., Amsterdam: John Benjamins Publishing Company, 2007, pp. 13–34. doi: 10.1075/lllt.17.07bes.

Cross-linguistic interference in the perception of L2 English fricatives by speakers of Chinese, Japanese, and Vietnamese

John Matthews [a], Takako Kawasaki [b], Kuniyoshi Tanaka [b]
[a] *Chuo University, Japan*
[b] *Hosei University, Japan*

Keywords — perception, L2 phonetics, fricatives, CLI, Asian languages

I. Background

It is widely known that L2 speech acquisition is influenced by the inventory of sounds in a learner's L1.[1],[2],[3] Kawasaki et al.[4] studied whether immersion in a study-abroad environment might induce a change in perceptual sensitivities to L2 segments that resemble familiar L1 segments but are nevertheless both articulatorily and acoustically distinct from them. They presented two groups of Japanese learners of English (JLEs), with and without study-abroad experience in an English-speaking country, with an identification task comprised of four different English fricatives in two vowel environments ([a] and [i]). Two of the fricatives do not occur in spoken Japanese ([f] and [θ]), and the two that do exist in Japanese ([s] and [ʃ]) are contrastive in some phonetic environments (e.g., before [a], [u], or [o]) but subject to phonological neutralization in others (i.e., before [i]). They analyzed the performance of the two groups tabulated in confusion matrices and found that the two groups differed in their susceptibility to perceptual confusion, which Kawasaki et al.[4] attribute to the increased exposure to target-language phonetic input available in an immersion environment. They further argue that this experience induces restructuring of a learner's perceptual map for speech sounds.

II. Experiment

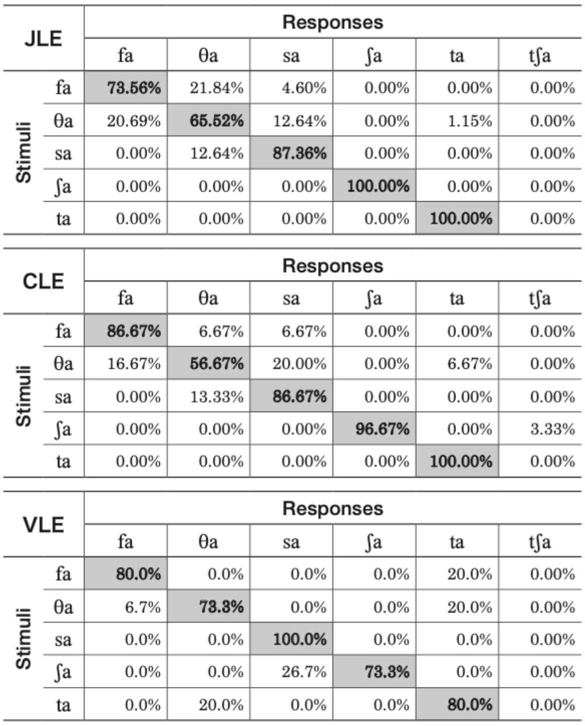
In this study we apply the Kawasaki et al.[4] approach to English learners whose native language is Vietnamese (VLEs) and compare their vulnerability to perceptual confusion with groups of Chinese learners of English (CLEs) as well as JLEs.[4],[5] All three of these languages contain contrastive inventories of fricatives, but they differ in their precise phonetic realizations.[6],[7] While they all include an alveolar fricative equal to English [s], none contains segments identical to [θ] or [ʃ], and Japanese further lacks [f]. If L2 learners experience perceptual confusions based on perceptual maps optimized for their L1 segmental inventory until sufficient exposure to L2 phonetic input can drive restructuring, then learners with different L1s should exhibit distinct perceptual confusion effects germane to their particular segmental inventories. We focus our investigation on post-alveolar fricatives, realized in English as palato-alveolar [ʃ] but in Japanese and Chinese as alveopalatal [ɕ]; and in Vietnamese retroflex [ʂ]. Our aim is to investigate the effects of familiarity with these L1 fricatives on perceptual sensitivities to L2 English fricatives.

III. Procedure

Native speakers of Vietnamese, Chinese, and Japanese performed an identification task with CV monosyllables containing one of four English fricatives [f], [θ], [s], [ʃ], or stop [t], and either the low vowel [a] or high vowel [i]. Concurrent with each audio stimulus was an array of English words displayed on a computer screen, each beginning with one of six obstruent spellings, corresponding to the four fricatives, the stop, and the affricate [ʧ] as well. For each trial they were then instructed to select the word that starts with the same sound as the audio stimulus they hear. Ten naturally produced items were presented three times at different signal-to-noise ratios using Millisecond Software's Inquisit.[8] We then computed confusion matrices (Figures 1 and 2) for each pair-wise combination of initial consonants in auditory stimuli and orthographic responses for each vowel environment separately.[9]

IV. Results

Results show that different native languages lead to different identification behavior with segments in a common second language. CLEs confuse [fi] and [θi] equally in both directions while VLEs display asymmetrical confusion, perceiving [fi] as [θi] but not vice versa. Accuracy for [ʃa] was lower among VLEs, who typically confuse it with [sa], than among either CLEs or JLEs. Contrastive [ɕa]–[sa] in Chinese and Japanese appears to lead speakers of those languages to substitute L1 [ɕ] for L2 [ʃ], having perceived the novel L2 segment as a familiar L1 segment through equivalence classification.[2] Vietnamese speakers are unencumbered by their own similar fricative, the retroflex [ʂ], which appears to be clearly differentiated from [ʃ]. Our analysis of these findings is based on the acoustic measurements of these segments when produced by speakers of the languages under investigation and on their distinct phonological behavior in each language.



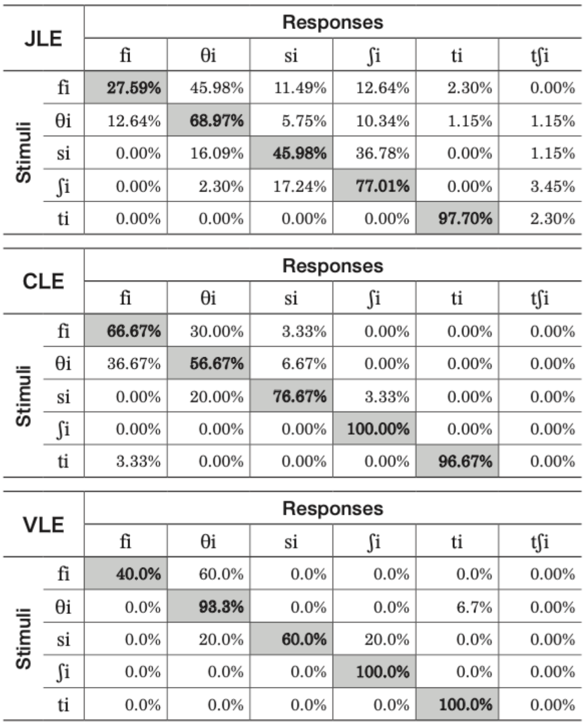
JLE		Responses					
		fa	θa	sa	ʃa	ta	tʃa
Stimuli	fa	**73.56%**	21.84%	4.60%	0.00%	0.00%	0.00%
	θa	20.69%	**65.52%**	12.64%	0.00%	1.15%	0.00%
	sa	0.00%	12.64%	**87.36%**	0.00%	0.00%	0.00%
	ʃa	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
	ta	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%

CLE		Responses					
		fa	θa	sa	ʃa	ta	tʃa
Stimuli	fa	**86.67%**	6.67%	6.67%	0.00%	0.00%	0.00%
	θa	16.67%	**56.67%**	20.00%	0.00%	6.67%	0.00%
	sa	0.00%	13.33%	**86.67%**	0.00%	0.00%	0.00%
	ʃa	0.00%	0.00%	0.00%	**96.67%**	0.00%	3.33%
	ta	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%

VLE		Responses					
		fa	θa	sa	ʃa	ta	tʃa
Stimuli	fa	**80.0%**	0.0%	0.0%	0.0%	20.0%	0.00%
	θa	6.7%	**73.3%**	0.0%	0.0%	20.0%	0.00%
	sa	0.0%	0.0%	**100.0%**	0.0%	0.0%	0.00%
	ʃa	0.0%	0.0%	26.7%	**73.3%**	0.0%	0.00%
	ta	0.0%	20.0%	0.0%	0.0%	**80.0%**	0.00%

J = Japanese, C = Chinese, V = Vietamese, LE = Learners of English

Fig. 1 Confusion matrices before [a]



JLE		Responses					
		fi	θi	si	ʃi	ti	tʃi
Stimuli	fi	**27.59%**	45.98%	11.49%	12.64%	2.30%	0.00%
	θi	12.64%	**68.97%**	5.75%	10.34%	1.15%	1.15%
	si	0.00%	16.09%	**45.98%**	36.78%	0.00%	1.15%
	ʃi	0.00%	2.30%	17.24%	**77.01%**	0.00%	3.45%
	ti	0.00%	0.00%	0.00%	0.00%	**97.70%**	2.30%

CLE		Responses					
		fi	θi	si	ʃi	ti	tʃi
Stimuli	fi	**66.67%**	30.00%	3.33%	0.00%	0.00%	0.00%
	θi	36.67%	**56.67%**	6.67%	0.00%	0.00%	0.00%
	si	0.00%	20.00%	**76.67%**	3.33%	0.00%	0.00%
	ʃi	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
	ti	3.33%	0.00%	0.00%	0.00%	**96.67%**	0.00%

VLE		Responses					
		fi	θi	si	ʃi	ti	tʃi
Stimuli	fi	**40.0%**	60.0%	0.0%	0.0%	0.0%	0.00%
	θi	0.0%	**93.3%**	0.0%	0.0%	6.7%	0.00%
	si	0.0%	20.0%	**60.0%**	20.0%	0.0%	0.00%
	ʃi	0.0%	0.0%	0.0%	**100.0%**	0.0%	0.00%
	ti	0.0%	0.0%	0.0%	0.0%	**100.0%**	0.00%

J = Japanese, C = Chinese, V = Vietamese, LE = Learners of English

Fig. 2 Confusion matrices before [i]

REFERENCES

[1] Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research*, pp. 171-204. Timonium, MD: York Press.

[2] Flege, J. E. (1995). Second language speech learning Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research*, pp. 233-277. Baltimore: York Press.

[3] Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, **39**: 456-466.

[4] Kawasaki, T., Tanaka, K., Takeuchi, M. and J. Matthews (2019). L2 shutoku-ni yoru musei masatsuon tikaku mappu-no saikoutiku [Restructuring the perceptual map for voiceless fricatives in L2 acquisition]. Proceedings of the 36th Annual Conference of the *Japan Society of Cognitive Science*, pp. 698-702.

[5] Kawasaki, T. and K. Tanaka (2021). L2 onseitikakumappu-ni okeru L1 mokuroku-no eikyou nihongo bogowasya-to tyuugokugo bogowasya-no eigo masatuon tikaku-no hikaku [Comparing the effects of L1 phonetic inventory on the perceptual map for L2 English fricatives among native speakers of Japanese and Chinese]. *Proceedings of the 38th Annual Conference of the Japan Society of Cognitive Science*, pp. 91-94.

[6] Hwa-Froelich, D., Hodson, B. W., and H. T. Edwards (2002) Characteristics of Vietnamese Phonology. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association* **11**(3): 264–73.

[7] Lee-Kim, S. (2014). Revisiting Mandarin 'apical vowels': An articulatory and acoustic study. *Journal of the International Phonetic Association*, **44**: 261-282.

[8] Inquisit 5 [Computer software]. (2016). Retrieved from https://www.millisecond.com.

[9] Miller, G. A., and P. E. Nicely (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**: 338–352.

Self-study of the MOOC *How to Teach Pronunciatio*n: Qualitative Method

Marta Nowacka
University of Rzeszow, Poland

Keywords — MOOC, How to Teach Pronunciation, self-study, teaching English phonetics, qualitative method

A plethora of research on teachers' beliefs on pronunciation teaching shows that teachers' own experience shapes their convictions and speech assessment. For example, pre-service more experienced teachers turn out to be more sceptical about pronunciation instruction and are harsher in the evaluation of accentedness than less experienced ones (Tsunemoto et al., 2023) [1]. Instructors with more training assign more value to pronunciation, set more pronunciation-related goals and eschew delaying a focus on this skill (Nagle et al., 2023) [2]. One of the recurrent findings in this research is that a large number of teachers feel unconfident and ill-prepared in their teaching as a result of insufficient or even a lack of training (Bay & Yuan, 2019 [3]; Couper, 2021 [4]; Jarosz, 2023 [5], Nguyen and Newton (2020) [6]). Couper (2021) shows that teachers report problems with determining priorities and setting goals for pronunciation instruction: native teachers admit to lacking knowledge of phonetics and phonology while outgroup teachers exhibit insecurity about their own pronunciation, which is also observed among outgroup teachers in Bay and Yuan (2019).[1] Nguyen and Newton (2020), and Jarosz (2023) revealed discrepancies between teachers' pronunciation beliefs and their classroom practice. The findings of the former research point to unplanned and reactive instruction, corrective feedback of segmental errors through recasts while those of the latter demonstrate that only a recent university graduate is more likely to integrate pronunciation instruction into all-skill learning and correct pronunciation mistakes mainly on word stress and accuracy, while more experienced teachers tend to avoid pronunciation teaching contrary to what they report believing. Therefore, the need for refresher pronunciation-oriented training programmes for in-service teachers is frequently signalled. Taking into consideration the fact that similar shortcomings in phonetic expertise have recently been reported by teachers worldwide, e.g. in Asia: Hong Kong (Bay & Yuan, 2019), Japan (Tsunemoto et al., 2023), Vietnam (Nguyen & Newton, 2020; Tran & Nguyen, 2020 [7]); Europe: Poland (Jarosz, 2023), Spain (Nagle et al., 2023); Oceania: New Zealand (Couper, 2021); South America: Uruguay (Couper, 2021), it is evident that there is a need for the application of clear guidelines and standards of pronunciation instruction in English language teaching institutions.

Undoubtedly, it is a duty of teacher educators to encourage future teachers to be positive about the teachability of L2 pronunciation by providing them with a sufficient training. One of the ways of learning pedagogical skills concerning pronunciation could be Language Fuel's (2023) [8] MOOC *How to Teach Pronunciation*. This course outlines the main principles in the realm of phonetics, suprasegmental and segmental aspects for accentedness and accuracy ratings, and it also targets communicative aspects of L2 speech such as fluency, intelligibility, and comprehensibility. MOOC stands for massive open and online course, which aims at a specific skill, including language (Bárcena and Martín-Monje, 2014) [9], allows for self-paced interactive study by means of discussions, user forums and immediate feedback. Nowacka (2023) [10] examined the impact of self-study of Rupp's (2018) [11] *MOOC Pronunciation in a Global World* on students' meta-awareness of English phonetics, at the beginning of a corresponding face-to-face university course, however, found no positive effects.

This paper aims to present qualitative results on the influence of Language Fuel's (2023) MOOC *How to Teach Pronunciation*, targeted at pre-service and in-service teachers of English, on students' phonetic expertise and preparation for future pronunciation teaching. This course follows a top-down approach to pronunciation teaching and covers four topics: overview of English phonetics (aspects, fluency and accuracy, principles and techniques of teaching pronunciation) and pronunciation at three levels: 1) the phrase and sentence: linking (CV, CC, VV), intonation (fall, rise, contrast/emphasis), pausing and speed; 2) the word (word stress and weak forms); 3) the sound (perception before production, consonants and vowels).

As regards the method, the participants were eighty-three first-year university Polish students of English. The experimental cohort (*n*=44) self-studied the course as an assignment, approximately one hour a week for a month. The students were tested weekly in the classroom by means of online forms, which included altogether 93 questions (54 close-ended, 4 justifications and 35 open-ended) on the Microsoft Teams platform. The pre-test checked the respondents' phonetic meta-competence and their knowledge about methodology of teaching selected aspects of English phonetics. The four successive tests corresponded to the above-mentioned MOOC's weekly topics. The post-test evaluated the MOOC's usefulness and attractiveness.

The quantitative results were discussed at the EPIP'8 conference in Santander, Spain in May 2024. This paper focuses on the qualitative part, in the form of open-ended questions (*n*=35), of which the largest number (*n*=23) concern teaching pronunciation,

[1] A term *outgroup*, a concept introduced in Tajfel and Turner's (1979) [12] social identity theory, is used as a substitute for *non-native* in contrast with *native*.

e.g., the choice of aspects for pronunciation practice, major principles in teaching pronunciation, techniques to develop muscle memory, types of lessons on pronunciation and embedded pronunciation instruction. Some of the enquiries ask directly about the ways of teaching such aspects as: intonation, pausing, linking, word stress, weak forms, sound distinctions (e.g., the difference between /e/ and /æ/, voiceless versus voiced consonants). Nine questions examine the respondents' know-how of English phonetics, e.g. main types, and rules for linking, e.g., "What is the rule behind linking in 'we are' and 'you are'? (Q.19.T2)" or coalescence, "What can final alveolar consonants /s/, /z/, /t/, /d/ change into in front of /j/? (Q.11.T0.)". The remaining three questions concern the attractiveness, usefulness of this tool and suggestions for improvement. Four additional justifications provide the reasons for recommending this tool but also for choosing the focus of a pronunciation-oriented lesson, either on sounds, units larger than sounds or both, as well as on fluency, accuracy, or both.

It is hypothesized that this introductory training on teaching phonetics should be beneficial to future English teachers as an initial examination of the results shows overall better phonetic meta-competence and greater awareness of pronunciation teaching principles for the experimental group. If proved effective, the said *MOOC* could help fill the gap in teachers' phonetic competence and expertise.

References

[1] A. Tsunemoto, P. Trofimovich, and S. Kennedy, "Pre-service teachers' beliefs about second language pronunciation teaching, their experience, and speech assessments," *Language Teaching Research* [Online], vol. 27, no. 1, pp. 115–136, 2023. Available: https://doi.org/10.1177/1362168820937273

[2] C. Nagle, R. Sachs, and G. Zárate-Sández, "Spanish teachers' beliefs on the usefulness of pronunciation knowledge, skills, and activities and their confidence in implementing them," *Language Teaching Research* [Online], vol. 27, no. 3, pp. 491–517, 2023. Available: https://doi.org/10.1177/1362168820957037

[3] B. Bai, and R. Yuan. "EFL teachers' beliefs and practices about pronunciation teaching," *ELT Journal* [Online], vol. 73, np. 2, pp. 134–143, April 2019. Available: https://doi.org/10.1093/elt/ccy040

[4] G. Couper, "Pronunciation Teaching Issues: Answering Teachers' Questions," *RELC Journal* [Online], vol. 52, no. 1, pp. 128–143, 2021. Available: https://doi.org/10.1177/0033688220964041

[5] A. Jarosz, "Exploring How Teachers' Pronunciation Beliefs Affect Their Classroom Practices," in *English Pronunciation Teaching: Theory, Practice and Research Findings*, V. G. Sardegna, and A. Jarosz, Eds. Bristol, Blue Ridge Summit: Multilingual Matters, 2023, pp. 168–184. Available: https://doi.org/10.21832/9781800410503-016

[6] L. T. Nguyen, and J. Newton, "Pronunciation Teaching in Tertiary EFL Classes: Vietnamese Teachers' Beliefs and Practices," *TESL-EJ* [Online], vol. 24, no.1, pp. 1–20, May 2020. Available: https://tesl-ej.org/wordpress/issues/volume24/ej93/ej93a2/

[7] D. P. T. Tran, and H. B. Nguyen, "EFL Teachers' Beliefs and Practices of Teaching Pronunciation in a Vietnamese Setting," *Universal Journal of Educational Research* [Online], vol. 8, no. 12, pp. 7022–7035, 2020. Available: DOI:10.13189/ujer.2020.081270.

[8] Language Fuel, *How to Teach Pronunciation* [MOOC], FutureLearn. 2023. Available: https://www.futurelearn.com/courses/how-to-teach-english-pronunciation

[9] E. Bárcena, and E. Martín-Monje, "Language MOOCs: An Emerging Field," in *Language MOOCs: Providing Learning, Transcending Boundaries*, E. Martín-Monje, and E. Bárcena, Eds. Warsaw: De Gruyter Open, 2014, pp. 1–15.

Available: https://www.degruyter.com/view/product/455678. Available: https://doi.org/10.2478/9783110420067.1

[10] M. Nowacka, "Self-study of the *MOOC English Pronunciation in a Global World*: metaphonetic awareness and English accent variation," *Research in Language*, vol. 21, no. 3, pp. 267–290, 2023. Available: DOI: 10.18778/1731-7533.21.3.04

[11] L. Rupp, *English Pronunciation in a Global World* [MOOC], FutureLearn. 2018. Available: https://www.futurelearn.com/courses/english-pronunciation

[12] H. Tajfel, and J. C. Turner, "An integrative theory of intergroup conflict," in *The social psychology of intergroup relations*, W. G. Austin, and S. Worchel, Eds. Monterey: Brooks/Cole, 1979, pp. 33–48.

Evaluating the effectiveness of asynchronous online pronunciation training in Spanish

Matthew Patience [a], Sonya Bird [b]
[a] *Florida State University, USA,*
[b] *University of Victoria, Canada*

***Keywords* —Pronunciation instruction, High Variability Phonetic Training, Shadowing, Asynchronous learning**

I. INTRODUCTION

It has been well established that pronunciation instruction is typically lacking in the language classroom [1], despite the fact that students are interested in learning and practicing their pronunciation (e.g., [2], [3]). One of the main reasons that pronunciation instruction is not included in the classroom is due to a perceived lack of time – teachers don't feel there is enough time in class to work on pronunciation [1]. One possible solution to this dilemma is to have students practice their pronunciation using online educational technology tools outside of class time. This ensures that students get the pronunciation practice that they need and want, but it avoids the necessity to spend substantial time on pronunciation training during class time. The goal of the present study was to investigate the effectiveness of two different types of online pronunciation practice: high variability phonetic training (HVPT) and shadowing. These two types of training were chosen because previous work has shown that they can be effective methods for improving pronunciation [4] – [7], and they are easily performed online. However, it has not yet been established if one type of training is more effective than the other.

HVPT is one of the more studied types of pronunciation instruction (see [4], [5] for reviews). Learners typically listen to minimal pairs (e.g., pick-peak), and have to select the word that they hear. Regarding shadowing, learners listen to and mimic the speech of a model. Shadowing has been studied less than HVPT, but studies [e.g., 6, 7] have found that it can also lead to improvements in pronunciation. Almost no work has examined these types of training for learners of Spanish – our aim was to extend these methods to Spanish pronunciation, with the intention of establishing if they are as useful for learning Spanish as they have been for other languages. In order to determine the effectiveness of these two types of training, we performed a pre and post-training analysis of individual sounds, in addition to more global measures of proficiency: accentedness, fluency, and comprehensibility. Our study was designed to answer the following questions:

RQ1: Is HVPT or shadowing a more effective training for improving Spanish pronunciation?

RQ2: Do different types of training (shadowing vs. HVPT) lead to different types of improvement?

Given that HVPT is focused on improving a speakers' perceptual system of specific contrasts, we expected the learners who completed the HVPT training to improve more than those who completed the shadowing training on the contrasts tested in the HVPT training ([ð, ɾ], [b, β], [g, ɣ], lexical stress; see Section II.B for details). On the other hand, we expected the learners who completed the shadowing training to show greater improvements on the paragraph reading task, which was analyzed based on improvements in accentedness, fluency, and comprehensibility.

II. METHODOLOGY

A. Participants

All students registered in three sections of intermediate Spanish at a North American university were included in the study (N = 74). Students were randomly assigned to one of the two types of pronunciation training.

B. Tasks and Stimuli

All students performed three tasks during the first week of classes (week 1): a paragraph reading, a free speech task (not discussed in this work), and a word reading task. The tasks were designed to determine the students' baseline pronunciation. In weeks 2-11, participants performed their weekly pronunciation practice – students were required to complete 8 of 10 of the weekly sessions, which lasted 15 minutes per week (total practice time = 120 – 150 minutes). In week 12, all students completed the same tasks as in week 1 – improvements in pronunciation were based on changes between week 1 and week 12. Students received 3% towards their final grade for completing the pronunciation practice. All work was done online through the university's LMS.

For the HVPT group, the sessions consisted of listening to sets of stimuli and deciding which sound/word they heard. Sets corresponded to contrasts between phonemes [ð, ɾ] (e.g., cero /seɾo/ 'zero' vs. cedo /seðo/ 'I give up), allophones [b, β] and [g, ɣ] (e.g., debe [deβe] vs. [debe] 'he/she/it must'), or lexical stress (e.g., número /ˈnu.me.ɾo/ 'number' vs. numero /nu.ˈme.ɾo/ 'I number' vs. numeró /nu.me.ˈɾo/ 'he/she/it numbered'). Allophonic contrasts are not typically included in HVPT tasks. However, for L1 English speakers, although there are few phonemic contrasts that are difficult in Spanish, some allophonic contrasts such as [β, b] and [g, ɣ] are also challenging [8], and lead to non-target realizations. One of the goals of this study was to determine whether HVPT can be effective for overcoming allophonic variation difficulty, by training students on the phonetic difference of Spanish compared to English intervocalic stop realizations. All three contrasts were produced by three native bilingual Spanish-English speakers. For the shadowing group, students listened to short recordings of native Spanish speakers from a variety of dialects. They first practiced repeating what they heard in the recording. When ready, they listened to the recording again, and shadowed the native speaker.

C. *Data Analysis*

Of the 74 students, 37 completed all requirements of the Spanish pronunciation practice. Of these 37, two were native speakers of other languages (Mandarin and French), and therefore not comparable to the other speakers. Consequently, we included in our analysis the data from the 35 native English speakers who completed all tasks. The paragraph readings were extracted and presented in random order to native Spanish speakers and near-native University Spanish instructors (N = 10). The judges rated the paragraph readings for accentedness, comprehensibility, and fluency (using a 9-point scale, with '1' reflecting the lowest score and '9' the highest score (e.g., native-like). The average score for each criteria was calculated and used to evaluate any improvements in pronunciation. The individual word productions were extracted and analyzed in Praat. Segments were marked as target-like or non-target-like according to whether the production matched the target sound's category (i.e., a bilabial approximant for [β], an alveolar tap for /ɾ/).

III. Results

A. *Paragraph reading*

The preliminary results reveal improvements across all three constructs for both types of training. In the HVPT group, we found increases of: 1.0 for accentedeness (4.61 -> 5.61), 0.3 for fluency (4.83 ->5.17), and 0.4 for comprehensibility (5.44 -> 5.83). For the shadowing group, we found slightly higher increases of 1.1 for accentedeness (4.88->5.94), 0.4 for fluency (5.12->5.47), and 0.4 for comprehensibility (5.76->6.18).

B. *Word production*

The preliminary results for the word production task also revealed improvements over time, with average accuracy rates improving by 22% (26% at T1 to 48% at T2). The improvements were observed in both groups, but were largest in the HVPT (26% to 54%) compared to shadowing group (26% to 41%). The improvement in the HVPT group was largest for the lexical stress contrast (40%), followed by the [ɾ, ð] (30%) and the stop-approximant contrasts (17%). In contrast, the improvements in the shadowing group were largest for the [ɾ, ð] contrast (25%), followed by the lexical stress (10%) and stop-approximant contrasts (8%).

IV. Discussion

Overall, our results suggest that both types of training are effective for improving pronunciation. However, future work would need to compare improvements in pronunciation over the course of a semester to a group that did not complete the training to better understand the general extent of the improvements, and whether they were due to training and not simply due to additional classroom experience speaking Spanish. With respect to our research questions, our results suggest that neither method is categorically better than the other. It depends on whether the goal is to improve learners' competence with specific contrasts (for which the HVPT training is more effective), or with more global measures (for which the shadowing training is most effective). The type of training used should therefore be considered in tandem with the specific goals of the curriculum related to pronunciation. Future work should determine at which point in the curriculum each training is most beneficial to learners.

V. References

[1] T. M. Derwing and M. J. Munro, *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins, 2015.

[2] N. Edo-Marzá, "Pronunciation and comprehension of oral English in the English as a foreign language class: Key aspects, students' perceptions and proposals", *Journal of Language Teaching and Research,* vol. 5, no. 2, pp. 262–273, Mar. 2014.

[3] L. T. Nguyen, B. P. Hung, U. T. T. Duong, and T. T. Le, "Teachers' and learners' beliefs about pronunciation instruction in tertiary English as a foreign language education", *Front. Psychol.*, vol. 12, Aug. 2021.

[4] R. I. Thomson, 'High variability [pronunciation] training (HVPT) A proven technique about which every language teacher and learner ought to know", *J. Second Lang. Pronunciation,* vol. 4, no. 2, pp. 208–231, Dec. 2018.

[5] T. A. Barriuso and R. Hayes-Harb, "High Variability Phonetic Training as a Bridge from Research to Practice", *CATESOL Journal,* vol. 30, no. 1, pp. 177–194, 2018.

[6] J. A. Foote and K. McDonough, "Using shadowing with mobile technology to improve L2 pronunciation", *J. Second Lang. Pronunciation,* vol. 3, no. 1, pp. 34–56, Apr. 2017.

[7] Y. Shao, K. Saito, and A. Tierney, "How does having a good ear promote instructed second language pronunciation development? Roles of domain-general auditory processing in choral repetition training", *TESOL Q.*, Jan. 2022.

[8] M. Patience, "Articulatory Difficulty in L2 Spanish," Ph.D Dissertation, Dept. Hisp. Linguistics, Univ. of Toronto, Toronto, Canada, 2022.

Language transfer and articulatory conflict in advanced L1 Spanish learners of English

Joaquín Romero
Universitat Rovira i Virgili, Spain

Keywords — articulatory conflict, tongue-tip consonants, language transfer

I. Introduction

One of the major factors in the acquisition of a foreign language's phonology is the transfer from the speaker's L1 [1]. Both quantity and quality differences between the sound inventories of the two languages, as well as other aspects such as syllable structure [2], can determine whether transfer has a positive or negative impact on the acquisition of the L2 system, though it is more commonly the effects of negative transfer that have been the focus of research in pronunciation learning. While much of the work done on the effect of transfer in pronunciation learning, especially in the area of vowels, is limited to phonemic transfer, the interplay between the phonemic and the subphonemic (allophonic) levels can cause additional difficulties when it comes to acquiring a foreign language's sound system. Thus, a phonemic distinction such as /d/ vs. /ð/ in English is problematic for Spanish speakers, who will commonly identify them with the Spanish allophones of /d/ [d] and [ð̞], respectively and, consequently, apply the Spanish allophonic rules that determine that [d] should be used in absolute initial position, after nasals and /l/, while [ð̞] is used elsewhere. This results in common mispronunciations such as [dɛr] for English /ðɛr/ *there* or [əˈð̞ɔr] for English /əˈdɔr/ *adore*. The fact that English /d/ is often articulated as an alveolar flap [ɾ] in General American, as in [ˈhɛɾɚ] *header*, adds another level of complexity for speakers of Spanish, whose consonantal inventory includes a nearly identical sound, often referred to as 'single r', but at the phonemic level, as in /ˈkoɾo/ 'coro' *choir*, which contrast with both a trill or 'double r' as in /ˈkoro/ 'corro' *circle* and the continuant allophone of /d/ [ð̞] as in /koð̞o/ 'codo' *elbow*. An additional factor, the different articulatory nature of Spanish rhotics /r/ and /ɾ/ vs. American English rhotics [ɹ] and [ɝ]/[ɚ] results in a degree of complexity that often leads to negative transfer.

The complex relationship between these crosslinguistic phonemic and allophonic distinctions is aggravated by the fact that the sounds involved are all coronal and, therefore, share the same basic articulator, the tongue-tip/blade. Coronal consonants have long been considered special both in phonological and phonetic terms [3], among other reasons because the number of distinct place and manner of articulation distinctions that can be produced by the same general articulator far exceeds what other articulators (lips, tongue dorsum) can do [4]. From apicodental Spanish [ð̞] to postalveolar or retroflex English /ɹ/, the variety and subtlety of articulations in either language is very high and can be further complicated by coarticulatory and assimilatory phenomena. Therefore, in situations where more than one of these sounds appear in close proximity of each other, the potential for articulatory conflict and, consequently, mispronunciations increases significantly.

The current study investigates the production of English /ð/, /ɹ/ and [ɾ] by advanced L1 Spanish learners of American English in contexts where these sounds appear within the same word, as in *parody*, *moderate*, *order*, *further*, etc. Impressionistic observations indicate a considerable degree of difficulty in achieving the correct articulatory targets, which is in contrast with the relative ease and accuracy with which the same sounds are produced in isolation, as in *other*, *story*, *bedding*, etc.

II. Method

The subjects in the experiment were 37 third-year university students majoring in English and with a C1 or C2 overall language level. Prior to the experiment, all participants had taken a two-semester English phonetics and phonology course with a large pronunciation component and intensive practice in phonemic and allophonic transcription. Many of these students were expected to go on to become English teachers at different levels of the educational system after graduation. Participants were recorded in a sound-proof booth reading a set of 14 English sentences that included instances of /ð/, /ɹ/-/ɝ/ and [ɾ] in real words and in a variety of vocalic and prosodic contexts, both in isolation (one single sound per word) and in combinations (two sounds per word). In those cases where more than one sound appeared in the same word, the design also controlled, whenever possible, for order, with a special focus in distinguishing between [ɹ]V[ɾ] and [ɾ]V[ɹ] sequences, as in *Florida* vs. *federal*. The different contexts are shown in Table I below. The words were analyzed acoustically in Praat and information was obtained for the target consonants in terms of relative intensity (intensity difference between the consonant and the adjacent vowel) and consonant duration. These measures were taken as indicators of the nature of the consonants, such that [ɾ] was expected to correlate with high relative intensity and short duration, /ð/ with medium relative intensity and long duration, and /ɹ/ with low relative intensity and medium duration. Mixed-models analyses were performed for each dependent variable (relative intensity and duration) and with context (single or double) and, for a subset of the data, order ([ɹ]V[ɾ] vs. [ɾ]V[ɹ]) as fixed factors; subject and word were included as random factors.

III. Results

The acoustic analysis of the data revealed numerous instances of mispronunciations as the result of negative transfer from Spanish to English. While the order factor did not produce significant differences for either relative intensity and duration, both dependent variables were significantly different for the context factor. These results illustrate that, while most speakers had little difficulty articulating /ð/, /ɹ/ and [ɾ] accurately when these consonants appeared only once in the word (*other*, story, *bedding*, for example) the presence of two of them (*parody*, *federal*) significantly increased the number of inaccurate pronunciations, with numerous instances of [ɾ] instead of /ɹ/ and /ð/ instead of [ɾ] (*federal* being pronounced [ˈfɛðəɾəl], and substitutions of /ɹ/ for [ɾ] ([ˈbæɹəɹi] for *battery*). Of particular difficulty were /ɹ/-/ɝ/+[ɾ] sequences as in *order* and *murder*, where the continguous production of a retroflex /ɹ/-/ɝ/ and alveolar [ɾ] proved articulatorily very challenging. In these instances, several different strategies were used to compensate for the difficulty, including complete deletion of [ɾ] ([ˈɔɹɚ]) in the word *order*, substitution of /d/-/ð/ for [ɾ] ([ˈɔɹdɚ]-[ˈɔɹðɚ]) or a pronunciation more in accordance with Spanish phonology such as [ˈɔɾðɚ], where English /ɹ/ and [ɾ] are directly replaced by Spanish /ɾ/ and [ð̞], respectively. Fig. 1 illustrates some of these mispronunciations.

The results of this study give support to the notion that sound transfer from an L1 to an L2 is a complex phenomenon that operates at different levels of the system, both phonemically and phonetically, but also has implications at the phonotactic level, especially when the productions of contiguous sounds result in sequences that require conflicting activation of the same articulator. Even for advanced learners of English, the need to distinguish between phonemic and/or allophonic representations in the two languages, coupled with the simultaneous physical demands of articulatory configurations that are only subtly different, can prove to be too difficult, leading to inaccuracies in their productions. While many of these mispronunciations may not necessarily result in communication problems, they can contribute to the perception of accented speech, which can be detrimental to the interests of future language teachers who are training to become accurate language models.

TABLE I. Contexts included in the design with example words

one consonant in word				
[ð]	[ɹ]	[ɾ]		
*o**th**er*	*sto**r**y*	*be**dd**ing*		
two non-consecutive consonants in word				
[ɹ] V [ɾ]	[ɾ] V [ɹ]			
*Flo**r**i**d**a*	*fe**d**e**r**al*			
two consecutive consonants / vowel+consonant / consonant+vowel				
[ɹ]+[ɾ]	[ɹ]+[ð]	[ɝ]+[ɾ]	[ɾ]+[ɚ]	[ɹ]+[ɚ]
*bo**rd**er*	*fu**rth**er*	*mu**rd**er*	*la**dder***	*manufactu**rer***

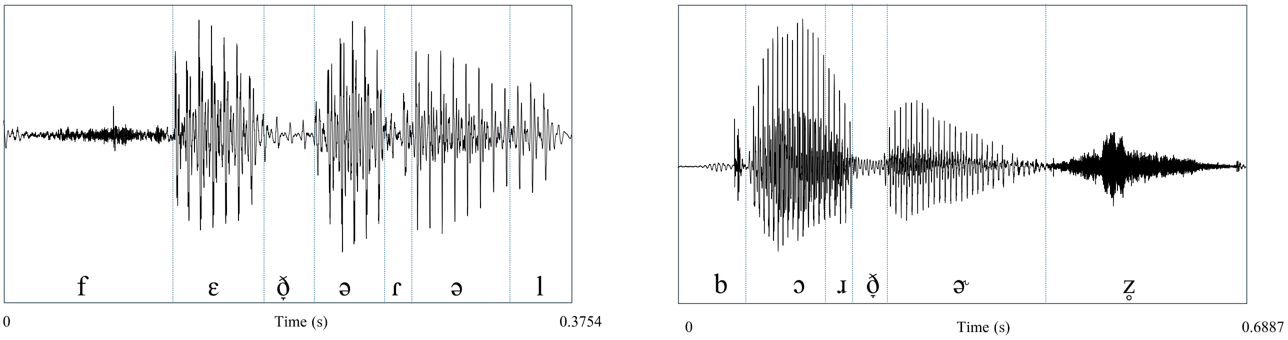


Fig. 1. Example of *federal* (left) and *borders* (right), illustrating inaccurate productions by two different subjects.

References

[1] B. Hammarberg, "Conditions on transfer in phonology," in Second Language Speech: Structure and Process, A. James, & J. Leather, Eds. Berlin: Mouton de Gruyter, 1997, pp. 161-180.

[2] I. Stockman and E. Pluut, "Segment composition as a factor in the syllabification errors of second-language speakers," Language Learning, 42 (1), 21-45, 1992.

[3] P. Keating, "Coronal places of articulation," in The Special Status of Coronals: Internal and External Evidence, C. Paradis, & J-F Prunet, Eds. Academic press, 1991, pp. 29-48.

[4] B. Gick, I. Wilson and D. Derrick, Articulatory Phonetics. John Wiley & Sons, 2012.

Extralinguistic and linguistic factors in single-word ASR of German atypical speech

Eugenia Rykova [a, b, c], Mathias Walther[a]
[a] *Technical University of Applied Sciences TH Wildau, Germany*
[b] *University of Eastern Finland, Finland*
[c] *Catholic University Eichstätt-Ingolstadt, Germany*

Keywords — ASR, atypical speech, digital health

I. Introduction

Automatic speech recognition (ASR) has become part of many everyday services, including digital health. In particular, speech and language therapy (SLT) can benefit considerably from ASR usage – for example, when in-person therapy is supplemented with digital therapy solutions used independently [1]. However, commercial systems with excellent results in applications for typical speakers demonstrate poor performance on the material of impaired speech [2].

Aphasia is a relatively common language disorder that occurs after completed language development because of a brain damage, which in 80% of the cases is caused by a stroke [3]. Generally deteriorated condition of speech, high variability among speakers, and insufficiency of data make it difficult to use ASR for aphasic speech. Imprecise articulation and phonemic structure distortions are mostly inconsistent and unpredictable, which hinders error modelling [4]. Aphasia can be also comorbid with motor speech disorders, which bring further disfluencies and decrease speech intelligibility [5]. Besides, age is a risk factor for stroke and aphasia, and older age per se can influences speech production (e.g., slower speech rate) [6]. Changes in acoustic features are reflected in poorer ASR performance for older speakers, which might be more drastic for female voices [7].

To the best of authors' knowledge, the option of including ASR for automated feedback in digital SLT solutions for German-speaking people with aphasia (PWA) is currently under research [8-10] but is not offered to the users yet (cf. [4]). For the current project [10], four open-source ASR solutions have been selected as the most suitable for PWA's speech recognition. Selection procedure was a complex process based on models' performance on atypical speech, both discussed in this paper and two small PWA's speech corpora [11]. In the absence of adequate data from PWA, test material from other corpora with atypical speech was considered for the purposes of the present evaluation, namely speech of adult cochlear implants (CI) users, which can be characterized by decreased vowel exactness and precision of articulators' movements [12], and speech under intoxicated condition, characterized by decreased speech rate and weakened speech motor control [13]. Respective changes are both captured by human perception and reflected in ASR rates. This paper presents the analysis of the models' robustness to extralinguistic factors and effects of linguistic features on single-word ASR rates.

II. Materials and methods

Four open-source ASR models [14-17] were tested with the help of selected material from ALC [18] and CI corpora [19]: words segmented out of the tongue-twisting lists uttered by sober (NA_words) and intoxicated speakers (A_words), and words segmented out of the sentences uttered by CI users (CI_words) and normal-hearing speakers (NORM_words).

Character Error Rate (CER) and HITS measurement (the number of precisely recognized words) were used for evaluation. Recognition results were analysed as influenced by atypicality, demographics, and linguistic and speech factors: duration of the segment, (in seconds), length of the segment in syllables, and speech rate (syllables/second, syll/s) (for reference values see [20]).

Statistical analyses used analysis of variance (ANOVA) with a post-hoc Tukey's Honest Significant Difference (Tukey's HSD) test, Pearson and Spearman correlation tests, pairwise Wilcoxon signed-rank test, and decision (regression) trees with ANOVA as a fit method [21].

III. Results & Discussion

A graphical representation of the robustness to demographic factors can be seen in Figure 1. The absence of a statistically significant difference in ANOVA and post-hoc Tukey's HSD tests (p-value > 0.05) between CER values of demographic groups is understood under robustness. The significant differences and the corresponding p-values are marked in orange. Mfleck and oliver9 are robust to gender, age, and their interaction in the experiments with NA_words. Jonatas53 is robust to gender, but not to age. Tukey's HSD shows that the underlying difference is CER values for the MO group, which are significantly higher than CER values for both FY and MY groups. In the experiments with A_words, jonatas53, mfleck, and oliver9 are robust to gender, but show significantly higher CER values for the older group as in the earlier study by [7]. With both datasets, nvidia2 is robust to age, but shows significantly higher CER values for the female group, for A_words, in particular, the difference between FY and MY groups is significant. The oliver9 model is robust to age in the experiments with NORM_words, and jonatas53 and mfleck are robust to age in the experiments with CI_words. In the rest of the comparisons, the CER values for younger speakers are significantly higher (cf. [7]).

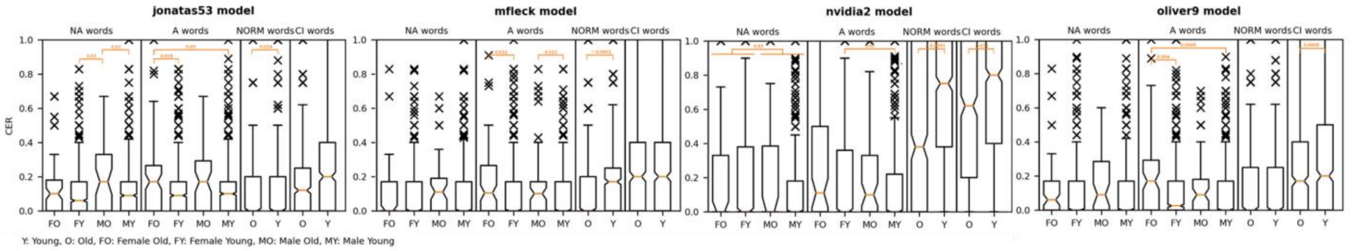


Fig. 1. Robustness of the four selected models to gender and age.

In CI corpus, the speech rate of younger speakers is greater than that of the older speakers, and the duration of the same words is longer when uttered by the latter, which could the underlying reason for the differences in CER values between age groups. Excluding nvidia2 (the weakest model for these datasets), ASR systems generally perform better on audio samples of greater duration (greater than 0.27 s) in combination with speech rates: the lower threshold for normal hearing speakers is 2.1 syll/s, and for the CI users it is 4.3 syll/s. Two-syllable words are recognized with lower CER values on average and are more likely to be recognized precisely, but they should not be uttered too quickly or too slowly.

The experiments with the three models (excluding nvidia2) on words from the ALC corpus confirm that for better single-word recognition the audio samples should be not too short and not too slowly pronounced: duration ≥ 0.44 s and speech rate ≥ 2.9 syll/s (values comparable with NORM_words). These datasets contain much longer words than the CI corpus, and there are more relatively shorter words among those that are precisely recognized.

Summarizing the above, one can expect that words of moderate length will be recognized better than one-syllable or long words. Speech samples uttered at the rates below average of the corresponding datasets, which are lower than intended "very slow" [20], are more likely to produce higher CER values. Faster speech rates – the maximum values in ALC and CI corpora are higher than intended "very fast" [20] – also lower the recognition quality. In the experiments with different datasets, recognition results show inconsistent, and sometimes contrasting, influence of the demographic factors, which might be a consequence of interaction with speech rate. In those datasets, where older speakers speak slower than the younger ones, the CER values of the former are also lower. In those with no difference, the CER values for older speakers are higher.

References

[1] F. Hönig & E. Nöth, „Automatic speech processing in speech therapy," in *Neue Technologien in der Sprachtherapie,* K. Bilda, J. Mühlhaus, and U. Ritterfeld, Eds. Georg Thieme Verlag KG, 2016, pp. 173-184.

[2] J. Green et al., „Automatic Speech Recognition of Disordered Speech: Personalized models outperforming human listeners on short phrases," *Proc. Interspeech 2021*, pp. 4778-4782, 2021.

[3] A. Wiehage & J. Heide, „*Aphasia: information for the affected and relatives*". German Federal Association of Academic Speech Therapists, 2016.

[4] A. Abad et al., „Automatic word naming recognition for an on-line aphasia treatment system," *Comput Speech Lang.*, 27, pp. 1235-1248, 2013.

[5] C.D. Qualls, „Neurogenic disorders of speech, language, cognition-communication, and swallowing," in *Communication Disorders in Multicultural and International Populations, 4th edition*, D. Battle, Ed. Mosby, 2012, pp. 148-163.

[6] L. Johnson et al., „Predictors beyond the lesion: Health and demographic factors associated with aphasia severity," *Cortex,* 154, pp. 375-389, 2022.

[7] R. Vipperla, S. Renals, and J. Frankel, „Longitudinal study of ASR performance on ageing voices," *Proc. Interspeech 2008*, pp. 2550–2553, 2008.

[8] J. Heide et al., "Improving lexical retrieval with LingoTalk: an app-based, self-administered treatment for clients with aphasia," *Front Commun*, 2023.

[9] Y. Lin et al., "*Automatic language assessment with artificial intelligence. for the neolexon aphasia app*," Poster session presentation at Sprachtherapie aktuell: Forschung - Wissen – Transfer 9(1): XXXIV. Workshop Klinische Linguistik e2022-11, April 2022.

[10] Martin-Luther-Universität Halle-Wittenberg, „aphaDIGITAL," Accessed: May 14, 2024. [Online]. Available: https://aphadigital.sprechwiss.uni-halle.de/

[11] E. Rykova & M. Walther, „Evaluation of German ASR solutions for speech and language therapy support of people with aphasia," submitted for publication.

[12] T. Arias-Vergara et al., „Adult cochlear implant users versus typical hearing persons: An automatic analysis of acoustic-prosodic parameters," *J Speech Lang Hear Res*., 65(12), 2022, pp. 4623-4636.

[13] E. Tisljár-Szabó et al., „The effect of alcohol on speech production,". *J Psycholinguist Res,* 43, 2014, pp. 737–748.

[14] M. Fleck, "*Wav2vec2-large-xls-r-300m-german-with-lm*". Accessed: September 12, 2022. [Online]. Available: bit.ly/3VOGZQQ.

[15] J. Grosman, "*Fine-tuned XLSR-53 large model for speech recognition in German*". Accessed: September 12, 2022. [Online]. Available: bit.ly/3vE1NQL.

[16] O. Guhr, "*wav2vec2-large-xlsr-53-german-cv9*". Accessed: September 12, 2022. [Online]. Available: bit.ly/3VOgH1a.

[17] NVIDIA, "*NVIDIA Conformer-Transducer Large (de)*". Accessed: September 12, 2022. [Online]. Available: bit.ly/3vFtpot.

[18] F. Schiel et al., „ALC: Alcohol Language Corpus," *Proc. LREC'08,* 2008.

[19] V. Neumeyer, „*Phonetic examination of the CI users' articulation,"* M.S. thesis, Ludwig-Maximilians-Universität, München, Germany, 2009.

[20] V. Dellwo, E. Ferragne, and F. Pellegrino, "The perception of intended speech rate in English, French, and German by French speakers," *Proceedings of the 3rd International Conference of Speech Prosody, Dresden, Deutschland*, 2006, pp.101-104.

[21] T.M. Therneau & E.J. Atkinson, "*An introduction to recursive partitioning using the RPART and routines,"* Technical report. Mayo Foundation, 2022.

Quantity perception among Estonian kindergarten children with developmental language disorder

Liis Themas[a,b], Pärtel Lippus[a], Marika Padrik[c] , Kairi Kreegipuu[b]
[a] *Institute of Estonian and General Linguistics, Estonia,*
[b] *Institute of Psychology, Estonia,*
[c] *Institute of Educational Science, Estonia*

Keywords — estonian, word prosody, quantity discrimination, child development, developmental language disorder

I. INTRODUCTION

Previous studies have shown delayed language processing in children with developmental language disorder (DLD) [1], including difficulties in prosody perception and other prosodic skills [2-9]. Estonian's unique three-way quantity distinction in prosody combines tonal and durational components [11]. There is scarce data about the perception of the three-way quantity system by children.

The distinction of short (Q1), long (Q2) and overlong (Q3) quantity degrees in Estonian is marked firstly by the duration of the stressed and the unstressed syllables and secondly by the pitch contour. The temporal pattern can be described by a reverse relation between the stressed and the unstressed syllables of a left-headed disyllabic foot, meaning that the stressed syllable is longer in the case of higher degrees of quantity while the unstressed syllable is compensatorily shortened. [11]

Developmental language disorder is a heterogeneous category that encompasses a wide range of problems [12]. Impairment can occur on some or all levels of speech perception and/or production [13]. There is some behavioural evidence that children with DLD have difficulties in distinguishing between the quantity degrees, pronouncing the quantities [14] and marking them correctly in orthography [15].

II. METHOD

This study explores the differences in perceiving Estonian's short (Q1), long (Q2), and overlong (Q3) quantity distinctions between children with DLD and typically developing (TD) peers. We examined children aged 4.6-6.5 years (DLD group N=25, TD control group N=25) using psychometric testing, sleep-EEG, auditory event-related potentials measure and computerized behavioural tasks. The first behavioural task being a quantity discrimination task, where children heard a train of a same word and had to press a button when the quantity of that word changed. The second task was a lexical decision task with aiding pictures. The participant saw a picture and heard eighter a word or a pseudoword, which were created by changing the quantity degree of the real word. The subject's task was to press the button if they hear a real word corresponding to the picture.

III. RESULTS

Here we present the behavioural data of the two phases of our longitudinal study. As we anticipated considerable variation at the individual level, encountered missing data, and because the two tasks measured a similar ability, a linear mixed model was used to analyze the data from the first phase. The dependet variable was the ratio of correct to incorrect button presses from both tasks and the independent variables chosen with previous correlational analysis were: age, sex, group, overall language ability and non-verbal intelligence. A random effect was included to account for variability at the participant level. The optimal model, as determined by model fit indices, incorporated an interaction between group and age. The model coefficients indicate significant difference between groups. More specifically, that the mean score for the DLD group on the behavioral discrimination measure is estimated to be 1.5 units higher than that of the EK group. However, in this context, a lower score signifies better performance. The impact of age varies depending on the group to which an individual belongs. Specifically, in the DLD group, age has a more pronounced negative effect on the quantity discrimination measure compared to the TD group. This may indicate that quantity discrimination is still developing in the DLD group. In contrast, in the TD group, this skill appears more stable across different ages and seems to be already acquired. This result may also reflect the development of other cognitive functions (e.g., sustaining attention) that are necessary to perform these kinds of behavioral tasks and may not directly reflect the ability to discriminate between quantities.

An analysis of the types of errors across the tasks revealed that both groups struggled with discriminating between the second and third quantities. However, the DLD group made significantly more errors than the TD group in this regard. Additionally, the DLD group also had nearly the same number of errors when differentiating between the first and second quantities—a task their typically developing peers accomplished with ease. While fewer errors occurred in discriminating between the first and third quantities, the difference between groups remained significant, with the DLD group making more mistakes.

The data of the second phase of the longitudinal study is still being collected therefore the result are not yet presented but will be by the time of the conference.

REFERENCES

[1] T. Kujala, and M. Leminen, “Low-level neural auditory discrimination dysfunctions in specific language impairment - A review on mismatch negativity findings,” Dev Cog Neurosci, vol. 28, pp. 65–75, 2017.

[2] C. Weber, A. Hahne, M. Friedrich, and A. D. Friederici, “Reduced stress pattern discrimination in 5-month-olds as a marker of risk for later language impairment: Neurophysiologial evidence” Cogn Brain Res, vol. 25(1), pp. 180–187, 2005.

[3] C. Cantiani, V. Riva, C. Piazza, R. Bettoni, M. Molteni, N. Choudhury, C. Marino, and A. A. Benasich, “Auditory discrimination predicts linguistic outcome in Italian infants with and without familial risk for language learning impairment,” Dev Cog Neurosci, vol 20, pp 23-34, 2016.

[4] H. Datta, V. L. Shafer, M. L. Morr, D. Kurtzberg, and R. G. Schwartz, “Electrophysiological Indices of Discrimination of Long-Duration, Phonetically Similar Vowels in Children With Typical and Atypical Language Development,” J of Speech, Lang, and Hearing Res, vol. 53(3), pp. 757–777, 2010.

[5] Y-Y. Cheng, H-C. Wu, H-Y. Shih, P-W. Yeh, H-L. Yen, and C-Y. Lee, “Deficits in Processing of Lexical Tones in Mandarin-Speaking Children With Developmental Language Disorder: Electrophysiological Evidence,” J of Speech, Lang, and Hearing Res, vol. 64(4), pp 1176-1188, 2021.

[6] N. Calet, M. Á. Martín-Peregrina, G. Jiménez-Fernández, and P. Martínez-Castilla, "Prosodic skills of Spanish-speaking children with developmental language disorder,” Int J Lang Commun Disord, vol. 56, pp. 784–796, 2021.

[7] R. Cumming, A., Wilson, and U. Goswami, “Basic auditory processing and sensitivity to prosodic structure in children with specific language impairments: A new look at a perceptual hypothesis,” Front Psychol, vol 6, 2015.

[8] L. B. Leonard, and J. B. Kueser, “Five overarching factors central to grammatical learning and treatment in children with developmental language disorder.” Int J Lang Commun Disord, vol. 54(3), pp. 347–361, 2019.

[9] S. Richards, and U. Goswami, “Impaired Recognition of Metrical and Syntactic Boundaries in Children with Developmental Language Disorders,” Brain Sci, vol 9(2), pp 33, 2019.

[10] S. Sundström, B. Lyxell, and C. Samuelsson, “Prosodic aspects of repetition in Swedish-speaking children with developmental language disorder,” Int J of Speech-Lang Pathology, vol. 21(6), pp. 623–634, 2019.

[11] P. Lippus, K. Pajusalu, and J. Allik, “The tonal component of Estonian quantity in native and non-native perception,” J of Phonetics, vol. 37(4), pp. 388–396, 2009.

[12] D. V. M. Bishop, M. J. Snowling, P. A. Thompson, T. Greenhalgh, and the CATALISE-2 consortium, "Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology,” J of Child Psychol and Psychiatry, vol. 58(10), pp. 1068–1080, 2017.

[13] L. B. Leonard, Children with Specific Language Impairment. 3rd ed. The MIT Press, 2014.

[14] M. Padrik, M. & M. Hallap, Kommunikatsioonipuuded lastel ja täiaksvanutel: märkamine, hindamine ja teraapia. Tartu Ülikooli Kirjastus. 2016.

[15] K. Karlep, Emakeele abiõpe I. Tartu Ülikooli kirjastus, 1999.

Communicating Dominance through Prosody in Deaf Speakers of Taiwan Mandarin

Tsung-Lun Alan Wan
National Yang Ming Chiao Tung University
Tla.wan@nycu.edu.tw

Keywords —clinical sociolinguistics, prosody, Mandarin, tone language, vowel shortening, pitch

Deaf speakers engage with prosody in distinct ways compared to their hearing counterparts especially in the frequency domain because variation in F0 is less accessible to many deaf individuals. For instance, deaf children with cochlear implants (CIs) exhibit smaller acoustic contrasts between emotions such as happiness and sadness in voice pitch, compared to hearing adults and deaf individuals who acquired deafness later in life [1]. Instead, deaf children seem to compensate for the lack of F0 variation by utilizing other prosodic cues, such as intensity [2]. A study on discourse prosody has further revealed that deaf children with CIs rely on temporal cues to express prosody, such as vowel length [3]. While deaf individuals may find pitch height more difficult to utilize, they can employ other acoustic cues like intensity and duration to realize prosody. This study argues that this is true also in sociolinguistic prosody, among deaf adults who speak Taiwan Mandarin.

Prosody plays a crucial role in social interaction. Pitch has been a well-studied variable in the research on sociolinguistic prosody, partly in response to the theory of frequency code [4]. Frequency code argues that, evolutionally, human beings associate high pitch and smallness, for animals with small sizes usually have higher pitch. However, sociophoneticians have reported that different cultures attribute varying social meanings derived from size to pitch. For instance, African American women employ a high pitch to communicate resistance and powerfulness [5]; body-building trainers speak faster and employ higher mean pitch than yoga instructors, indicating the former acoustic cues communicate a higher energy [6]. Research on politeness prosody also reveals that high pitch is not cross-culturally linked to a high level of politeness [7], [8]. In Mandarin spoken in China, speech rate (slower speech), rather than pitch height, leads to a perception of increased politeness [9]. The previous studies were primarily focused on deaf children, laying the emphasis on the potential effects of early cochlear implantation or hearing aid fitting on prosody development, often driven by clinical implications for early intervention. Limited attention has been given to what happens after deaf children transition to adulthood. This study explores how deaf adults engage with prosody to communicate social dominance.

This study recruited fourteen deaf adults to take part in a role-playing task (average age = 21.14). All the participants were prelingually deafened and orally educated. Gender is not balanced; 10 identified as women and 4 identified as men. The role-play task consists of six sentences, each with a designated role. The participants read aloud the sentence as if it's their line, after a practice trial. There are six roles: a customer at a bubble tea shop, a passenger at a taxi, a boss, a barber, a waiter, and a tenant looking for a flat. The first three are the target of this study and are expressed as declarative sentences. The last three serve as fillers, taking the form of interrogative sentences. The speech act performed by the customer is to order a bubble tea; the passenger informs the taxi driver their destination, and the boss asks their subordinate to submit a report. The first two can be grouped as a request; both roles are referred to as "the customer". The last should be seen as an order, exerting more dominance.

The participants also took part in a sentence-reading task, where they were required to read aloud seven sentences without any designated role. Compared with the role-playing task, the first, second, and sixth sentences from the sentence-reading task are used as a control group.

This study looks at two linguistic variables: pitch and vowel duration. Tokens with creakiness and receive tracking errors were removed. Duration is log-transformed, for humans perceive the difference in duration in a logarithmic way. For each syllable, only three F0 measurements were obtained: the first available measurement, the temporal-midpoint measurement, and the last measurement. These three values were averaged; each syllable (lexical tone unit) occupies one datapoint. The F0 value was transformed into semitone, with 100 hertz as the reference value.

Linear mixed effects models were applied. To exclude the possibility that the difference between the customer and the boss stems from a simple sentence order effect, the models explore the interaction between the position of sentences and the task. The position of sentences is coded as either the beginning or end of the task. For the sentence-reading task, they refer to the first two sentences, or the sixth sentence. In the role-playing task, they individually refer to the customer utterances or the boss utterances. If there is a significant interaction, it indicates that the role effect observed in the role-playing task is not simply driven by the order of sentences in the task.

Linguistic factors included lexical tone category (four level: high-level, rising, mid-falling, and high-falling), and position in an intonational phrase, defined by pauses (continuous: 1, 2, …). By-speaker random slopes of all the fixed effects were included. The pitch model is defined as *Pitch~LogDuration+SentencePosition*Task+PhrasalPosition+Tone+Gender+(1+Tone+Task|Speaker)*. The duration model is defined as *LogDuration~Tone+ SentencePosition *Task+(1+Task+Tone|Speaker)*.

In terms of pitch height, there is no significant interaction between the task and the sentence position, meaning that the difference between the customer utterances and boss utterances is not different from the difference between the first two sentences and the sixth sentence in the sentence-reading task. In addition, mid-falling and rising tones receive lower pitch, as expected. A significant effect of phrasal position indicates a down-step effect (Table I) [10]. That is, on a group level, the deaf participants engage with the lexical tone and global intonation as the mainstreamed speakers do.

In terms of duration, the deaf participants shifted to shortened vowels in the role of the boss, compared to the role of the customer. It is demonstrated by a significant effect of sentence position (Estimate = 0.194, SD = 0.033, $t = 5.735$, $p < 0.001$) and a significant interaction between sentence position and task (Estimate = -0.307, SD = 0.049, $t = -6.245$, $p < 0.001$; Fig. 1). The current research indicates that deaf adults make use of acoustic cues in the temporal domain to communicate social dominance, in line with previous research on Mandarin rude speech among hearing speakers.

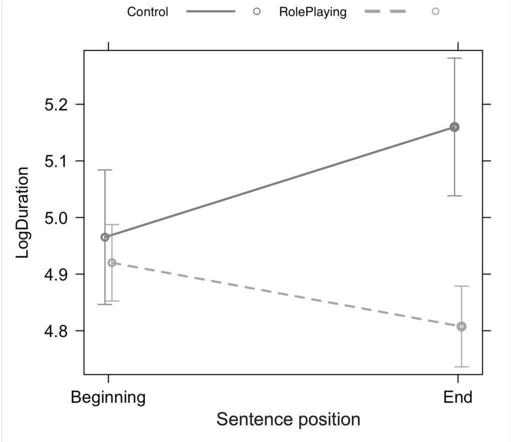


Fig. 1. The interaction between sentence position and task on vowel duration

TABLE I. SUMMARY OF PITCH HEIGHT MODEL

	Estimate	*SD*	*t value*	*p value*
(Intercept)	11.521	1.587	7.259	<0.001
LogDuration	-0.678	0.207	-3.268	0.001
SentencePosition=End	0.667	0.234	2.85	0.004
Tone = high-level	n.s.			
Tone = mid-falling	-2.584	0.328	-7.874	<0.001
Tone = rising	-2.402	0.253	-9.494	<0.001
Phrasal position	-0.229	0.020	-11.140	<0.001
Task = Role-Playing	n.s.			
SentencePosition=End:Task=Role-Playing	n.s.			

REFERENCES

[1] M. Chatterjee *et al.*, “Acoustics of Emotional Prosody Produced by Prelingually Deaf Children With Cochlear Implants,” *Front. Psychol.*, vol. 10, p. 2190, 2019, doi: 10.3389/fpsyg.2019.02190.

[2] T. J. de Jong, M. M. Hakkesteegt, M. P. van der Schroeff, and J. L. Vroegop, “Communicating Emotion: Vocal Expression of Linguistic and Emotional Prosody in Children With Mild to Profound Hearing Loss Compared With That of Normal Hearing Peers,” *Ear Hear.*, Online ahead of print.

[3] J. Yu, Y. Liao, S. Wu, Y. Li, and M. Huang, “Discourse Prosody in Children’s Rhyme Speech Produced by Prelingually Deaf Mandarin-Speaking Children With Cochlear Implants,” *J. Speech Lang. Hear. Res.*, vol. 63, no. 6, pp. 1736–1751, Jun. 2020

[4] J. J. Ohala, “Cross-language use of pitch: An ethological view,” *Phonetica*, vol. 40, no. 1, pp. 1–18, 1983

[5] R. J. Podesva, “Gender and the social meaning of non-modal phonation types,” *Annu. Meet. Berkeley Linguist. Soc.*, vol. 37, no. 1, p. 427, Jun. 2011

[6] L. Esposito and C. Gratton, “Prosody and ideologies of embodiment: Variation in the use of pitch and articulation rate among fitness instructors,” *Lang. Soc.*, vol. 51, no. 2, pp. 211–236, 2020

[7] J. Holliday, A. Walker, M. Jung, and E. S. R. Cho, “Bringing indexical orders to non-arbitrary meaning: The case of pitch and politeness in English and Korean,” *Lab. Phonol.*, vol. 14, no. 1, Feb. 2023, doi: 10.16995/labphon.9112.

[8] S. Shuju, T. Chiharu, F. Xiaoli, Z. Jinsong, and M. Nobuaki, “Acoustic correlates and gender effects in production and perception of Japanese polite speech,” in *2016 10th ISCSLP*, IEEE, Oct. 2016, pp. 1–5.

[9] P. Fan and W. Gu, “Prosodic cues in polite and rude Mandarin speech,” in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Oct. 2016, pp. 1–4.

[10] B. Wang and Y. Xu, “Differential prosodic encoding of topic and focus in sentence-initial position in Mandarin Chinese,” *J. Phon.*, vol. 39, no. 4, pp. 595–611, Oct. 2011

Backchannel and Filled Pause by Taiwan Mandarin Speakers with Autism

Vanessa Shih-Han Wu [a], Ying Hsun Liu [a], Shaoren Lyu [a], Hohsien Pan [a], Susan Shur-Fen Gau [b]
[a] *National Yang Ming Chiao Tung University, Taiwan,*
[b] *National Taiwan University Hospital, Taiwan*

Keywords — Autism Diagnostic Observation Schedule, spontaneous dialogues, frequency, f0

I. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by difficulties in social interaction, as classified by the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) [1]. Previous research suggests that autistic individuals exhibit relatively intact grammatical aspects of prosody, while demonstrating difficulties primarily in pragmatic and affective language functions [2], [3], [4]. Effective communication relies on dialogue coordination between interlocutors, where speakers often perform self-repairs, repetitions, and hesitations, using either silences or filled pauses (FPs), while listeners provide backchannel responses (BC) to signal understanding, attention, agreement, or encouragement. Fillers can serve distinct pragmatic functions based on the speaker's intent, such as facilitating turn-taking, holding, or yielding the floor, helping speakers manage interactions and engaging their listeners [5]. Recent studies [6] revealed that German speakers with autism had a lower rate and less diversity in the use of BCs compared to the control group. Conversely, autistic speakers produced fewer FPs with canonical level intonation but showed more diversity in the intonational realization of FPs. This study investigates the two roles of Mandarin toneless "en" in spontaneous dialogues. Mandarin “en” serves three pragmatic functions, including positive response, BC, and FPs. In Mandarin, "en" function as “yes” to positively respond to a yes-or-no question. The usage of Mandarin "en" as a positive response differs fundamentally from the pragmatic functions of using Mandarin "en" as either backchannels or filled pauses. The backchannel function of Mandarin "en" signals attention engagement to other conversation partners, whereas the filled pause function of Mandarin "en" indicates hesitancy to hold the floor by the speaker during conversation. We disentangled the use of Mandarin "en" and compared the frequency, normalized duration and f0 to determine whether differences exist between non-autistic and autistic speakers in the production of BC and/or FP.

II. Method

A total of 104 individuals participated in the study, including 30 male non-autistic speakers (age range: 7-24, Mean=14.93, SD=6.21), 17 female non-autistic speakers (age range: 7-21, Mean=17.18, SD=3.99), 50 male ASD speakers (age range: 4-26, Mean=15.88, SD=5.71), and 7 female ASD speakers (age range: 6-18, Mean=13.14, SD=3.91). Phonetically transcribed dialogues from the emotion sessions in Module 3 and Module 4 of the Autism Diagnostic Observation Schedule (ADOS-G) interviews were conducted. Module 3 and 4 are for verbally fluent children / adolescents or adults respectively. A total of 781 "en" tokens were analyzed for f0, and 742 tokens were analyzed for duration after removing outliers. Likelihood ratio tests were used to identify the factors, age and gender, to be included into the model to optimize the goodness of fit. The normalized log semitone mean f0 of BC and FP produced by autistic and non-autistic individuals were compared with Linear Mixed Effects Regression Models (LMER) using the lmer() function in R. The LMER models (autistic state) with speaker as a random effect and non-autistic data as baseline analyzed the mean f0 of FP and BC.

III. RESULTS

The ANOVA results showed no significant difference in the percentage of filled pauses between autistic and non-autistic speakers (F(1, 102) = 0.766, p = 0.384). However, autistic speakers produced a significantly lower percentage of backchannels compared to non-autistic speakers (F(1, 102) = 8.654**) (Figure 1).

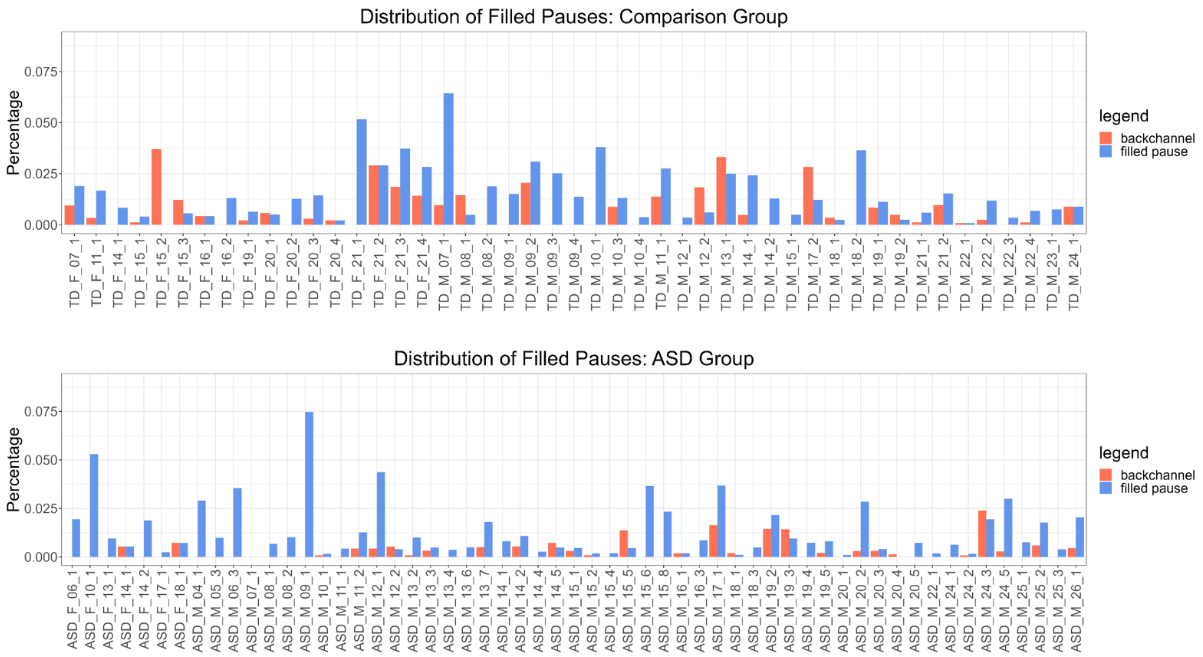


Fig. 1. Individual speakers’ frequencies of Mandarin “en” as backchnnel and filled pauses.

Likelihood ratio tests reveal that the inclusion of gender (χ^2 = 17350.2 ***), and age (χ^2 = 1445.60***) significantly improve the goodness of fit of the model on mean f0. The LMER models (autistic state) on f0 means of BC reveal that autistic male youths exhibited significantly lower f0 than non-autistic speakers (male youths: β=-1.98**; boys: β=-1.57*). As for FP, autistic female youths produced lower f0 in FP than their non-autistic peers (β=-2.92*) (Figure 2).

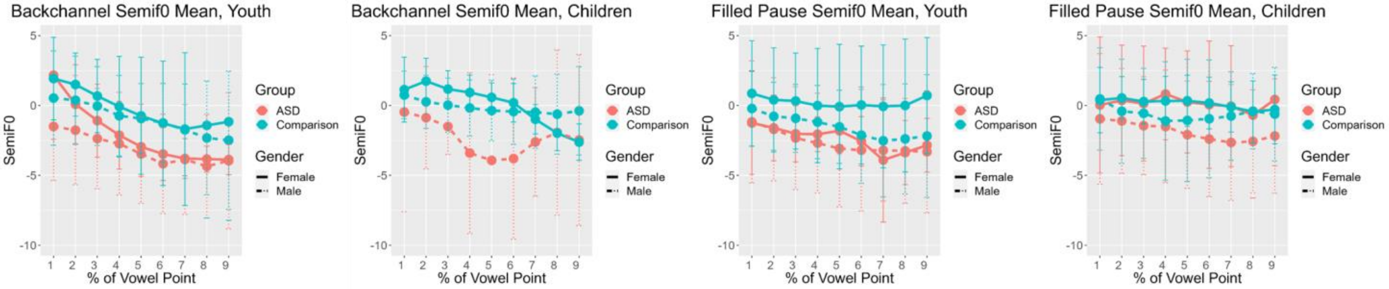


Fig. 2. Semitone f0 contours of backchannel and filled pauses 'en' across subgroups by autistic state and gender, with age group distinctions

In sum, autistic speakers produced significantly fewer backchannels compared to non-autistic speakers. Autistic male speakers produced BC and autistic female youths produced FP with autism had significantly lower mean f0.

References

[1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition. American Psychiatric Association, 2013. doi: 10.1176/appi.books.9780890425596.

[2] R. Paul, L. D. Shriberg, J. McSweeny, D. Cicchetti, A. Klin, and F. Volkmar, “Brief Report: Relations between Prosodic Performance and Communication and Socialization Ratings in High Functioning Speakers with Autism Spectrum Disorders,” *Journal of Autism and Developmental Disorders*, vol. 35, no. 6, pp. 861–869, Dec. 2005, doi: 10.1007/s10803-005-0031-8.

[3] E. Shriberg, “To ‘errrr’ is human: ecology and acoustics of speech disfluencies,” *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153–169, 2001, doi: 10.1017/S0025100301001128.

[4] H. Tager-Flusberg, “On the nature of linguistic functioning in early infantile autism,” *Journal of Autism and Developmental Disorders*, vol. 11, no. 1, pp. 45–56, Mar. 1981, doi: 10.1007/BF01531340.

[5] H. H. Clark and J. E. Fox Tree, “Using uh and um in spontaneous speaking,” *Cognition*, vol. 84, no. 1, pp. 73–111, May 2002, doi: 10.1016/S0010-0277(02)00017-3.

[6] S. Wehrle, “A Multi-Dimensional Analysis of Conversation and Intonation in Autism Spectrum Disorder,” PhD Thesis, University of Cologne, 2022. [Online]. Available: https://osf.io/6vynj

German and Hungarian long and short vowels in fast speech

Andrea Deme[a,b], Kornélia Juhász[a,b,c], Zsuzsa Szánthó[a,b], Szabina Zsoldos[a], Reinhold Greisbach[c]
[a]*ELTE Eötvös Loránd University, Hungary*
[b]*MTA–HUN–REN NYTK Lendület "Momentum" Neurophonetics Research Group, Hungary*
[c]*HUN–REN Hungarian Research Centre for Linguistics, Hungary*
[d]*University of Cologne, Cologne, Germany*

Keywords — speech rate, vowel length contrast, fast speech, vowel quantity, Euclidean distances

I. Introduction

Fast speech is the result of speech sounds produced shorter. However, it is expected that in terms of duration, not each segment may be reduced to the same extent, especially if duration serves linguistic functions, as in e.g., German, and Hungarian, where vowel length contrast is phonologically distinctive. In Hungarian, phonologically short and long high vowels, e.g., /i iː/ /u uː/, are traditionally assumed to be distinguished primarily by duration, while short and long low vowel pairs /ɒ aː/ and /ɛ eː/ differ also in their quality [8]. In German, the situation is the other way round: (in accented syllables) we find no quality difference between the low vowels /a aː/, while there is a simultaneous durational and quality difference in high vowels, e.g., in /ɪ iː/, and /ʊ uː/ [3].

In vowels, temporal reduction in fast speech is also expected to be accompanied by some degree of spectral changes, due to target undershoot [4,5]. As a result, increased speech rate is expected to endanger the vowel length contrast both in the temporal and the spectral domain. To this hypothesis, we find scarce, and to some extent, inconclusive evidence in Hungarian and German. With respect to **duration**, in Hungarian, there is some evidence that long vowels reduce to a higher degree than short vowels [5, 2]. Similarly, in German, long (tense) vowels were found to reduce more than short (lax) ones [3]. In **vowel quality**, German short and long vowel spaces were found to be affected by fast speech in a similar fashion: they were both reduced, or (as opposed to expectations) increased (in different dialectal regions) [7]. In Hungarian, we find no systematic analysis of the spectral changes in fast speech. In the present study, we hypothesize that increased speech rate induces reduction of the vowel length contrast in Hungarian and German. Our aim is to explore if this reduction emerges differently in these two, typologically unrelated languages, where the phonological vowel length contrast is expressed using similar means, but in a different implementation.

II. Methods

We analyzed short and long vowel pairs in monosyllabic words in the production of 15 Hungarian and 14 German speaking females. In the Hungarian material, target vowels consisted of /uuː iiː ɒaː/. In German, we used the corresponding /ʊuː ɪiː aaː/ pairs. Consonants preceding and following target vowels were controlled for place of articulation. Target sequences did not constitute minimal pairs, hence did not facilitate exaggeration of the contrast at hand. Speakers produced target words in carrier sentences, where the target word bore sentence level accent. We recorded samples in two speech rate conditions: at i) comfortable/”normal” speech rate, and ii) maximum/”fast” speech rate. To achieve maximum speech rate, we instructed speakers to repeat each target sentence several times starting with a comfortable tempo (“normal” speech), and by each repetition, increase speed. Each participant produced 6 of these sets for each target word resulting in 72 sets (144 tokens) per speaker in total. We segmented all words, and vowels manually, and labelled the shortest words as the fast speech variants. We analyzed temporal and spectral measures in the two speech rate conditions using Praat [1]: a) target vowel durations, and duration ratio of short-long pairs, and b) Euclidean distances of the pairs in the F1×F2 vowel space. Data were submitted to linear mixed effects modeling separately for the two languages.

III. Results

On average, in fast speech, speakers produced words in half the length of that found in normal speech in both languages. Fig. 1 shows vowel durations as a function of vowel quantity, vowel type, and speech rate. In both of the languages, these three factors had a significant interaction effect on the data (HUN: $F(2, 2095) = 16.5$; $p < .001$; GER: $F(2, 1960) = 11.5$; $p < .001$). According to the post hoc tests, durational reduction in fast speech was significant in all vowels, and members of the vowel pairs differed in both speech rate conditions in Hungarian, while the contrast disappeared in /ʊ uː ɪ iː/ pairs in fast speech in German. In duration ratios (Fig. 2), as opposed to expectations, we found /i iː/ contrast to be the greatest of all contrasts in normal speech in Hungarian, while in German, /ɪ iː/ differed the most and /ʊ uː/ the least. In fast speech, in Hungarian, all contrast reduced so that they all became similar to one another; while in German, the differences found in normal speech between the three pairs were preserved, and /u/ pairs reduced in a manner that the contrast seem to reach complete neutralization. According to pair-wise comparisons, in Hungarian, we found a significant decrease of the contrast only in /i iː/ (vowel type*speech rate interaction: $F(2, 75) = 16.9$; $p < .001$), while in German, in all vowel pairs (significant speech rate [$F(1, 14) = 61.4$; p < .001] and vowel type [$F(2, 19) = 22.9$; $p < .001$] main effects).

Euclidean distances of short-long vowel pairs (Fig. 3) showed significant interaction effect of vowel type and speech rate in Hungarian [F(1, 75)= 11.24, p < .01], and speech rate [F(1, 70) = 5.3; p < .05] and vowel type [F(2, 70) = 193.0; p < .001] main effects in German. Post hoc analyses revealed that spectrally, in line with expectations, in Hungarian /a/ contrast was greater than /u/ and /i/ contrasts (while /u/ and /i/ contrasts did not differ from each other significantly) in both speech rate conditions. Further, as expected, in German, /i/ contrast was the largest, followed by /u/ and /a/ in both speech rate conditions. In Hungarian, only the contrast of (i.e., distance between) /ɒ/ and /a/ was reduced in fast speech. In German, however, none of the differences between fast and normal speech contrasts differed significantly, that is, spectral distinction between the pairs did not reduce.

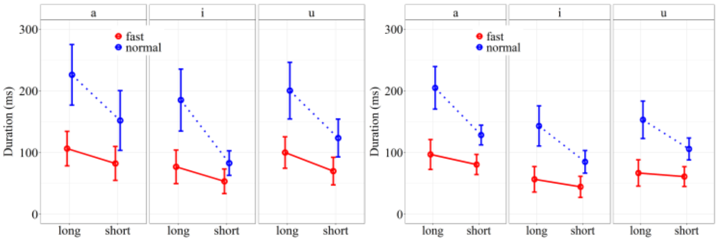


Fig. 1. Short and long vowels' duration in Hungarian (left) and German (right)

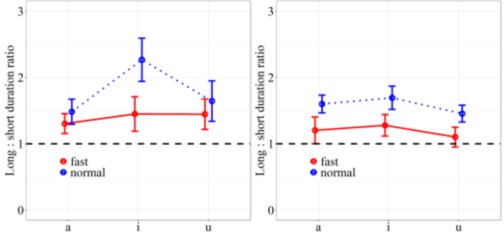


Fig. 2. Short and long vowels' duration ratio in Hungarian (left) and German (right)

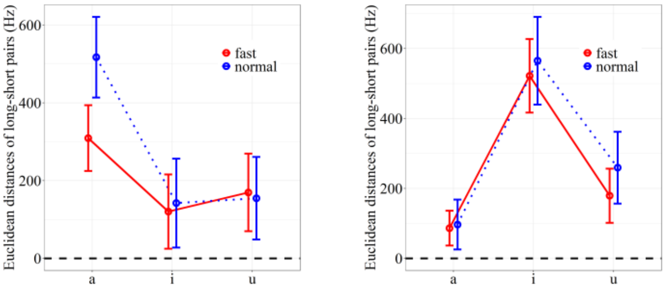


Fig. 3. Short and long vowels' Euclidean distances based on F_1 and F_2 in Hungarian (left) and German (right)

IV. Conclusions

In comfortable tempo, German long and short vowel pairs were extensively distinguished by quality and duration, while in fast speech, all vowels and their distinction reduced in duration, and distinction of vowel pairs was not reduced in quality. In fast speech, temporal distinction of /ʊ uː/ and /ɪ iː/ disappeared. In comfortable speech, contrast was the greatest in /ɪ iː/ and the smallest in /ʊ uː/ durationally, while it was the greatest in /ɪ iː/ and the smallest in /a aː/ spectrally. In Hungarian, quality and duration equally distinguished vowels in both speech rates. In fast speech, all vowels reduced in duration, but in duration distinction only /i iː/ showed reduction. Spectrally, only the /ɒ aː/ distinction reduced. In normal speech, /i/ pairs were distinguished the most (not /a/ pairs), but qualitatively /a/ pairs differed the most even in fast speech. Results showed that duration cues reduced more extensively in German than in Hungarian, while vowel quality seemed to change more in Hungarian in general (due to the changes in /ɒ aː/).

References

[1] P. Boersma, D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.3, 2022. http://www.praat.org/

[2] A. Deme, K. Juhász, Zs. Szánthó, Sz. Zsoldos, R. Greisbach, "Segmental durations and the vowel length contrast in fast speech in Hungarian," In: C. Fougeron, P. Perrier (eds.) Proc. 13th ISSP. Autrans, 2024. pp. 37-40.

[3] P. Hoole, C. Mooshammer, H. G. Tillmann, "Kinematic analysis of vowel production in German," Proc. ICSLP 94, Yokohama, vol 1, pp. 53–56, 1994.

[4] B. Lindblom, "Spectrographic study of vowel reduction", J Acoust. Soc. Am. vol. 35. pp. 1773-l78l, 1963.

[5] B. Lindblom, "Explaining phonetic variation: a sketch of the H & H theory," in Speech production and speech modeling, W. J. Hardcastle, A. Marchal, Eds. Dordrecht: Kluwer, 1990, pp. 403–439.

[6] K. Magdics 1969. "A magyar beszédhangok időtartama nyugodt és gyors beszédben," Nytud. Ért. vol. 67. pp. 45–63.

[7] B. Siebenhaar, M. Hahn, "Vowel space, speech rate and language space," in: Proc. 19th ICPhS. Melbourne, 2019, pp. 879-883.

[8] P. Siptár, M. Törkenczy, "The Phonology of Hungarian," Oxford University Press, 2007.

[9] R Core Team 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Acknowledgements

The research was supported by the TKA–DAAD grant No. 177375., the NKFIH grant No. FK128814, and the ÚNKP-23-3-I grant (Zs. Sz.).

Decoding L1 Influence: Navigating Phonotactic Barriers in L2 German Pronunciation by Bulgarian Speakers

Denitsa Dimitrova
Sofia University “St. Kliment Ohridski”, Bulgaria

Keywords — Phonotactics, Second Language Acquisition, Phonetic Interference, Lateral Consonants, Palatalization

I. Introduction

Bulgarian speakers often exhibit a 'hard' accent when speaking German, attributed to the phonotactic differences in lateral consonants between Bulgarian (L1) and German (L2). The German /l/ is consistently palatalized across different positions, unlike in Bulgarian, which has [ɫ] and [l], with the latter appearing only before front vowels. However, there's a noticeable trend towards pronouncing the velarized allophone [ɫ] as [w] or [ɰ] [1], [2], [3], leading to interference, especially before or after back vowels when Bulgarians pronounce the German lateral sound.

This study applies the Perceptual Assimilation Model (PAM) [4] and the Speech Learning Model (SLM) [5] to understand these pronunciation challenges. These models suggest perceptual strategies from L1 acquisition can cause foreign sounds resembling L1 phonemes to be perceived as their L1 counterparts. This effect, called "perceptual assimilation” in PAM and "equivalence classification" in SLM, is evident in the assimilation of the German palatalized [l] before back vowels to the familiar Bulgarian combination [l] + [j] + back vowel, as in the word любов [ljubof] (‘love’).

II. Methodology

The study involved three main experiments focusing on phonological analysis, perception and production tests. First, the phonological systems of Bulgarian and German were analyzed to understand the phonotactic constraints faced by Bulgarian speakers learning German. Next, a perceptive-productive test from the SETK 3–5 battery [6] was conducted with 76 second and third-grade students, focusing on phonological working memory using nonsense words (logatoms). Participants were divided into two groups: one heard and repeated the logatoms, while the other heard and wrote them down. In two logatoms the palatalized [l] at the end of the syllable following the back vowel [a] violated Bulgarian /l/ allophone rules.

The second experiment examined the perception and production of [l] before a rounded back vowel by Bulgarians with varying German proficiency, involving first and second-grade students and first-year German Philology students. Participants were presented with 32 German words, including nonsense words, recorded by a native speaker. Group 1 listened and repeated all stimuli, Group 2 selected from forced-choice options, and Group 3 listened and wrote the stimuli down. The recordings of Group 1 were digitized and analyzed using Praat software, focusing on the first two formants (F1 and F2) to investigate the perception and production of the palatalized German [l] before rounded back vowels.

Finally, an additional production test involved seven-year-old first-grade students from two schools with intensive German programs. The stimuli consisted of two-syllable words in the CV-CV format with stress on the first syllable, containing [l] + a rounded vowel in the first syllable, and a plosive and schwa in the second. The stimuli were recorded, digitized, randomized, and analyzed using Praat.

III. Results

A. First Experiment

Participants frequently substituted the palatalized German [l] with the second component of the diphthong /aɪ/ when it appeared at the end of the syllable following the back vowel [a]. This occurred in 32% of Group 1, 62% of Group 2, and 37.5% of Group 3.

For example, in "Waltikosander," the <al> combination was replaced with <ai> by 68% of Group 1 and 43% of Group 2, with only one substitution in Group 3. Follow-up tests showed mispronunciations of familiar words like "Wald" as "weit" and "bald" as "beid," indicating phonological confusion.

B. *Second Experiment*

Three pronunciation variants of [l] + rounded back vowel were observed in Group 1: correct production, substitution with [lju], and replacement with [ly], with [lu] to [ly] being the most common substitution. Group 2's forced-choice test showed higher correct response rates due to their age and language experience. Group 3 mainly substituted back vowels with front rounded vowels in writing, demonstrating their higher proficiency in German.

C. *Third Experiment*

For rounded front vowels following [l], participants showed a high percentage of correct realizations, with formant values matching those of the native speaker. However, for rounded back vowels, participants often substituted front rounded vowels or added [j] before the back vowel, resulting in elevated F2 values. This indicates greater difficulty in accurately producing rounded back vowels following [l], highlighting the influence of L1 phonological constraints on L2 pronunciation.

IV. Discussion

The experiments indicate that L1 phonotactic rules are crucial in predicting the difficulty of acquiring L2 segments, similar to the phonetic proximity between L2 and L1 segments. PAM and SLM provide insights: PAM explains that Bulgarian learners often substitute the German palatalized [l] with native sounds or sound combinations due to the absence of a similar consonant in Bulgarian. SLM posits that L2 sounds are perceived and produced based on their similarity to L1 sounds, but also on the experience with L2, explaining varying success rates. The forced-choice test showed higher correct responses in older students due to refined perceptual categories. The results of the German Philology students clearly demonstrate that focused phonetic training can lead to improved discrimination and production. Formant analysis supports both models, indicating participants struggled with the combinations of [l] and rounded back vowels, reverting to familiar L1 articulations.

Knowledge of the rule for combining the palatalized allophone of /l/ in L1 hinders its correct discrimination in phonetic distributions not allowed by L1, specifically after a rounded back vowel. In these instances, participants apply two strategies that utilize the same acoustic feature of [l], namely the high frequencies of F2 indicating a narrowing of the vocal tract. In the first scenario, they interpret these high values as an indication of inserting the glide [j] between [l] and a back rounded vowel. In the second scenario, they use the F2 values as the primary information for the place of articulation of the vowel and replace the back rounded vowel with a front vowel, thus preserving the rounded feature.

Overall, the study highlights the significant impact of L1 phonotactic rules on L2 pronunciation, demonstrating the utility of targeted phonetic training.

References

[1] V. Zhobov, "Zvukovete v bŭlgarskiya ezik (The Sounds in the Bulgarian Language)", Sofia, Bulgaria: Sema RSh, 2004.
[2] S. Burov, "Dve normi na bŭlgarskata ustna knizhovna rech (Two Norms of Bulgarian Oral Literary Speech)", in LiterNet, electronic journal, Nov. 7, 2012, no. 11 (156). http://liternet.bg/publish28/stoian-burov/dve-normi.htm
[3] G. Padareva-Ilieva and S. Mitsova, "Is Bulgarian Language Losing Its Alveo-Dental Consonant [l]?" International Journal of Linguistics and Communication, vol. 2, no. 1, pp. 45–65, New York, 2014. ISSN 2372-4803.
[4] C. Best and T. M. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in Language Experience in Second Language Speech Learning: In honor of James Emil Flege, O-S. Bohn and M. Munro, Eds. Amsterdam: John Benjamins, 2007, pp. 13–34.
[5] J. E. Flege, "Second Language Speech Learning: Theory, Findings, and Problems," in Speech Perception and Linguistic Experience: Issues in Cross-Language Research, W. Strange, Ed. Timonium, MD: York Press, 1995.
[6] H. Grimm, "Sprachentwicklungstest für drei- bis fünfjährige Kinder (SETK 3–5)," Göttingen: Hogrefe, 2001.

The sociophonetics of coda /t/ in Dublin English

Chloé Diskin-Holdaway [a], Debbie Loakes [a], Kirsty McDougall [b]
[a] *The University of Melbourne, Australia,*
[b] *University of Cambridge, United Kingdom*

Keywords — Dublin English, /t/, sociophonetics, frication

I. Introduction

The variable /t/ in Dublin English (DubE) has been of particular interest among scholars of Irish English (IrE), and is an "immensely variable sound in the accents of English" [1: 728] more generally. A fricated realisation of /t/ (also known as a slit-t), particularly in coda position, has been described as "one of the most conspicuous features" of IrE ([2: 429]; see also [3] and [1]). DubE /t/ is quite heterogenous, however, with documented categories including an aspirated alveolar plosive [t^h] (henceforth aspirated stop), glottalised variants, a full glottal stop [ʔ], a tap [ɾ], [h], dental [t̪], a dental fricative [θ], an affricate [ts], [ɹ] [see 3: 40-41], and deletion [4: 122]. IrE is also noted to have several fricative variants, including a voiced fricative and a fricative flap [1]. The sociolinguistic factors of age, social class, neighbourhood and gender have been found to govern /t/ variability in DubE, with the fricated variant [t̞] most common among women and in younger people living on the Northside or Southside of the city [see 5]. All other variants, especially the aspirated stop and glottalised variants, are more common among males (see [6] on its indexing toughness and masculinity) and in low socio-economic status areas of Dublin, such as the Inner City [5], or among low income or unemployed speakers [4]. The age effect is likely to reflect the impact of language change, as [4] has shown via a small-scale study of Irish radio shows from 1930-2011 that /t/ has changed from being realised primarily as the aspirated stop to the fricated variant. Fricated /t/ may thus be a more recent incursion into the variety, indicative of increased endonormativity, with the exonormative target being the aspirated stop from British English. We investigate sociophonetic variability in coda /t/ in DubE, which has been given limited scholarly attention since [7] and [8], with the exception of [4]. In doing so, we aim to contribute to the understanding of /t/ variability in English varieties and IrE more generally [see also 1]. We pose the following research questions:

1. What patterns of /t/ variability are found in present-day DubE?
2. What effect do extralinguistic (i.e. gender, location, age) factors have on /t/ variability in DubE?

II. Method and materials

The first author, a native speaker of IrE, collected the data in 2019 in Dublin with 21 DubE speakers (11 female; 10 male), born and raised in or near the city. Eight participants were from the Southside (SS in the participant codes), eight from the Northside (NS), three from Terenure, to the south-west (W) of the city and two from satellite towns in the neighbouring county of Wicklow (WW). Participants were students (*n*=15) or professionals (*n*=5), with mean age 26 (range: 18-57; *SD*=10), also represented in the participant codes. They were recorded reading aloud a wordlist of 60 real words with varying lexical frequency with three repetitions, and also completed other tasks not reported on here. We draw on 17 /hV(r)t/ items with differing preceding vowel environments (*bite, boat, boot, bought, bout, hat, hate, heart, heat, het, hit, hot, hurt, hut, let, put, that*), resulting in 1,133 tokens (54 per speaker), noting one exclusion, and *hot* appeared twice erroneously in the printed wordlist, leading to six *hot* tokens per speaker.

Sound files were uploaded to Webmaus [9] with the corresponding wordlist text file for autosegmentation, and then exported into Praat textgrids [10]. Each token was subject to an auditory and visual analysis, with the corresponding spectrogram inspected in Praat. Coding, conducted by author 1 and a subset checked by author 2, was bottom-up and data-driven, with 13 categories established and operationalized with the aid of definitions in [1, 11]. The majority of tokens were either fricated [t̞] or aspirated stop [t^h], with related categories of fricated+voiced, fricated+stop, aspirated stop+dental and dental fricative. Other categories included an affricate [t^s] and an affricate+aspirated stop variant, a pre-glottalised variant and glottal stop variant, and three categories of ejective: ejective, ejective affricate and ejective dental.

III. Results

Overall, 45.6% of the data was fricated [t̞], constituting the largest category. The word *bought* was most likely to be fricated (61.9% of the time), and *het* the least likely (28.6%). Other tokens with high (52%+) rates of frication were *boat*, *bout, hat* and *put.* Three younger speakers, 062F_SS_21, 064F_SS_21 and 071M_WW_27, had a 100% frication rate for all 54 of their tokens. The average frication rate per speaker was 45.5%, but four speakers had zero fricated variants, e.g., 065M_W_19 with 94% pre-glottalisation, and 074M_NS_18 with 66.6% aspirated stop and the remainder affricate variants. The grammatical word *that* was fricated 47.6% of the time (see [1] for a comparison of frication in lexical and grammatical words in IrE).

The next most common category was the aspirated stop, at 41.6%. *Het* was most likely to have the aspirated stop (61.9%) and *hat* the least likely (28.6%). Other words with high (52%+) rates of aspirated stop were *let*, and the two words containing /t/ following a rhotic, *heart,* and *hurt*. Two older speakers had a 100% aspirated stop rate: 069F_SS_45 and 070F_NS_57. The remaining 12.8% of variants in the dataset were distributed among eleven categories, including pre-glottalised (5.6%) and affricate (2.7%). Both of these categories were restricted to a few speakers, with just four speakers using affricate variants, the most common of which were *let* and *hut* (both at 9.5% of variants), and three speakers using pre-glottalised or glottal stop variants, the most common of which were *hat* (11.1%) and *bout* (9.5%). There were at least two pre-glottalised variants for all tokens across the dataset except for *heart*.

We examined the effect of three extralinguistic factors on rates of frication: gender, location and age. Fig. 1 (left panel) shows that females had significantly higher percentages of fricated variants (mean=61.6%) as compared to men (27.8%) (one-way ANOVA (F(1)=4.42, *p*<0.05), although with considerable interspeaker variation. Fig. 1 (right panel) shows that speakers in the West have the lowest frication rates (mean 21%) and the highest rates are among Wicklow (61.1%) and Southside speakers (55.3%), again with considerable interspeaker variation. A regression analysis found no significant differences by location, but we note that the numbers of participants in the West (*n*=3) and Wicklow (*n*=2) are comparatively few. Finally, a Pearson correlation analysis showed that age and frication were weakly negatively correlated (*r*(19) =-0.2, *p*=0.37, i.e. younger participants do not necessarily have higher frication rates), while noting that our sample is skewed towards younger participants.

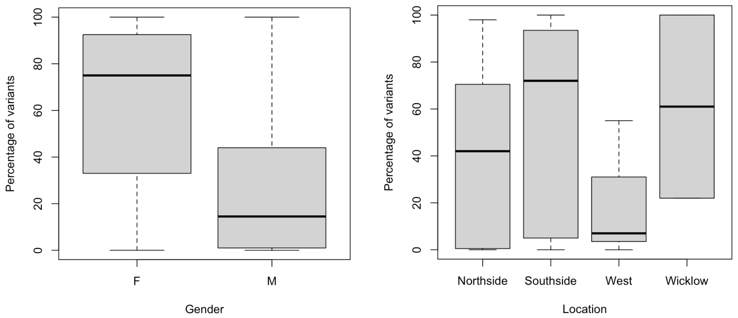


Fig. 1. *Variation in /t/ frication by gender (left panel) and location (right panel)*

IV. Discussion and conclusion

Our findings reveal 13 categories of /t/ across 1,134 tokens, but the majority of tokens are either the fricated or aspirated stop variants. We find alignment with [1] that fricated variants are particularly common in IrE pre-pausal contexts, or at the ends of prosodic phrases, noting that our participants took a pause between each of their three repetitions of the tokens. We found a voiced fricated variant, as documented for the first time in [1], but no examples of [h], noting that [h] may be restricted to particular lexical items in IrE, such as *Saturday* [1] or *scarlet* [4]. We also found three categories of ejective, with ejectives being noted in [1] but not in [4], which could be related to the effect of the wordlist task. As regards the sociolinguistic patterning of /t/, our findings align with [4] whereby fricated variants appear to be the new endormative standard among young Dubliners, with this phenomenon being particularly widespread in the Northside and Southside communities, rather than other areas such as the West (see also [5] for lack of frication in Inner City speakers). We have also shown that certain prevocalic environments appear to favour frication, including back vowels (*bought*) or diphthongs (*bout*), which warrant further investigation. Other lexical items, such as *het*, were favoured with the aspirated stop variant, suggesting a word frequency effect (*het* is a comparatively rare word, potentially eliciting careful production). Our future research aims to investigate the role of voice quality, particularly creak in the pre-glottalised variants [see 1, 11] , as well as word frequency effects, and the same participants' spontaneous speech [see 1].

References

[1] R. Skarnitzl and D. Rálišová, "Phonetic variation of Irish English /t/ in the syllabic coda," *Journal of the International Phonetic Association,* vol. 53, no. 3, pp. 728-747, 2023, doi: 10.1017/S0025100321000347.

[2] J. Wells, *Accents of English 2: The British Isles*. Cambridge: Cambridge University Press, 1982.

[3] R. Hickey, *Dublin English: Evolution and Change*. Amsterdam: John Benjamins, 2005.

[4] M. Schulte, *The Sociophonetics of Dublin English*. Amsterdam: John Benjamins, 2023.

[5] J. Lonergan, "An acoustic and perceptual study of Dublin English phonology," Unpublished PhD thesis, University College Dublin, 2013.

[6] F. O'Dwyer, *Linguistic Variation and Social Practices of Normative Masculinity*. London: Routledge, 2020.

[7] R. Hickey, "The phonology of Irish English," in *Handbook of Varieties of English. Vol.1: Phonology*, B. Kortmann Ed. Berlin: Mouton de Gruyter, 2004, pp. 68-97.

[8] J. L. Kallen, *Irish English, Volume 2: The Republic of Ireland*. Berlin: Mouton de Gruyter, 2013.

[9] T. Kisler, U. D. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language,* vol. 45, pp. 326–347, 2017.

[10] *Praat: doing phonetics by computer [Computer program]*. (2024). Accessed: 17 March 2024. [Online]. Available: http://www.praat.org/

[11] D. Loakes, K. McDougall, and A. Gregory, "Variation in /t/ in Aboriginal and Mainstream Australian Englishes," in *Australasian Speech Science and Technology Conference (SST*, Canberra, R. Billington, Ed., 2022: ASSTA, pp. 61-65.

Prosodic challenges in Italian learners’ L2 Mandarin production: a preliminary analysis

Davide Francolino, University for Foreigners of Siena

Keywords — Phonology, Mandarin Chinese, Italian, Prosody, L2 production

I. Introduction

The traditional method of Mandarin phonetic instruction has long focused primarily on the acquisition of isolated word tones, unfortunately overlooking crucial suprasegmental elements like utterance-level prosody [1]. This exclusive attention on word-level tone acquisition has led to a notable gap in the development of students' connected speech production, as well as their proficiency in intonation and rhythm [2]. The persistence of the issue can also be attributed to the scarcity of research on L2 Mandarin prosodic production, further hindering a comprehensive understanding of the intricacies involved [3].

The present study is configured as a preliminary investigation on some prosodic aspects of L2 Mandarin pronunciation among intermediate-level Italian university learners. Considering the non-tonal nature of the participants' native language (L1) and the dissimilarities in intonational realization between Mandarin and Italian, we aim to provide initial observations on the following points: 1) recurring difficulties encountered by Italian learners in L2 Mandarin prosodic production distinct from those at word-level tone production out of context; 2) identifying possible instances of negative transfer from intonational and focus realizations that may affect comprehensibility in L2 Mandarin [4]; and 3) instructional interventions on L2 Mandarin prosodic production that may also benefit word-level tone instruction.

II. METHOD

A. Stimuli and participants

The analysis centers on minimal prosodic units, i.e. disyllabic phrases in dialogic context where not only "local" tonal variations (e.g., sandhi) can be found, but also "global" variations (e.g. sentence intonation, prosodic focus) influencing the prosodic realization of the utterance [5]. Specifically, target phrases involved all possible dysillabic stressed tone combinations (excl. neutral tone), each set in short dialogues to require two intonational realizations (statement and interrogative) and two contrastive focus realizations (on the first and on the second syllable, i.e. pre-focus and post-focus, respectively). The stimuli consist of 32 short dialogues, each including two target phrases: a statement and an unmarked echo question with the same focus position; below is an example of a short dialogue with two tone 4-tone 2 (T4T2) post-focus target phrases (focus in bold):

28. Stim_T4T2_post
A: 你从云南回来，打算给你妈带茶还是带咖啡？
A: nǐ cóng Yúnnán huílái, dǎsuàn gěi nǐ mā dài chá háishì dài kāfēi?
A: [Are you planning to bring tea or coffee for your mom when you come back from Yunnan?]

B: 带**茶**。
B: dài **chá**.
B: [I’ll bring **tea**.]

A: 带**茶**？
A: dài **chá**?
A: [Bring **tea**?]

B: 对啊，毕竟茶叶才是云南真正的特产。
B: duì a, bìjìng cháyè cái shì Yúnnán zhēnzhèng de tèchǎn.
B: [Yeah, after all tea leaves are the true specialty of Yunnan.]

The participants (n=6, all female) include five native Italian bachelor third-year students, aged 21 to 26, and a 22-year-old native Chinese informant (Ch) with a Putonghua Proficiency Exam certification level 1-A. The total number of target phrases is 384 [16 tone combinations * 2 intonational realizations * 2 focus realizations * 6 participants].

B. Data acquisition and extraction

The data for this analysis consist of recordings obtained during a reading session. Recordings were made using a Shure MVL connected to a 7th-gen. iPad. The audio tracks were saved in .wav at 48 kHz/24-bit. Participants recorded their readings in the presence of the researcher alone, after completing a preliminary questionnaire and engaging in a brief conversation primarily aimed at reducing the “affective filter” [6]. The recordings of the target phrases productions were extracted and saved individually in .wav using Logic Pro X. The study employs a dual-tiered approach, incorporating pitch-contour analysis and a perceptual test. Pitch-contour analysis was conducted using Praat (v6.3.03), focusing on two main features: 1) f0 variation and 2) relative syllable duration. The results were then compared with [7], and with those obtained from the native speaker (Ch). The f0 variation was obtained by measuring three points (P) from the tone bearing unit (TBU) of each syllable[1]: onset point, midpoint, and endpoint [5]. The f0 values were extracted in Hz and converted to St with a f_{ref} of 80 Hz [8]. Relative syllable duration was calculated as a percentage including the entire segment of the syllable. The perceptual test, conducted using PsychoPy6 (v2023.2.3), involved an auditory decision task designed to test the perception of the intonational information conveyed by the target phrases, presented without context or visual cues. For each sentence, participants chose between "statement," "question," or "I don't know". Each sentence was played twice before a response was required. The informants (n=3) were native Chinese speakers aged 24 to 26. The audio material included target phrases produced by Ch as a control (positive).

III. BIREF DISCUSSION

Participants faced interesting challenges deviating from native prosodic strategies, as evidenced by the pitch-contour analysis, and hindering the conveyance of the prosodic information, as suggested by the perceptual test results. One notable phenomenon observed was tone-intonation interference. Fig.1 illustrates tone-intonation interference in students’ productions (S1, S2 and S5) of the pre-focus interrogative "jiāo kè? 教课？ " (T1T4), a compatible tone sequence[2].

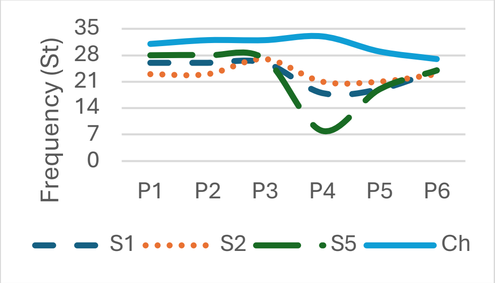


Figure 1 Example of tone-intonation interference

In students’ productions, the contour of T4 (P4-P6, HL in its citation form) rises, influenced by intonational demands. Notably, in the production of S5, the value of the starting point of T4 (P4) significantly lowers compared to the first syllable (P1-P3), further emphasizing the final rising contour (LH). The above-mentioned L2 productions differ from L1 production (Ch) as in the latter a change in the tone register is employed instead of a change in the tone contour. In fact, in Ch T4 retains its falling contour. Further and more in-depth results will be discussed also in relation to focus positioning on target phrases, offering valuable insights into the way speakers of a non-tonal language, like Italian, perceive and produce Mandarin prosody. The author will also explore the potential benefits of research-informed instructional interventions on L2 Mandarin prosodic production to word-level tone instruction.

REFERENCES

[1] C. Yang, “The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice”, Amsterdam: John Benjamins, 2016.

[2] H. Třísková, “Acquiring and Teaching Chinese Pronunciation”, in Explorations into Chinese as a Second Language, 2017, pp. 3–30.

[3] C. Yang, “Teaching Chinese Intonation and Rhythm”, in C. Shei, M. E. McLellan Zikpi, & D.-L. Chao (Eds.), The Routledge Handbook of Chinese Language Teaching, Abingdon-New York: Routledge, 2020, pp. 180-194.

[4] S.-A. Jun and M. Oh, “Acquisition of second language intonation”, J. Acoust. Soc. Am., vol. 107, fasc. 5_Supplement, pp. 2802–2803, mag. 2000.

[5] W. Cao, “Hànyǔ jiāodiǎn zhòngyīn de yùnlǜ shíxiàn 汉语焦点重音的韵律实现 [Prosodic realization of focus in Mandarin]. Běijīng: Běijīng yǔyán dàxué chūbǎnshè, 2010.

[6] S. D. Krashen, “Principles and practice in second language acquisition”, Oxford: Pergamon Press, 1982.

[7] R. Wang and W. Cao, “Hànyǔ pǔtōnghuà shuāngyīnjié jù shíyàn yánjiū 汉语普通话双音节句实验研究 [Experimental Research on Bisyllabic Phrases in Standard Chinese]”, Qīnghuá dàxué xuébào (zìrán kēxué bǎn), 49(S1), 2009.

[8] F. Nolan, Intonational equivalence: An experimental evaluation of pitch scales, Proceedings of the 15th ICPhS 2003, 2003, pp. 771-774.

[9] Y. Xu, “Production and perception of co-articulated tones”, J. Acoust. Soc. Am., vol. 95, , 1994, pp. 2240–2253.

[1] All the syllables of the target phrases were designed to begin with a voiceless sound and end with a voiced sound, thereby constraining the TBU to the final sound.

[2] A compatible tone sequence is characterized by similarity between the tone target at the offset of the preceding tone and the onset of the following tone (e.g., T1's offset target is H, and T4's onset target is also H). Such sequences are generally easier to produce at the word level for both L1 and L2 Mandarin speakers [9]. Therefore, the study mainly focuses on these sequences to better observe utterance-level prosody-related phenomena.

The Relationships Among Inhibitory Control, Auditory Integration, and L2 Perception/Production Accuracy

Amanda Huensch

University of Pittsburgh, USA

Keywords — L2 perception/production, L2 phonology, cognitive control

I. Introduction

The question of the relationship between language use and cognitive skills (e.g., attention, working memory, inhibition) is a current and promising line of inquiry in SLA [1, 2] with direct implications for theorizing in connection to usage-based accounts of L2 learning [3, 4]. In the field of L2 pronunciation, these questions have been explored with regard to both general cognitive mechanisms [5] as well as those specifically related to domain-general auditory processing [.g., 6, 7]. In these separate lines of inquiry, both inhibition (i.e., the ability to suppress an automatic or dominant response) and auditory integration (i.e., the ability to repeat melodic or rhythmic strings) appear to be promising in their ability to explain individual differences in L2 perception and production abilities for classroom learners. What remains unknown is how inhibition and auditory integration might be related to one another, and whether different types of inhibition might be better suited to explain pronunciation abilities.

II. Methods

A. Participants and Materials

The current study explored the relationships among inhibition, auditory integration, and L2 perception/production in a sample of L1 English learners of Spanish (*n*=58). Replicating Darcy et al. (2016) [5] perception was measured using a speeded ABX categorization task and production was measured using a delayed sentence repetition task. Three distinct tasks were chosen to measure inhibition to tease apart whether different types of inhibition (e.g., resisting a dominant response vs. proactive interference) might be more or less related to auditory integration and L2 pronunciation skills. Auditory integration was measured using the melodic and rhythmic reproduction tasks in [7].

B. Results

The first underlying theoretical question at the core Darcy et al. (2016) and the current replication regards the extent to which L2 phonological skills are related to general cognitive abilities, specifically inhibitory control. The logic behind the hypothesized relationship is that those with greater inhibitory control might be better at suppressing their L1 during the processing of L2 acoustic-phonetic input, ultimately resulting in more accurate segmental categories. Having these more accurate categories would, in turn, result in more accurate perception and production of segments during language use. Put another way, variability in outcomes in L2 pronunciation could be the result of individual differences in inhibitory control. Bringing together the findings from DM&D and the current replication, no strong, clear, or consistent relationship emerges between inhibitory control and L2 perception/production skills. Multiple explanations for why there were discrepancies between the results of the two studies are explored.

The second underlying question of the extension portion of this study is the relationship between L2 phonological skills and domain-general auditory processing. An intriguing body of research is providing mounting evidence for the role of domain-general auditory processing in second language learning [7] such that “the ability to precisely encode auditory input [e.g., information about frequency, duration, amplitude] may be a bottleneck for the establishment of knowledge about segmental and suprasegmental linguistic categories” [8, p. 480]. The current analysis for the relationship between L2 phonological skills and domain-general auditory processing is currently ongoing. The findings will provide a better understanding which types of skills are most robustly related to L2 perception/production and more importantly whether they relate to auditory integration.

C. Authors and Affiliations

Amanda Huensch is Assistant Professor in the Department of Linguistics at the University of Pittsburgh.

REFERENCES

[1] Luque, A., & Morgan-Short, K. (2021). The relationship between cognitive control and second language proficiency. *Journal of Neurolinguistics, 57*. https://doi.org/10.1016/j.jneuroling.2020.100956

[2] McManus, K. (2021). Crosslinguistic influence and second language learning. Routledge.

[3] Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27,* 164–194.

[4] MacWhinney, B. (2008). A unified model. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 341–371). Routledge.

[5] Darcy, I., Mora, J. C., & Daidone, D. (2016). The role of inhibitory control in second language phonological processing. *Language Learning, 66,* 741–773. https://doi.org/10.1111/lang.12161

[6] Saito, K., Sun, H., Kachlicka, M., Alayo, J. R. C., Nakata, T., & Tierney, A. (2022). Domain-general auditory processing explains multiple dimensions of L2 acquisition in adulthood. *Studies in Second Language Acquisition*, *44*, 57–86. https://doi.org/10.1017/S0272263120000467

[7] Shao, Y., Saito, K., & Tierney, A. (2022). How Does Having a Good Ear Promote Instructed Second Language Pronunciation Development? Roles of Domain-General Auditory Processing in Choral Repetition Training. *TESOL Quarterly*.

[8] Zheng, C., Saito, K., & Tierney, A. (2022). Successful second language pronunciation learning is linked to domain-general auditory processing rather than music aptitude. *Second Language Research*, *38*, 477–497. https://doi.org/10.1177/0267658320978493

English vowel perception by Kichwa–Spanish bilingual speakers and Ecuadorian Spanish speakers

Daniel Sebastian Romero Jaramillo, Joaquín Romero and Cristina Belen Crison Chavez
Universitat Rovira i Virgili, Spain

Keywords* — *Kichwa, English vowel perception, language learning.

I. Introduction

The current study aims to evaluate the perception of English high-front vowels, low vowels, and high-back vowels by Kichwa – Spanish bilingual speakers and Ecuadorian Spanish speakers. In Ecuador, Spanish is the official language, while Spanish, Kichwa, and Shuar are official languages of intercultural relation. Because of that, there are EFL learners who are Kichwa–Spanish bilinguals as opposed to EFL learners who are monolingual Spanish native speakers. The phonemic inventory of Kichwa contains twenty-seven phonemes [1] including three vowels, /i/, /a/ and /u/; the vowels /e/ and /o/ are used in loanwords from Spanish [2]. The phonemic inventory of Spanish distinguishes vowels in terms of tongue height and frontness, resulting in five vowels: /i/, /e/, /a/, /u/ and /o/ [3]. The inventories of English and Spanish exhibit more widely dispersed peripheral vowels compared to the less dispersed vowel inventory of Kichwa [4]. The Speech Learning Model [5] suggests that the phonological systems of a bilingual's two languages exist within the same 'phonological space'. It hypothesizes that a new phonetic category can be established for an L2 sound that differs phonetically from the closest L1 sound if learners discern at least some of the phonetic differences between the L1 and L2 sounds. Thus, the research question explored in this study investigated the influence of perceived phonetic differences between an English sound and its closest L1 counterpart and the ability to distinguish phonetic variations between the two sounds in bilingual Kichwa-Spanish and monolingual Ecuadorian Spanish learners of English.

II. Experimental desing

The participants in this study consisted of forty-eight undergraduate students from various majors at Universidad Técnica de Ambato in Ecuador, all of whom were required to attain a B1 level of English proficiency as a prerequisite for graduation. On average, these students had received one year of formal English instruction at the university level. The Kichwa-Spanish bilinguals had been exposed to English during their high school education, and half of them also received exposure during primary school. In contrast, the Spanish monolinguals had enrolled in English instruction throughout both primary and secondary education. It is important to note that none of the participants reported having lived in an English-speaking country prior to this study. The participants were divided into two groups: Kichwa-Spanish bilinguals and Spanish monolinguals. They were exposed to two experiments in which /ɪ- i/, /æ- ɑ/, and /u- ʊ/ phonetic contrasts were evaluated. These contrasts were tested since they would likely fall within the range of variations of the closest corresponding Spanish or Kichwa phoneme. The first task involved an ABX discrimination task using E-Prime 2.0 software. Participants listened to twenty-one pairs of words containing /ɪ- i/, /æ- ɑ/, and /u- ʊ/ phonetic contrasts. After hearing each pair twice, they were presented with a third word and asked to identify if the vowel sound was similar to the first or second word they heard. The second experiment developed an AX task where participants were exposed to twenty-one pairs of words containing one of /ɪ- i/, /æ- ɑ/, and /u- ʊ/ contrasts, and they had to determine if the vowel sounds were the same or different. The entire experiment was completed within fifteen minutes for each participant.

III. RESULTS

To evaluate the data collected, two generalized linear mixed model analyses were run. Based on the results of the ABX discrimination task shown in Table 1, a significant effect of Vowel was observed ($p < .001$), indicating that participants discrimination performance varied within the different vowel contrasts presented in the task. Indeed, participants may have greater difficulty discriminating between /ɪ- i/ and /æ- ɑ/ vowel sound pairs than between /u- ʊ/ vowel sound pairs. However, there was no significant effect of Group ($p = 0.408$), suggesting that there were no overall differences in discrimination performance between the Kichwa-Spanish bilinguals and Spanish monolinguals. Additionally, the interaction effect between Vowel and Group was not significant ($p = 0.822$), indicating that the relationship between vowel discrimination performance and language background did not differ significantly within the different vowel contrasts. This may imply that, although Kichwa speakers have expanded their vowel inventory because of the influence of Spanish, they are still not able to differentiate English vowels consistently. Based on the results of the AX discrimination task in Table 2, it can be evidenced that there was no significant main effect of Vowel ($p = 0.372$), indicating that participants' discrimination performance did not significantly differ across the different vowel contrasts presented in the task. However, a significant main effect of Group was observed ($p = 0.024$), suggesting that there were overall differences in discrimination performance between the Kichwa-Spanish bilinguals and Spanish monolinguals, which indicates that the Spanish monolinguals performed better in this discrimination task. This group was able to recognize /u- ʊ/ more effectively than Kichwa-Spanish bilinguals. Furthermore, the interaction effect between Vowel and Group was not significant ($p = 0.391$). Overall, the results from the ABX task suggest that participants could distinguish different vowel sounds, but whether someone was bilingual, or monolingual did not make a big difference in the manner they achieved it. Besides, the language background was not determinant when it comes to differentiate sounds in this task (Figure 1). On the other hand, overall findings of the AX task suggest language background may play a role in the discrimination of sounds in this task (Figure 2). Further analysis might be necessary to fully understand the relationship between language background and vowel discrimination in this context.

TABLE I. RESULTS OF THE GENERALIZED MIXED-MODELS ANALYSES FOR ABX DISCRIMINATION TASK

ANOVA Summary			
Effect	**df**	**ChiSq**	**p**
Vowel	2	21.603	< .001
Group	1	0.684	0.408
Vowel ✻ Group	2	0.392	0.822

TABLE II. RESULTS OF THE GENERALIZED MIXED-MODELS ANALYSES FOR AX DISCRIMINATION TASK

ANOVA Summary			
Effect	**df**	**ChiSq**	**P**
Vowel	2	1.989	0.372
Group	1	5.060	0.024
Vowel ✻ Group	2	1.881	0.391

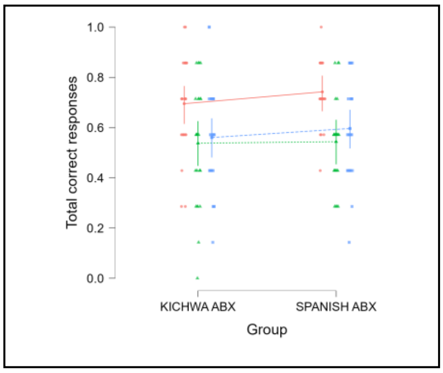


Fig. 1. Scatterplot for the ABX discrimination task (High-back vowels in red, high-front vowels in green, low vowels in blue)

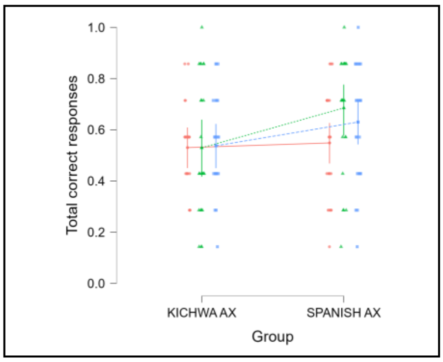


Fig. 2. Scatterplot for the AX discrimination task (High-back vowels in red, high-front vowels in green, low vowels in blue)

References

[1] C. Orr, “Ecuador Quichua Phonology” in *Ecuadorian Indian languages: I,* USA: Summer Institute of Linguistics of the University of Oklahoma, 1962, pp. 60-77.

[2] F. Chango and S. Marlett. (2008, Jul). Salasaca Quichua. *Journal of the International Phonetic Association*. *vol. 38*, pp. 223-227. DOI: https://doi.org/10.1017/S0025100308003332.

[3] C. Salcedo, (2010, Oct) The phonological system of Spanish. *Revista de Linguística y Lenguas Aplicadas. vol. 5*, pp. 195-209. DOI: https://doi.org/10.4995/rlyla.2010.769

[4] S. Guion, (2003). The Vowel Systems of Quichua-Spanish Bilinguals. *Phonetica*, *vol. 60(2)*, pp. 98-128. DOI: https://doi.org/10.1159/000071449

[5] J.E. Flege, “Second language speech learning: Theory, findings and problems” in *Speech perception and linguistic experience: Issues in Cross-Language Research*, Timonium, MD: York Perss, 1995, ch. 8, pp. 233-277.

Enhancing Computer-Assisted Pronunciation Training (CAPT) with Hybrid and End-to-End Children ASR Models

Aditya Kamlesh Parikh, Cristian Tejedor-García, Catia Cucchiarini and Helmer Strik
Center For Language Studies (CLS), Radboud University Nijmegen, The Netherlands

Keywords — computer assisted pronunciation training (CAPT), hybrid ASR, end-to-end ASR, feature extraction, MFCC, XLS-R

I. Introduction

Computer-assisted pronunciation training (CAPT) for non-native children involves using speech technology to help improve the pronunciation of non-native speakers, particularly children [1]. In CAPT, children are given specific sentences or phrases to practice pronouncing [2].

Hybrid automatic speech recognition (ASR) [3] models combine neural network techniques with statistical methods, offering high accuracy and lower latency. With a more limited search space for words, they are ideal for tasks like CAPT. However, in general, the amount of non-native children's speech data is very limited, and to create a better phoneme recognition model, a large amount of data is required [4]. On the other hand, self-supervised pretrained models have demonstrated superior performance when fine-tuned on smaller amounts of training data compared to hybrid models. Self-supervised pretrained models leverage large-scale unlabeled data to learn robust language representations [5]. When fine-tuned on smaller labeled datasets, they adapt effectively, achieving competitive performance with less data. While self-supervised pre-training models offer advantages, they can generate non-lexical words and hallucinations when dealing with under-resourced languages and limited text and speech data. This undermines their effectiveness in phoneme recognition for CAPT.

Hybrid ASR models use Mel-frequency cepstral coefficients (MFCCs) as acoustic features, which are based on a model of the adult vocal tract and based on how the human auditory system works. These features may not capture the nuances of the child vocal tract as accurately because of their inherent information loss [6]. On the other hand, self-supervised learning (SSL) end-to-end models like XLS-R, which have been fine-tuned on a small amount of children's speech data can generate high-quality vector representations of input speech audio sequences and that can offer features that are more tailored to the characteristics of children's speech [7].

In this research, our aim is to improve the Phoneme Error Rate (PER) and consequently enhance pronunciation error detection in CAPT systems by leveraging the strengths of both end-to-end and hybrid ASR systems. We address the following research questions: RQ1: Can the internal layers of self-supervised pre-training models provide any insights into the development of phonetic and phonological properties of speech? With this understanding, RQ2: Can we create a tailored and robust hybrid phoneme recognition model?

In this study we will use JASMIN [8]; a children’s read speech dataset of the Dutch language as a training and testing material. We will compare two feature extraction approaches:

(1) The first method involves using Mel-frequency cepstral coefficients (MFCC features), which are traditional acoustic features commonly used and create a phoneme recognition with hybrid ASR. We consider this as our baseline model.

(2) The second method uses features from a self-supervised model called XLS-R, fine-tuned on a small amount of children's speech. Then, these features are used to train a hybrid ASR model for phoneme recognition.. This method has shown improvement on word based under-resourced languages in ASR tasks, and can be well implemented in phoneme recognition with a hybrid ASR model in CAPT.

Acknowledgement

This work was supported by the NWO research programme AiNed Fellowship Grants under the project Responsible AI for Voice Diagnostics (RAIVD) - NGF.1607.22.013.

References

[1] M. G. Ambra Neri Ornella Mich and D. Giuliani, ‘The effectiveness of computer assisted pronunciation training for foreign language learning by children’, Computer Assisted Language Learning, vol. 21, no. 5, pp. 393–408, 2008.

[2] F. Pichette, L. de Serres, and M. Lafontaine, ‘Sentence Reading and Writing for Second Language Vocabulary Acquisition’, Applied Linguistics, vol. 33, no. 1, pp. 66–82, 09 2011.

[3] D. Povey et al., ‘The Kaldi speech recognition toolkit’, in IEEE 2011 workshop on automatic speech recognition and understanding, 2011.

[4] T. Patel and O. Scharenborg, ‘Improving End-to-End Models for Children’s Speech Recognition’, Applied Sciences, vol. 14, no. 6, 2024.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, ‘wav2vec 2.0: A framework for self-supervised learning of speech representations’, Advances in neural information processing systems, vol. 33, pp. 12449–12460, 2020.

[6] P. Mermelstein, ‘Articulatory model for the study of speech production’, The Journal of the Acoustical Society of America, vol. 53, no. 4, pp. 1070–1082, 1973.

[7] L.-M. Lam-Yee-Mui, L. O. Yang, and O. Klejch, ‘Comparing Self-Supervised Pre-Training and Semi-Supervised Training for Speech Recognition in Languages with Weak Language Models’, in Proc. INTERSPEECH 2023, 2023, pp. 87–91.

[8] C. Cucchiarini, J. Driesen, H. Van Hamme, and E. P. Sanders, ‘Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus’, 2008.

Examining L2 Vowel Contrast Perception in the EFL Classroom – A Study on Spanish/Catalan Speaker Acquisition of English Tense and Lax Vowels

Adrián Salcedo, Joaquín Romero
Universitat Rovira i Virgili, Spain

Keywords — desensitization hypothesis, cue-weighting of duration, phonological acquisition, tense, lax

I. Introduction

Non-native cue-weighting of acoustic features poses a significant obstacle to accurate vowel perception. While language transfer is responsible for a great many non-native perception and production errors, the speaker's L1 does not represent the sole source of non-native errors [1]. The literature has largely explored non-transfer production errors that can be found across speakers of different languages. However, further studies are yet to examine perception errors that do not hinge on the speaker's L1. This line of research requires more attention, especially in accounting for the non-native reliance on acoustic cues that are not phonologically contrastive in either the speaker's L1 or L2. Specifically, studies have shown that non-natives with different L1 backgrounds rely on duration contrasts to distinguish between tense and lax vowel pairs, instead of spectral contrasts [2, 3, 4]. This phenomenon is accounted for in Bohn's desensitization hypothesis, which posits that non-native speakers will resort to vowel length as the main contrastive feature for vowel pairs that non-native speakers lack in their phonological vowel inventory or for which they only have one vowel category.

Bohn's hypothesis has garnered empirical support and recent studies have researched the non-native cue-weighting of duration in vowel perception and the effectiveness of training in redressing this issue [5, 6]. However, the literature has largely neglected the population of non-adult, naïve learners who have not received significant L2 input or pronunciation training.

The current study examines the cue-weighting of duration and spectrum in the perception and production of English tense and lax vowels and the effect of three pronunciation training approaches (implicit, quality-based, length-based) on the perception and production of these vowels. The experiment included 43 bilingual Catalan/Spanish subjects (13- 14-year-olds) with an A2 or B1 language level, according to the CEFR [7], who had not received any prior pronunciation training. The subjects were EFL students at a secondary school and were divided randomly into three groups according to pronunciation training approaches.

II. Methodology

A. Stimuli and procedure

The experiment aimed to gather perception and production data following a pre-test/post-test design. The perception tests (pre-test and post-test) consisted of a forced-choice identification task with 50 randomized stimuli containing acoustically manipulated tense and lax vowels /i/ and /ɪ/ ("beat" and "bit"). The stimuli were recorded by a native English speaker and analyzed with Praat to measure the F1 and F2 values as well as duration of the tense and lax vowels. These measurements were used to manipulate the stimuli in five equal steps in duration and quality (F1 and F2), which resulted in a continuum of 25 stimuli ranging from the original F1, F2, and duration values of "beat" to those of "bit". The production tests consisted of recording the subjects' pronunciation of eight minimal pairs containing tense-lax vowels /i/-/ɪ/ and /u/-/ʊ/. In this paper, only the results of the perception tests are reported.

B. Pronunciation training

After the pre-tests, all groups were administered four sessions of one-hour-long pronunciation training. The pronunciation instruction was divided into three approaches and groups: length-based, quality-based, and implicit. All the groups received the same instruction but the first and second groups were introduced to "long" and "short" vowels and "tense" and "lax" vowels, respectively. The length-based approach aimed to assess how classifying vowels as "long" and "short", as in the British tradition, can impact vowel perception. Each training consisted of a short presentation and activities to practice perception and production. The sessions aimed to raise awareness of contrasts between English and Spanish/Catalan pronunciation, provide articulatory and acoustic descriptions of English vowels, and practice pronunciation through production activities and vowel discrimination activities.

III. Results

A linear mixed-models analysis on the perception test showed a significant reliance on vowel duration at specific steps of the continuum across all groups. The length manipulation had a significant effect on vowel perception at the first (*p = .029*), third (*p = .016*), and fourth (*p = .014*) steps of the continuum. Moreover, the pronunciation instruction also significantly affected vowel perception (*p = .033*) of all groups at the first step of the continuum. This indicates that participants correctly identified manipulated vowels in the first step more often in the post-test than in the pre-test, even though vowel duration still influenced their vowel perception. The results also showed a significant interaction of Time * Quality for the third (*p = .027*) and fourth (*p = .024*) steps of the duration manipulations of the stimuli. This means that vowel quality had a more significant effect on the subjects' vowel perception of some stimuli after the pronunciation training sessions.

Table I shows the pre-test's estimated marginal means of the subjects' vowel identification of the stimuli. The table illustrates how the subjects' vowel identification changes as the duration of the vowel shortens. Table II shows how this trend seems to hold in the post-test, albeit to a marginally lesser extent, with slightly more accurate identification of vowel quality towards the end of the duration axis.

These results seem to point to a relationship between length and non-native vowel perception which cannot have resulted from L1 transfer, since neither Spanish nor Catalan have contrastive phonological duration. Likewise, this reliance on duration cannot be ascribed to L2 experience because the subjects were naïve learners who had no knowledge of English phonetics or, specifically, the British tradition that classifies vowels by length ("long" and "short" vowels). Thus, the duration contrasts were more salient than the spectral ones, contrary to what L1 transfer accounts would predict. These findings are consistent with Bohn's hypothesis in that, even though the subjects rely solely on spectra, not duration, to distinguish native vowels, they resorted to length contrasts in their perception of English tense and lax vowels. According to the desensitization hypothesis, in the absence of salient spectral contrasts, speakers will rely on duration to distinguish non-native vowels because they have not been sensitized to such qualitatively different categories.

The analysis of the production data gathered will produce further insights on the non-native appreciation of duration as a significant feature in producing tense and lax vowels in English.

Table I. Estimated marginal means for pre-test (three groups)

Quality	Duration				
	long				short
beat	1.004	1.052	0.863	0.7	0.774
	1.189		1.152	0.926	0.852
	1.293	1.374	1.167		0.904
	1.326	1.356		0.993	
bit	1.374	1.319	1.1	1.237	1.03

● beat 2

○ bit 0

Table II. Estimated marginal means for post-test (three groups)

Quality	Duration				
beat	1.167	1.144	1.278	1.07	0.863
	1.185	1.341	1.03	1.156	1.07
	1.044	1.193	0.956	1.215	0.993
	1.233	1.007	0.804	1.181	1.041
bit	1.326	1.133	1.189	0.933	1.122

● beat 2

○ bit 0

References

[1] O. S. Bohn, "Cross language speech perception in adults first language transfer doesn"t tell it all", in *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, W. Strange, Ed. New York Press, pp. 279-304, 1995.

[2] C. Aliaga-García, The effect of Auditory and Articulatory Phonetic Training on the Perception and Production of L2 Vowels by Catalan/Spanish Learners of English (Doctoral dissertation, Universitat de Barcelona), 2017.

[3] H. K. D. Souza, A. Carlet, I. A. Jułkowska and A. Rato, "Vowel inventory size matters: Assessing cue-weighting in L2 vowel perception". Ilha do Desterro, 70, pp. 33-46. 2017.

[4] E. Cerviño-Povedano and J. C. Mora, "Spanish EFL learners' categorization of /i:-I/ and phonological short-term memory". Procedia-Social and Behavioral Sciences, 173, pp. 18-23 2015.

[5] A. Carlet and H. K. D. Souza, "Improving L2 pronunciation inside and outside the classroom: Perception, production and autonomous learning of L2 vowels". Ilha do Desterro, 71, pp. 99-123, 2018.

[6] L. Yu, The Perceptual Cue Weighting of English Tense-lax Vowel Contrasts by First Language (L1) and Second Language (L2) Speakers of English (Doctoral dissertation, Indiana University). 2023

[7] Council of Europe (2001). Common European Framework of Reference for Language Learning, Teaching, Assessment. Cambridge University Press. https://rm.coe.int/1680459f97

Variation of vowel /ɤ/ in the speech of teenagers in Saaremaa: A case study

Pire Teras, Kristiina Praakli
University of Tartu, Estonia

Keywords — the vowel /ɤ/, vowel variation, teenager's talk, Saaremaa Dialect and Accent, Estonian

I. Introduction

As innovative and creative language users, teenagers are of interest to language researchers for many reasons [1]. In the case of language usage from dialectal areas, spoken language data also allow for observing local pronunciation features. As a pilot study, this paper examines the language usage of two 12-year-old teenagers from Saaremaa. Saaremaa is situated in the Baltic Sea in the western part of Estonia and is the largest island of Estonia.

In a situation where the dialects of the Estonian language are levelling, features that define the accent of regional language use, can still be found. The accent of Saaremaa is known for some phonetic features, e.g. labialization of the vowel /ɤ/, and sing-songy melody [2].

The main focus of this research lies in one of the phonetic features characteristic of the Saaremaa accent: the variation of the unrounded back vowel /ɤ/. The vowel /ɤ/ occurs in most language varieties within the Estonian language area (see [3]). In the dialects spoken on the western Estonian islands, the usage of /ɤ/ varies. The variation of the /ɤ/ is also the most notable pronunciational difference between eastern Saaremaa, and western and central Saaremaa (see [2], [4]). Ellen Niit's acoustic-phonetic study on the pronunciation of Saaremaa vowels by older speakers showed that /ɤ/ has merged with the labial front vowel /ø/ [5].

This study aims to find out if the merger of these two vowels can also be considered a characteristic feature of the Saaremaa accent among teenage speakers. How much does the pronunciation of /ɤ/ vary? Can there be found variants in the language use of teenagers, where /ɤ/ distinguishes from /ø/ and has not merged with it?

II. Material and Method

Analysed data for the pilot study were extracted from the spoken corpus compiled in 2019–2022 within the "Teen Speak in Estonia" (TeKE) project [6]. The citizen science approach in data collection was used [7]. The corpus contains 97 hours of conversations from 131 participants aged 10–18 from various regions in Estonia. About 26 hours of recordings have been collected in Saaremaa. ELAN annotation software [8] was used to transcribe the audio files of the conversations.

For this study, we used an everyday conversation between two 12-year-old boys from Saaremaa, which lasted approximately one hour. The main topics discussed during the conversation were computer games, football and school. The informants themselves made the recording. However, the conversation was led by Speaker 1, taking the role of interviewer, which explains the between-speaker difference in the number of analysed words containing vowels /o/, /ɤ/, /ø/. The vowels were searched for using Praat [9]. Table I provides the number of analysed vowels. The vowels were analysed acoustically, and the values of the first three formants from the middle of the vowels were extracted using Praat's script. Statistical analysis was carried out in R [10]. Average formant values and standard deviations were calculated, and both short and long vowels were analysed together. For the analysis, the ggplot2 package was used [11].

TABLE I. Number of analysed vowels by speaker

Vowel	**Speaker**	
	SP1	***SP2***
/o/	12	31
/ɤ/	27	76
/ø/	3	21

III. Results

Some results are presented next. Fig. 1 presents formant values of F1 and F2 of all analysed vowels: *o* /o/, *õ* /ɤ/, and *ö* /ø/. Average values have also been provided, along with standard deviation ellipses. The standard deviation ellipse could not be shown for the vowel /ø/ of SP1 due to the limited number of tokens.

The quality of the vowel /ɤ/ varies notably in the pronunciation of both speakers. However, Speaker 1 had only three tokens containing the long vowel /ø/. Nevertheless, as seen in Fig. 1, his pronunciation exhibits both variants of /ɤ/ that are close to /ø/ or even more front than /ø/, as well as variants that are close to the back vowel /o/. The standard deviation also shows considerable variation in the values of F2 of /ɤ/ (around 180 Hz). The average formant values for F2 of /ɤ/ and /ø/ are 1406 and 1535 Hz, respectively.

The variation in the pronunciation of Speaker 2 is quite large: standard deviations of F2 for /ɤ/ and /ø/ are 141 Hz and 99 Hz, respectively. However, the standard deviation ellipses do not completely overlap. Some variants of /ɤ/ are close in quality to /ø/, while others have been pronounced more backward, resembling the quality of [ɤ]. The average formant values for F2 of /ɤ/ and /ø/ are relatively close: 1384 Hz and 1458 Hz, respectively.

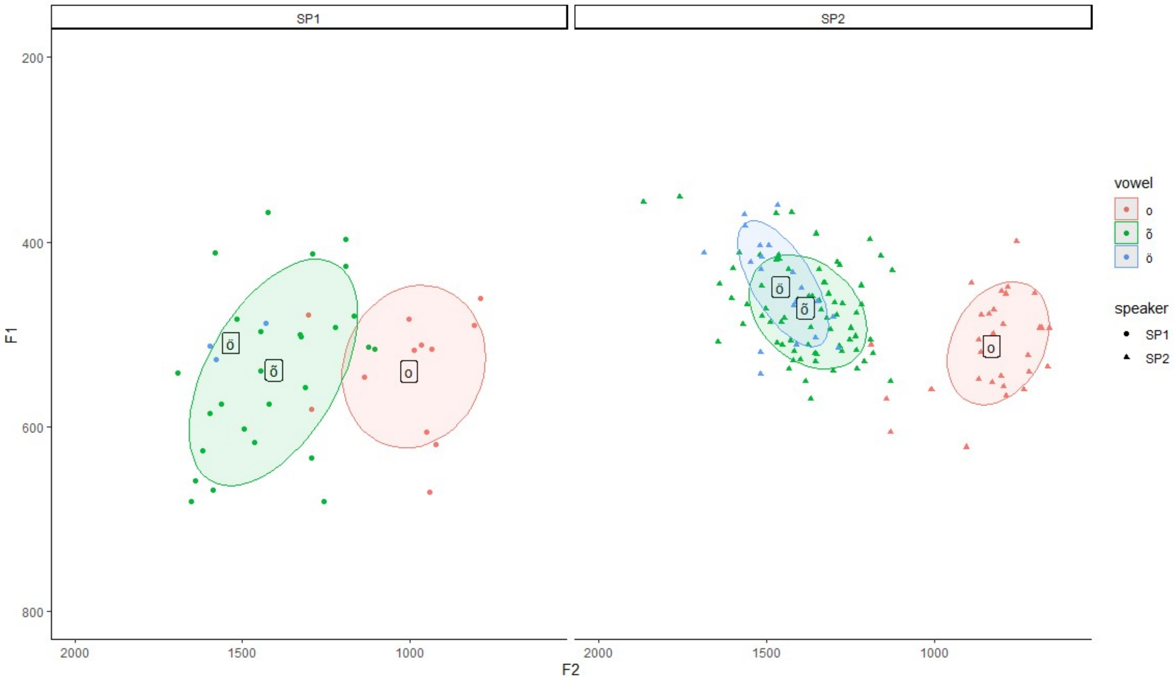


Fig. 1. Vowels *o* /o/, *õ* /ɤ/, and *ö* /ø/ in the F1–F2 (Hz) space: all vowel points, mean values and standard deviation ellipses (left panel – Speaker 1, right panel – Speaker 2)

IV. Conclusions

Some conclusions can be drawn from the analysis so far. It appears, that the merger of the vowel /ɤ/ with the vowel /ø/ can still be considered a characteristic feature of the Saaremaa accent. However, in the pronunciation of these two teenage speakers, the quality of /ɤ/ varies considerably: some variants are pronounced close to [ø], while others are pronounced more backward and are closer in quality to [ɤ]. These results will be discussed in more detail in the paper.

References

[1] P. Eckert, *Language Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Hoboken: Wiley-Blackwell, 2000.

[2] E. Niit, 'Eesti keele piirkondlikud aktsendid', *Lähivõrdlusi. Lähivertailuja*, vol. 15, pp. 50–61, 2004.

[3] P. Teras, '/õ/ hääldlduse varieerumine', *Emakeele Seltsi aastaraamat*, vol. 63, pp. 221–239, 2018, doi: 10.3176/esa63.10.

[4] K. Praakli and T. Pae, 'Õ', *Keel ja Kirjandus*, vol. 64, no. 12, pp. 1035–1057, 2021, doi: 10.54013/kk768a1.

[5] E. Niit, 'Vowels in Initial Syllables in Saaremaa', *Linguistica Uralica*, vol. 41, no. 2, pp. 113–122, 2005, doi: 10.3176/lu.2005.2.05.

[6] V.-A. Vihman, K. Praakli, M.-L. Pilvik, and M.-L. Korkus, 'Kas noored on inglise keelega "obsessed"? Millest räägivad korpusandmed?', *Philologia Estonica Tallinnensis*, no. 7, 2022, doi: 10.22601/PET.2022.07.11.

[7] K. Koreinik, A. Mandel, M.-L. Pilvik, K. Praakli, and V.-A. Vihman, 'Outsourcing teenage language: a participatory approach for exploring speech and text messaging', *Linguistics Vanguard*, vol. 9, no. 4, pp. 389–398, 2023, doi: 10.1515/lingvan-2021-0152.

[8] Max Planck Institute for Psycholinguistics, The Language Archive, 'ELAN [Computer software].'. Nijmegen, 2022. [Online]. Available: https://archive.mpi.nl/tla/elan

[9] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer [Computer program].' 2022. Accessed: Dec. 17, 2022. [Online]. Available: http://www.praat.org/

[10] R Core Team, 'R: A Language and Environment for Statistical Computing'. Vienna, Austria, 2022. [R Foundation for Statistical Computing]. Available: https://www.R-project.org/

[11] H. Wickham, 'ggplot2: Elegant Graphics for Data Analysis'. Springer-Verlag, New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

Duration and Declination in L2 Reading

Bojia Wang[a], Dafydd Gibbon[b]
[a] *Xi'an International Studies University, China*
[b] *Bielefeld University, Germany*

Keywords — duration, declination, prosody, second language learning, computer assisted language assessment, CALA, speech modulation theory, global intonation contour, Support Vector Machine, SVM

1. Background

In the extensive computer assisted language learning (CALL) literature on L2 intonation, much attention has been paid to the teaching and learning of stress and pitch accents, terminal contours and rhythm, from both phonological [1] and phonetic [2] perspectives. Computer assisted language assessment (CALA), can offer support for diagnostic prosody assessment: while communication goals are paramount, communication deteriorates with poor pronunciation. Differences in L1 and L2 utterance duration and fundamental frequency (F0) slope ('declination', 'inclination' etc.) are compared automatically in the present study, with a 'no difference' null hypothesis. This contrasts with earlier studies of declination:duration ratio [3, 4], pitch accents [5], or global height and range parameters [6].

A CALA pipeline was constructed, from F0 estimation and modelling through descriptive statistics to ML classification of L2 and target (e.g. L1) productions, with the medium-term aim of providing assistance for grading L2 speakers with probabilistic information [7] on L1-likeness, using (among other techniques) Support Vector Machine (SVM) classification. A similar SVM method has been applied to F0 contours in order to classify personalities [8].

2. Data and Method

L1 British English data are sourced from the IVIE corpus [9], and consist of 64 readings of 5 disjunctive questions ('OR-questions', chosen here for their complex 2-level intonation structure) by 13 female native speakers. The L2 data consist of 120 recordings of the same sentences by 24 female advanced Chinese EFL students with 12 years of school and university English, who are native speakers of Mandarin and Shaanxi dialect. The speakers are of the same sex so that the same F0 estimation parameters could be used for all participants.

The questions are: Q1, *Are you growing limes or lemons?*; Q2, *Is his name Miller or Mailer?*; Q3, *Did you say mellow or yellow?*; Q4, *Do you live in Ealing or Reading?*; Q5, *Did he say lino or lilo?* The sentences were recorded independently by the L2 students on equipment of their choice, mainly Praat [10] on laptops, a familiar testing scenario for them. Sampling rates varied, so recordings were resampled to 16 kHz. Initial and final silences were cropped to about 100ms and recordings were normalised to unit amplitude. Q4, which is voiced throughout, is used for illustration in this contribution. Several conspicuous L1-L2 differences are already apparent in Figure 1 (e.g. amplitude pattern, in addition to longer duration and flatter or uptrend global slope for L2).

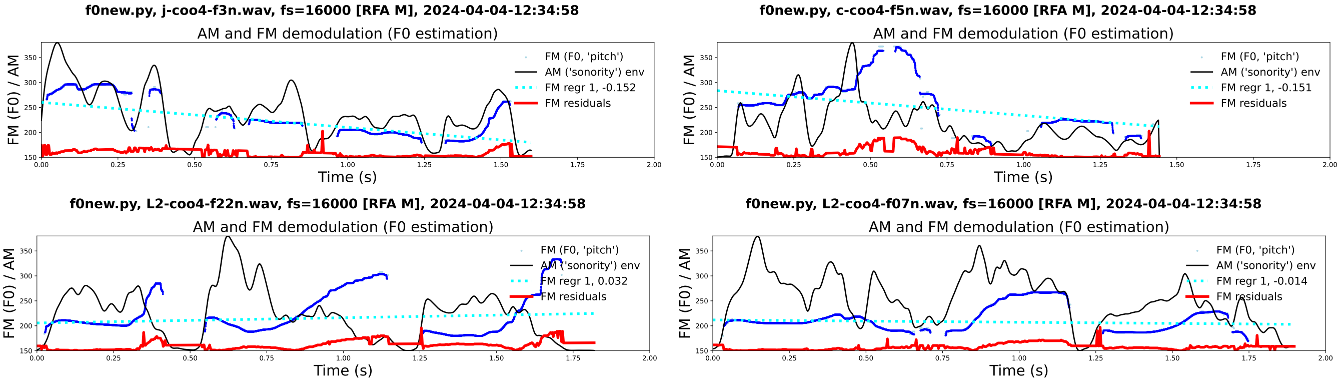


Figure 1: Q4, L1 (top), L2 (bottom), different readers: AM envelope, FM envelope, FM regression, residuals.

The methodological foundation is speech modulation theory [11], in which a carrier frequency is modulated by information signals, and which provides an integrated framework for all areas of phonetics. Modulation theory is as old as radio, and the terminology is the same as the labels on radio sets, but with a prosodic phonetic interpretation: FM (for frequency modulation of fundamental frequency, F0, in the larynx; a tone, pitch accent and intonation correlate) and AM (for amplitude modulation of speech sound and rhythm formants, by naso-oral filter; a sonority correlate):

$$\boldsymbol{Speech} = \boldsymbol{A_{AM}} Acos(2\pi(f + \boldsymbol{A_{FM}})t + \phi)$$

For example, rhythms use LF (low frequency) AM information signals, and intonation, tone and pitch accent use LF FM information signals, both with modulation frequencies which are in general below about 5 Hz. The focus in this study is on the FM information signal (f=frequency, t=time, A_{FM}=FM amplitude, A_{AM}=AM amplitude, ϕ=phase); phase is not treated further. The demodulation and modelling pipeline has the following steps:

1. AM demodulation: Amplitude envelope, extracted and smoothed, for visual time-frequency alignment.
2. Preliminary noise reduction: centre and peak clipping (10%).
3. Tuning: 3rd order Butterworth bandpass filter 120...380 Hz.
4. FM demodulation: custom time-domain F0 estimation with contour smoothing.
5. Model: linear regression line with interpolation of voicing gaps.
6. Feature extraction: residuals, and duration, slope, intercept values.

Table 1: Averaged results for L1 and L2 F0 contours.

Var:	Dur(s)	Slope	Intercept	SDF0	SDabsres
L1 mn:	1.807	-0.074	232.538	32.801	18.817
L2 mn:	2.629	-0.028	237.272	29.559	18.528
L1 SD:	0.348	0.051	26.040	9.988	4.623
L2 SD:	0.542	0.025	23.616	7.687	4.260
t-test:	$p < .01$	$p < .01$	$p < .01$	$p < .05$	$p > .05$
dur:x (corr):	-	0.507	-0.446	-0.484	-0.302

Selected global properties (duration, slope, intercept, SD of F0, SD of absolute values of residuals) were analysed for the L1 and L2 groups. Q-Q plots and Shapiro-Wilk tests showed near-normal distributions and T-tests were applied (Table 1). Duration, slope, intercept and SD F0 variables showed significant L1-L2 differences and thus refutation of the null hypothesis. The SD of absolute residuals did not. Slope correlated moderately with duration, confirming previous peakline studies [4]. Slope results also showed a tendency for steeper slopes in L1 than in L2 (see also Figure 1), a sociophonetic register factor. These differences found by basic statistical analysis suggest that a more general ML classification scheme could be used in order to discover whether the differences can be seen as a potential model for computer-assisted intonation grading, using a generalised target set rather than one individual, for example a teacher, as reference.

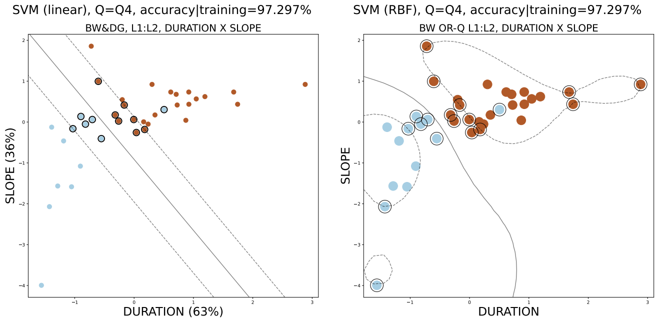


Figure 2: Q4 SVM graphs: linear (left), non-linear (right)

SVM classifiers were trained for the L1 and L2 classes and z-score standardised duration and slope features (Figure 2). The scatter plot alone shows L1-L2 differences: skewed, variable downward slope, shorter duration for L1 (vertical scatter), but flat or upward slope, skewed, variable longer duration for L2 (horizontal scatter). SVM analysis shows that duration contributes 63%, slope 36% to the difference in Q4; both slope and duration are thus potential indices of L1-L2 differences. Hyperplane accuracy is well above chance, again refuting the ‘no difference’ null hypothesis: Q1: 94.44%, Q2: 81.08%, Q3: 83.78%, Q4: 97.30%, Q5: 91.89%, full dataset: 90.22%.

3. Discussion and conclusions

Duration and global F0 slope in L1 and L2 readings of disjunctive ‘OR-questions’ are distinguished with good accuracy in this small database, statistically and by SVM. Reasons for the difference, such as 20-year L1-L2 time lapse, L1 interference, L2 uncertainty, need further research. Practical applications will require larger and more varied datasets and more features [6], and large language models will become more relevant. However, the results suggest F0 properties for learner feedback, with likelihood of prosodic target-likeness as an index of intonation proficiency.

References

[1] I. Mennen, “Beyond segments: Towards a l2 intonation learning theory,” in *Prosody and languages in contact: L2 acquisition, attrition, languages in multilingual situations*, E. Delais-Roussarie and M. A. . S. Herment, Eds. Springer Verlag, 2015, pp. 171–188.

[2] A. Suni, H. Kallio, Štefan Benuš, and J. Šimko, “Characterizing second language fluency with global wavelet spectrum,” in *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne: International Phonetic Association, 2019, pp. 1947–1951.

[3] D. Kocharov, N. Volskaya, and P. Skrelin, “F0 declination in Russian revisited,” in *18th International Congress of Phonetic Sciences*, 2015.

[4] J. Yuan and M. Liberman, “F_0 declination in English and Mandarin broadcast news speech,” *Speech Communication*, vol. 65, pp. 67–74, 2014.

[5] A. Rosenberg, “Classification of prosodic events using quantized contour modeling,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 2020, p. 721–724.

[6] H. Ding, R. Hoffmann, and D. Hirst, “Prosodic transfer: a comparison study of F0 patterns in L2 English by Chinese speakers,” in *Speech Prosody 2016*, 2016, pp. 756–760.

[7] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[8] U. D. Reichel, “Personality prediction based on intonation stylization,” in *18th International Congress of Phonetic Sciences*, 2015.

[9] E. Grabe and B. Post, “Intonational variation in the British Isles,” in *Proceedings of the International Conference on Speech Prosody*, P. Gilles and J. Peters, Eds., 2002.

[10] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[11] H. Traunmüller, “Conventional, biological, and environmental factors in speech communication: a modulation theory,” *Phonetica*, vol. 51, no. 1-3, pp. 170–183, 1994.

The impact of visual and auditory training on the production of r-colored vowels by Catalan/Spanish bilingual learners of English as a foreign language

Cristina Crison, Joaquín Romero, Leonardo Oliveira, Daniel Romero

Rovira i Virgili University, Tarragona, Spain

Keywords* — *r-colored vowels, ultrasound, pronunciation teaching.

I. Introduction

The current study aims to evaluate visual and auditory training on the production of English r- colored vowels /ɝ/ and /ɚ/ by EFL learners. In r-colored vowels, the coordination of the tongue movement related to /r/ occurs simultaneously to a central vowel [1] which makes its accurate production difficult for EFL learners. Research has suggested the benefits of multimodal feedback, which can be provided using technologies such as ultrasound in the context of second language pronunciation training [2]. A crucial contribution of ultrasound in second language acquisition is that it allows teachers and learners to see practical components of complex articulatory tasks [3]. Thus, using ultrasound to train students on the articulation of /ɝ/ and/ɚ/ might foster their accurate production. Two main research issues were explored: a. the potential for and extent of improvement that training provides on the accurate production of r-colored vowels; b. what the most effective training technique is to achieve accurate production.

II. Experimental design

The three subjects were exposed to 147 items containing stressed /ɝ/ and unstressed /ɚ/, which were presented in different contexts within a carrier phrase. The stimuli were presented in /ɝ/+C and /ɚ/+C contexts, where C were consonants produced with different manners of articulation. Data was recorded using ultrasound and audio simultaneously. Each experimental session consisted of a pre-test, a short 15-minute training and a post-test. Each participant was exposed to a specific type of training: auditory (AT), visual (VT) and a combination of auditory–visual (AVT). AT consisted of presenting 30 recordings containing /ɝ/ and /ɚ/, repeated three times. VT consisted of presenting 30 videos with no auditory component portraying the articulation of /ɝ/ and /ɚ/ as produced by a native speaker and the orthographic representation of the words, repeated three times. AVT consisted of presenting thirty videos with audio that portrayed the articulation of /ɝ/ and /ɚ/ as produced by a native speaker and the orthographic representation of the words; each video was repeated three times. During the training period the participants exposed to visual and visual–auditory input were allowed to see their articulation by using the ultrasound. In addition, data were obtained from a native speaker of North American English for comparison.

III. RESULTS

In order to reduce the number of dependent variables in the statistical tests, a factor (principal components) analysis was performed that helped identify the X and Y positions along the tongue surface that best corresponded to the back section of the tongue dorsum, the front section of the tongue dorsum and the tongue blade. Results from a series of linear mixed models analyses for the X and Y positions that resulted from the factor analyses are shown in Table 1. Significant main effects were obtained for fixed factors Training type (AT, VT and AVT) and Time (pretest vs. posttest), as well as a significant Training type x Time interaction for the X positions but only for Training type for the Y positions. Pairwise comparisons reveal that on most occasions non-native production deviates significantly from that of the control native speaker, whether in the pre-test or the post-test. This is illustrated by the tongue contours shown in Figure 1. The AT and VT groups show a lack of accuracy in the production of r-colored vowels. In most cases there is a clear sequence of back vowel + /r/ instead of the low central vowel target. There were also no cases of low central vowels without r-coloring, which might be an expected outcome for subjects aiming to hit the correct articulatory target. The post-tests evidenced similar difficulties in the achievement of the correct articulatory targets. However, there are aspects of the training period that are worth noting. In AT, the participant spontaneously adapted their speech after exposure to the items, while the VT participant evidenced varied tongue movements, possibly driven by the absence of audio in the video. Finally, the participant in the AVT group did not attempt to imitate the articulation from the videos. Overall, these findings suggest a general lack of awareness regarding the articulation of r-colored vowels that could be tackled using visual training with ultrasound. In that sense, data from additional subjects will inform as to the most effective methodology to achieve that objective.

TABLE I. General results of the mixed-models analyses for X and Y positions at three points on the tongue roughly corresponding to back dorsum (X1,Y1), mid dorsum (X2,Y2) and tongue tip/blade (X3,Y3)

X1	***df***	***F***	***p***		***Y1***	***df***	***F***	***p***
Group	3,2	11.549	0.081		Group	3,2.22	365.867	0.002
Time	1,646	67.599	<.001		Time	1,1.61	0.002	0.966
Group ✻ Time	3,2	22.251	0.043		Group ✻ Time	3,2	0.677	0.642
X2	***df***	***F***	***p***		***Y2***	***df***	***F***	***P***
Group	3,630	74.926	<.001		Group	3,2	36.074	0.027
Time	1,3.16	61.585	<.001		Time	1,1.35	0.061	0.838
Group ✻ Time	3,7.13	17.996	0.001		Group ✻ Time	3,2	3.569	0.227
X3	***df***	***F***	***p***		***Y3***	***df***	***F***	***P***
Group	3,2.88	40.053	0.007		Group	3.2	108.820	0.009
Time	1,3.58	72.232	0.002		Time	1.1.03	0.863	0.520
Group ✻ Time	3,2	28.606	0.034		Group ✻ Time	3.2	3.175	0.249

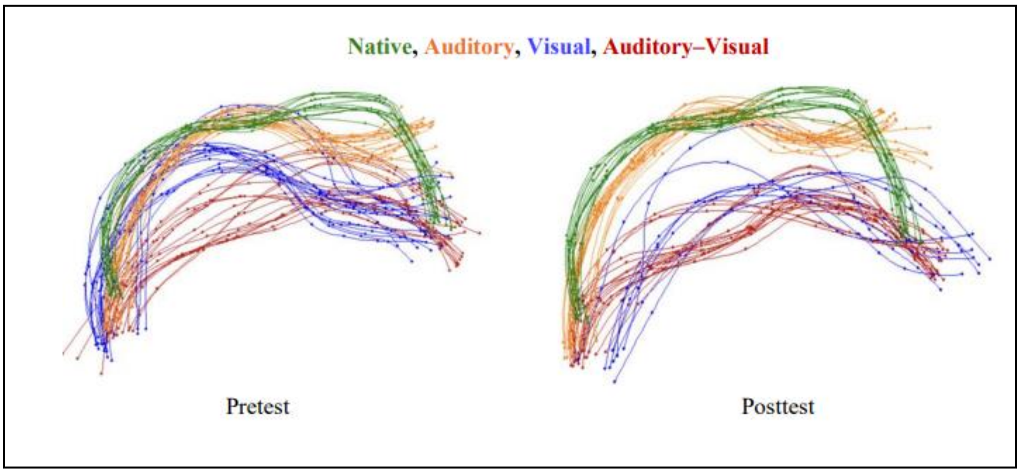


Fig. 1. Tongue contour splines comparing productions for native speaker vs. the three different training groups at pretest and posttest

References

[1] Clark, J., Yallop, C., & Fletcher, J. An Introduction to Phonetics and Phonology. Wiley-Blackwell, 2007.

[2] Abel, J., Allen, B., Burton, S., Kazama, M., Kim, B., Noguchi, M., Tsuda, A., Yamane, N., & Gick, B. "Ultrasound-enhanced multimodal approaches to pronunciation teaching and learning," Proceedings of Acoustics Week in Canada. Canadian Acoustics, 43(3), 124– 125, 2015.

[3] Gick, B., Bernhardt, B., Bacsfalvi, P., & Wilson, I. "Ultrasound imaging applications in second language acquisition". In J. G. Hansen Edwards & M. L. Zampini (Eds.), Phonology and Second Language Acquisition (pp. 309-322). John Benjamins, 2008.

Perception of Voicing Contrast of Japanese Word-Medial Plosives by Japanese and Chinese Listeners

Yixuan Huang [a], Mariko Kondo [a]
[a] *Waseda University, Japan*

Keywords — Japanese, plosive voicing, perception, VOT, closure duration

I. Introduction

Japanese plosives contrast in voicing, while Chinese plosives differ in aspiration. Previous studies have demonstrated that Chinese learners of Japanese are unable to hear prevoicing in Japanese voiced plosives, which causes difficulty in identifying the voicing of plosives in Japanese [1]. However, in addition to this problem, which is related to Voice Onset Time (VOT), the closure duration of Japanese word-medial voiceless plosives is longer than that of voiced plosives [2, 3]. This difference in closure duration has not been observed in Chinese plosives [4], and it has been shown to affect the closure duration of Japanese plosives by Chinese learners [5]. It is also likely that it will affect perception of Japanese by Chinese learners.

The role of closure duration in the voicing identification of Japanese word-medial plosives has received relatively little attention. Previous studies [1, 5] have emphasized the importance of closure duration in perception by native Japanese listeners. However, the small numbers of subjects in these studies mean that further clarification is needed to demonstrate the perceptual pattern. Nonetheless, it is hypothesized that native Japanese speakers rely on closure duration information to identify the voicing of intervocalic plosives. However, it is not clear whether native Chinese speakers can perceive differences in closure duration in a similar way to native Japanese speakers.

Therefore, this study aims to investigate the influence of VOT and closure duration on the perception of Japanese word-medial plosive voicing by native Japanese and Chinese listeners. By comparing perception patterns between Japanese and Chinese listeners, this study aims to identify difficulties faced by Chinese learners in the voicing identification of Japanese word-medial plosives.

II. Methods

Three Japanese minimal pairs of three-mora words were selected to contrast voiced and voiceless plosives in word-medial position: bilabial (*apai* 'nonsense' vs. *abai* 'nonsense'), alveolar (*jitai* 'font' vs. *jidai* 'era'), and velar (*sokai* 'evacuation' vs. *sogai* 'inhibition'). The stimuli were recorded by a female native Japanese speaker. The voiceless words in each pair were selected as the base stimuli for further manipulation, ensuring that other acoustic features, such as pitch, remained consistent between the voiced and voiceless plosives within each pair. For each pair, a continuum was generated with differing VOT and closure duration. The durations of the vowels before the target plosives were adjusted to 100 ms. For /p, b/ and /t, d/ contrasts, there were two steps in the VOT continuum (-40 ms and 10 ms). For the /k, g/ contrast, a 30 ms VOT step was added to match the longer VOTs that occur in the production of velar plosives. Closure duration steps ranged from 10 ms to 100 ms (in 15 ms increments) for positive VOTs and from 40 ms to 100 ms (in 15 ms increments) for negative VOTs. A total of 43 tokens were generated (VOT step 1 (-40 ms): 3 contrasts x 5 CD steps; VOT step 2 (10 ms): 3 contrasts x 7 CD steps; VOT step 3 (30 ms): 1 contrast x 7 CD steps).

Twenty native Chinese speakers (aged 18 to 40), with Japanese proficiency levels ranging from beginner to advanced, participated in an online experiment, along with 14 native Japanese speakers (aged 18 to 40). All participants were instructed to complete a forced-choice identification task using earphones or a headset in a quiet place. In the identification task, two visual stimuli representing the two possible responses written in Japanese *kana* syllabary were shown on the screen for each sound stimulus. The 43 generated tokens were repeated three times in random order.

III. main results

A generalized linear mixed model was fitted to the binomial voiced/voiceless response data, with *VOT* (1-3 steps, numeric), *CD* (closure duration, 1-7 steps, numeric), *NL* (native language of the listener, with Japanese as the reference level), and the *VOT* × *CD* × *NL* interaction included as predictors. *Participants* and *places of articulation* were included as random effects. Significant results were observed for the *VOT*, *NL*, *VOT* × *CD*, and *VOT* × *NL* variables. Particularly noteworthy is the significant contribution of the interaction between *VOT*, *CD*, and *NL* to the model, suggesting that the interaction of VOT and closure duration acts differently for Japanese and Chinese listeners.

To examine detailed perception patterns, voiced identification curves were plotted as a function of VOT steps (Fig. 1.) and CD steps (Fig. 2.). As indicated in Fig. 1, the voiced response rate remained high across VOT steps for the Chinese listeners, except for a slight drop at VOT step 3 (30 ms) for the velar contrast. This suggests that the Chinese listeners had higher sensitivity to aspiration. For the Japanese listeners, the voiced response rate was less than 100% at -40 ms VOT (step 1). Also, the Japanese listeners' voiced response rate varied with changes in the VOT, but the direction of change differed depending on the closure duration. Compared to the VOT of 10 ms (step 2), the VOT of -40 ms increased the Japanese listeners' voiced response rate when the closure duration was long at 100 ms (step 7) but decreased the voiced response rate when the closure duration was short at 40 ms (step 3). On the other hand, VOT of 30 ms (only for /k, g/) consistently lowered the voiced response rate regardless of the changes in closure duration.

Fig. 2. illustrates the different perception patterns for the Japanese and Chinese listeners. The voiced response rate for the Chinese listeners remained consistently high and did not appear to be affected by changes in closure duration. However, voicing identification by the Japanese listeners seems to have been greatly influenced by closure duration. The shortest closure duration step 1 (10 ms) almost always resulted in voiced responses for Japanese listeners. Generally, Japanese listeners gave fewer voiced responses as the closure duration increased.

The perception patterns of the Japanese listeners varied depending on the place of articulation of the plosives. In Fig. 2., VOT step 2 and closure duration step 7 resulted in more than 80% voiceless responses for the /t, d/ contrast, whereas there were only around 50% voiceless responses for the /p, b/ contrast. Moreover, for the /k, g/ contrast, the distance between the plots of VOT steps 2 and 3 was larger than the distance between the plots of VOT steps 1 and 2, indicating that aspiration had a greater effect than voicing. These results are supported by the post-hoc pairwise comparisons of each VOT step contrast and closure duration step contrast.

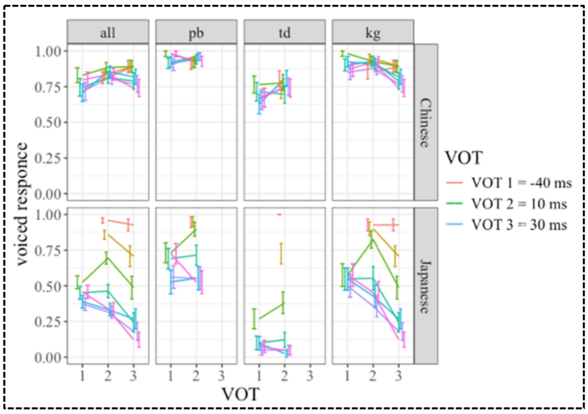


Fig. 1. Plot of voiced response rate as a function of VOT step.

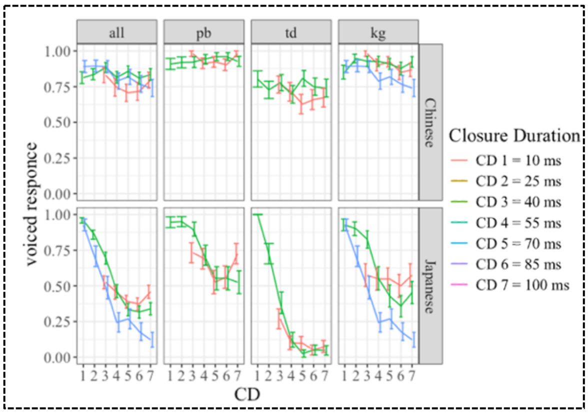


Fig. 2. Plot of voiced response rate as a function of CD step.

IV. DISCUSSION AND CONCLUSIONS

The Japanese and Chinese listeners showed different perceptual patterns in the voicing identification of Japanese word-medial plosives. Overall, the Chinese listeners gave predominantly voiced responses regardless of the VOT or closure duration. In contrast, the Japanese listeners' identification was influenced by both VOT and closure duration.

The reduction in voiced response rate by the Chinese listeners at 30 ms VOT suggests that they were more sensitive to aspiration than voicing. For the Japanese listeners, VOT strongly interacted with closure duration. The more consistent voiced response rate at 30 ms VOT compared to -40 ms VOT suggests a more consistent effect of aspiration than prevoicing. The Chinese listeners had difficulty detecting differences in closure duration, whereas the Japanese listeners were highly sensitive to these changes. As the closure duration increased, the Japanese listeners tended to give fewer voiced responses.

The insensitivity of Chinese listeners to duration differences may be due to the lack of such distinctions in the Chinese language. Similar difficulties have been observed in native Chinese speakers using vowel durational cues to identify English word-final consonants [6]. While previous research has extensively documented the inability of Chinese listeners to hear prevoicing [1], closure duration has often been overlooked. This study has demonstrated the critical role of closure duration in Japanese listeners' identification, while the lower accuracy of the Chinese listeners in differentiating Japanese word-medial plosives may be attributed in part to their reduced sensitivity to closure duration differences.

References

[1] C. Zhu, "Tyuugokugo no yuuki/muki siin to nihongo no musei/yuusei siin no seiriteki/onkyouteki/tikakuteki tokutyou to kyouiku [Educational and the Physiological, Acoustic and Perceptual Characteristics of Chinese Aspirated/Unaspirated Consonants and Japanese Voiced/Voiceless Consonants]," *Onsei gakkai kaihou* [Bulletin of the Phonetic Society of Japan], vol. 205, pp. 34-62, Apr. 1994. (in Japanese)

[2] J. Gao, T. Arai, "Plosive (de-)voicing and f0 perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery," *Journal of Phonetics*, vol. 77, p. 100932, Nov. 2019.

[3] Y. Homma, "Durational relationship between Japanese stops and vowels," *Journal of Phonetics*, vol. 9, no. 3, pp. 273–281, Jul. 1981.

[4] I. Yun, "A study of the preconsonantal vowel shortening in Chinese," *Phonetics and Speech Sciences*, vol. 10, no. 4, pp. 39–44, Dec. 2018.

[5] M. Sugito and Y. Kanda, "Nihongo wasya to tyuugokugo wasya no hatuwa ni yoru nihongo no musei oyobi yuusei haretuon siin no onkyou teki tokutyou [An acoustic study on the production of Japanese voiceless and. voiced plosives by Chinese speakers]," *Oosaka Shouin Joshi Daigaku Ronsyuu* [Journal of Osaka Shoin Women's University], vol. 24, pp. 67-89, Mar. 1987. (in Japanese)

[6] J. E. Flege, "Chinese subjects' perception of the word-final English /t/–/d/ contrast: Performance before and after training," *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1684–1697, Nov. 1989.

Stop production in Dhuwaya: Implications for literacy teaching

Kathleen Jepson [a], Rasmus Puggaard-Rode [a]
[a] *Institute of Phonetics and Speech Processing, Ludwig Maximilian University of Munich, Germany*

Keywords — stop contrasts, VOT, literacy, Australian Indigenous languages, Dhuwaya

I. Introduction

In Dhuwaya, like other Yolngu languages of northern Australia, there are six places of articulation for stops (bilabial, dental, alveolar, retroflex, palatal, velar) [1]. In related languages, e.g., Djambarrpuyŋu, there is a fortis/lenis contrast for all places [2]; however, in Dhuwaya the contrast exists only for alveolar and retroflex stops [1]. Generally, in Yolngu languages, the contrast carries a low functional load and is positionally restricted. In Dhuwaya, the contrast is limited to intervocalic position [1, 2]. The loss of the contrast for the other four places in Dhuwaya is attributed to regular processes of lenition of lenis stops to approximants /j/ and /w/. When the contrast does occur, it is thought that closure duration, not voicing or aspiration, is the primary cue [3].

Across the Yolngu languages, the same orthography is used, and is, on the whole, fairly transparent with a one-to-one correspondence between phonemes and graphemes [1]. In Dhuwaya, orthographically <b, dh, d, d̲, dj, g> are used for word-initial and after-nasal stops as well as stops that would be lenis in varieties where there is a contrast, while <p, th, t, t̲, tj, k> are used for word-final and stop-neighbouring stops, as well as stops that would be fortis in varieties where there is a contrast. Therefore, the relationship between stop graphemes and phonemes is obscured, so students must have knowledge of spelling rules for writing stops in most contexts [1]. In this paper we investigate the acoustic phonetic nature of stops in Dhuwaya across positions, focusing on the duration of stops as well as the contribution to duration of individual elements of voice onset time (VOT)—voiced and voiceless closure, release/aspiration. We also offer some qualitative remarks on stop production.

Dhuwaya is a koine variety that developed from contact between a number of related Yolngu languages spoken by people in and around Yirrkala, a community in east Arnhem Land, N.T., Australia [4]. Dhuwaya is the language of schooling in Yirrkala, which has a current population of around 660 people [5]. A bilingual school was established in the 1970s, with the goal of teaching children literacy in their own language and at that time Gumatj was determined to be the language of the school, though even then, most children spoke Dhuwaya [4]. Today, there is a vibrant teaching team working with linguists to develop literacy tools to complement the existing collection of reading materials, including an app and adapting the cued articulation approach to assist students in gaining phonological awareness [1]. Many non-Yolngu teachers in Yirrkala are first-language Australian English speakers. Therefore, there is a need to document the acoustics of stops to assist in students gaining phonological awareness and teachers in understanding the acoustics of stops because 1) the orthography reflects contrasts that are not phonemic in the language, and 2) non-Yolngu educators associate graphemes with different (English) phonemes and acoustic realisations. The outcomes of this study can assist educators in developing their approach to early-years literacy education that reflects the Dhuwaya stop system.

II. Methods

A. *Speakers and recordings*

Three Dhuwaya education assistants (women) recorded a range of words in isolation and frame sentences for a literacy tablet application [1]. Audio were recorded in stereo mp3 format and converted to mono wav format (128kb/s, 44.1 kHz). This is very low audio quality compared to most phonetic studies, but should not affect the temporal properties we are interested in here. Words uttered in isolation (n = 259) are examined here, which amounts to 391 stops. Isolation forms were selected for this preliminary analysis for ease of data processing. However, analysis of data from three framing sentences is forthcoming.

B. *Data processing and analysis*

Data were force aligned using Web MAUS [6] using the Australian language setting, and segmentation boundaries were manually corrected. An additional VOT tier was created and VOT element interval boundaries were placed following the recommendations by [7]. These included, where relevant, the voiced closure, voiceless closure, and release/aspiration, allowing us to capture the rate at which different stops are voiced, the proportion of voicing during the closure, and the duration of stop releases, including (positive) voice onset time in non-final stops; these measures can all be extracted from the annotated landmarks. Since we are not interested in how place of articulation or following vowels affect VOT, we take the onset of higher formants after the stop release as the relevant landmark to determine the end of the stop release, following [8]. An EMU-SDMS database [9] was created and queried in R [10]. A qualitative approach is taken to analysis due to small dataset and speaker number.

III. Results

Figure 1 Left shows the overall duration of stops in different categories, encompassing the entire (audible portion of) the closure and release phases. This shows that stops essentially cluster into two categories, corresponding to the difference between graphemes, on the basis of overall duration. Figure 1 Right shows the duration of stop releases (combining the ‘release’ and ‘aspiration’ phases). The across-group differences in release duration are fairly minor, especially considering that VOT differences below 10 ms likely fall below the just noticeable difference threshold [11]. Two groups stand out: final stops, which can vary quite freely in their release durations (when releases are audible), and stops neighbouring other stops, which are often not released at all. Apart from these groups, and disregarding outliers, no clear categories emerge from the release duration data. This suggests that stops form two clusters on the basis of overall stop duration, and in particular audible closure duration, but not on the basis of release duration or positive VOT.

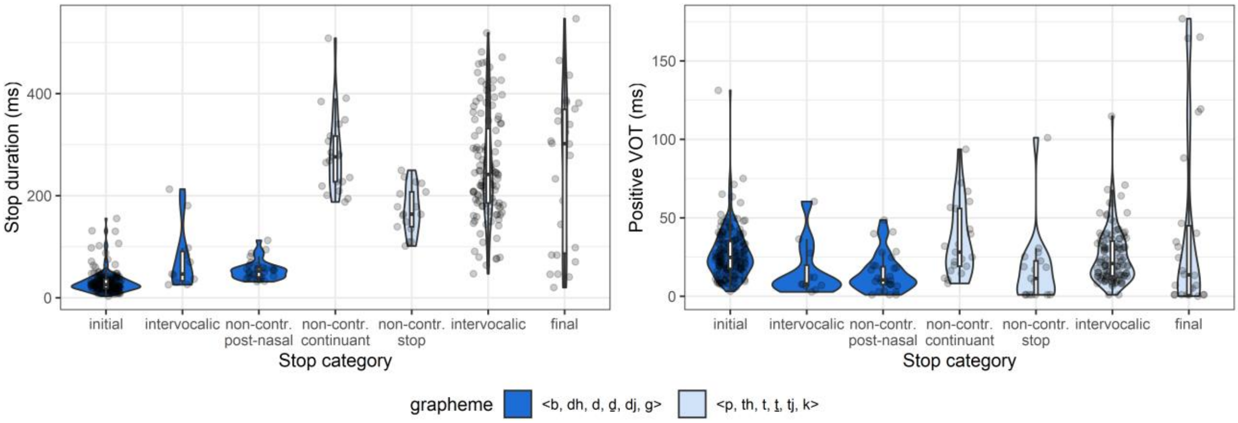


Fig. 1. Left: Stop durations by category. Right: Positive voice onset time by category. Coloured by grapheme class.

IV. Discussion

Duration patterns show Dhuwaya stops form two groups which reflect the orthographic representation of these speech sounds. Comparing total audible stop durations with release durations shows that the phonetic distinction between the graphemes relies primarily on audible closure duration: stops written with the <b> class of graphemes either have no audible closure (in the ‘initial’ case) or a short audible closure; stops written with the <p> class of graphemes have a long audible closure, and exhibit much more variation in closure duration. Future work may include other measures, such as Voice Termination Time [12]. The current results suggest a two-way division based primarily on duration could be used in a Dhuwaya literacy programme, irrespective of the functional load of the fortis/lenis contrast.

Acknowledgements

We thank the ARC Centre of Excellence for the Dynamics of Language (Project ID: CE140100041), Yirrkala Community School, Merrki Ganambarr, Yalmay Yunupingu, and Melanie Wilkinson. KJ was supported by the Alexander von Humboldt Foundation.

References

[1] G. Wigglesworth, M. Wilkinson, Y. Yunupingu, R. Beecham, and J. Stockley, "Interdisciplinary and intercultural development of an early literacy app in Dhuwaya," *Languages*, vol. 6, no. 2, 2021, doi: 10.3390/languages6020106.

[2] M. Wilkinson, *Djambarrpuyŋu: A Yolŋu variety of northern Australia*. Munich: LINCOM Europa, 2012.

[3] E. Round, "Segment inventories," in *The Oxford guide to Australian languages*, C. Bowern Ed. Oxford: Oxford University Press, 2023, pp. 96-105.

[4] R. Amery, "A new diglossia: Contemporary speech varieties at Yirrkala in north east Arnhem Land," Masters, ANU, Canberra, Australia, 1985.

[5] Australian Bureau of Statistics. "Yirrkala: 2021 Census All persons QuickStats." https://www.abs.gov.au/census/find-census-data/quickstats/2021/SAL70299.

[6] T. Kisler, U. D. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326-347, 2017, doi: 10.1016/j.csl.2017.01.005.

[7] A. S. Abramson and D. H. Whalen, "Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions," *Journal of Phonetics*, vol. 63, pp. 75-86, 2017, doi: 10.1016/j.wocn.2017.05.002.

[8] E. Fischer-Jørgensen and B. Hutters, "Aspirated stop consonants before low vowels. A problem of delimitation," *Annual Report of the Institute of Phonetics, University of Copenhagen* vol. 15, pp. 77-102, 1981, doi: 10.7146/aripuc.v15i.131752.

[9] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392-410, 2017, doi: 10.1016/j.csl.2017.01.002.

[10] *R: A language and environment for statistical computing*. (2023). R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[11] L. L. Elliott, L. A. Busse, R. Partridge, J. Rupert, and R. DeGraaff, "Adult and child discrimination of CV syllables differing in voicing onset time," *Child Development*, vol. 57, no. 3, pp. 628-635, 1986, doi: 10.2307/1130341.

[12] R. Mailhammer, S. Sherwood, and H. Stoakes, "The inconspicuous substratum: Indigenous Australian languages and the phonetics of stop contrasts in English on Crocker Island," *English World-Wide*, vol. 41, no. 2, pp. 162-192, 2020, doi: 10.1075/eww.00045.mai.

Vowel Space Areas in Varieties of German

Caroline Kleen, Marina Frank, Alfred Lameli
Research Center Deutscher Sprachatlas, Philipps-Universität Marburg, Germany

Keywords — vowel spaces, vowel space area, regional variation, German

I. Introduction

In this paper we present methods for the analysis of vowel space areas (VSA) in varieties of German (in Germany). VSAs are defined as two-dimensional areas where vowels are represented by the coordinates of their first and second formant frequencies (F1, F2). These values are determined by the size and shape of the vocal tract, i.e., the tongue position and the degree of jaw opening. They provide information on vowel position and distribution and illustrate the configurations of different vowel systems. Knowledge of the regional characteristics of VSA can be important for language teaching, but also for the assessment of language disorders.

In German dialectology, a distinction is made between vertical and horizontal variation [1]. The latter describes variation in terms of geographical space, such as in Low German and Bavarian dialects. Vertical variation examines the range of varieties between dialect and Standard German (registers). This type of variation can be influenced by regional factors or by an individual's linguistic proficiency and may additionally reflect social identity, age, and the level of formality in a communicative context [2]. Previous studies have explored variation in VSAs of different languages. For instance, studies on American English have revealed distinct VSAs corresponding to dialect and gender [3]. Regional differences have also been studied for Low German varieties [4], as well as between various locations in Bavaria [5]. In [6], differences in vertical variation have been observed for speakers from Leipzig, depending on whether individuals employ standard-intended or dialect-intended speech. Conversely, another study with crowdsourced data from speakers from Germany found no clear effect on VSA variation [7]. Given the high degree of variability in varieties of German, we assess systematic cross-dialectal comparisons of VSAs. We expect that formant frequency values will vary among regions (horizontal variation) and among registers (vertical variation), resulting in different shapes and area measurements of the VSAs.

II. Methods

A. Data

For the analysis, we use data from the corpus "Regionalsprache.de" (REDE) [8]. This corpus contains audio recordings of 148 locations in Germany. We follow the selection of [9] of one location in each of seven dialect regions, which are Northern Low German, East Franconian, Upper Saxon, Mecklenburgish-West Pomeranian, Middle Bavarian, High Alemannic, and Moselle-Franconian. For each region, two speakers are examined (middle-aged: 45–55 years; older: 65+ years). Differences regarding gender cannot be analyzed because the corpus only contains male speakers. The recordings include standard-intended speech, elicited by a translation task of the "Wenker sentences" (dialect translated into Standard German), and read speech, elicited by reading the fable "The Northwind and the Sun" aloud. The "Wenker sentences" are 40 sentences which were proposed by Georg Wenker in the late 19th century with the aim to cover as many phonetic phenomena as possible in the dialects of German. A questionnaire with the sentences was then sent to all schools and the results were mapped in the "Sprachatlas des Deutschen Reichs". Since the days of Wenker, the sentences are often used in German dialectology because they provide ample material for phonetic analysis and comparability between different studies. For the analysis of vertical variation, we include dialect-intended speech, also elicited by a translation task of the "Wenker sentences" (Standard German translated into dialect).

B. Methods

The methodology occurs in four steps. First, orthographic transcriptions of the standard-intended "Wenker sentences" and the read-out text are needed. For the dialect-intended speech, the orthographic transcriptions are based on the standard-language equivalents of the produced dialect words. In a second step, these orthographic transcriptions can be uploaded together with the recording to the web service WebMAUS [10]. WebMAUS carries out an automated grapheme-to-phoneme conversion. The resulting TextGrid includes the segmentation and labels of words and phones. The segmentation is then manually corrected. In the third step, we extract F1 and F2 values at 21 measurement points per vowel automatically using a Praat [11] script. This allows us to analyze F1/F2 values at the vowel midpoint of all vowels. For future research, we plan to include the formant trajectories as well. In this study we analyze long and short monophthongs. Vowels followed by a vocalized /r/ were treated as diphthongs and are thus not included. The fourth step includes data normalization, visualization, and analysis using R [12]. The F1/F2 values of all vowels are Lobanov-normalized. To calculate the size of the VSAs, we implemented a polygon-formula in R.

We determine all VSAs regarding the region (n=7), the age group (n=2), the communicative setting (n=3), and vowel length (n=2). In total, we will have 84 VSA values, which we can analyze with a mixed effects regression model (lmer). This provides an objective statistical measurement that allows predictions about the usage of the vocal tract depending on the above-mentioned factors. A greater VSA value refers to greater use of the vocal tract. With this data, we can compare the values among different regions, the variation within one region, and the variation within each speaker. Additionally, we calculate a centroid of each polygon. Starting from this point, we calculate triangles formed by two vowels and the centroid, e.g., /i/ and /u/. This allows the calculation of the distance between vowels. Furthermore, the different sizes of the triangles can be compared regarding inter- and intra-speaker variability.

III. Results

First descriptive results can be found in the following figures. In Fig. 1 all long monophthongs of the standard-intended and read speech of 14 individuals (two for each location) are plotted. Fig. 2 shows the short monophthongs.

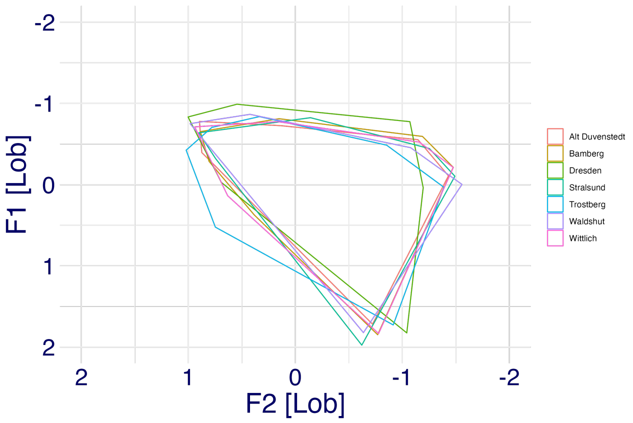


Fig. 1. Vowel space of long monophthongs per region; each polygon contains the middle-aged and older speaker of the standard-intended speech and read speech

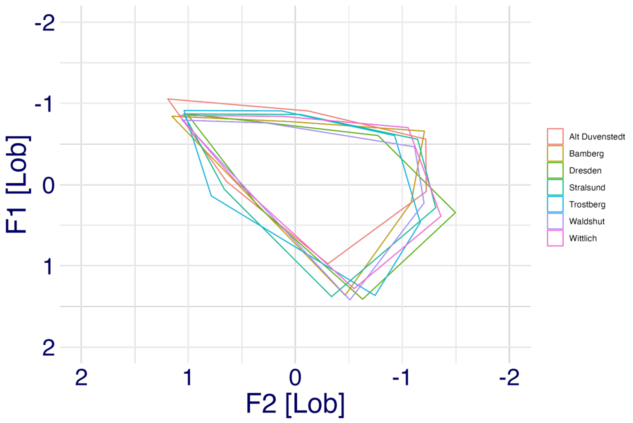


Fig. 2. Vowel space of short monophthongs per region; each polygon contains the middle-aged and older speaker of the standard-intended speech and read speech

When examining the long vowels (Fig. 1), the VSA of the speakers from Trostberg (Middle Bavarian) stands out. Their front vowels /i:/ and /e:/ are produced more towards the front of the vowel space, compared to speakers from other regions. The short vowels (Fig. 2) appear to be more centralized than the long monophthongs. In contrast, the long monophthongs are positioned closer to the periphery. For instance, the short /a/ (Fig. 2) of the speaker from Alt Duvenstedt (Northern Low German) shows more centralization compared to other regions. Additionally, the short /o/ of the speakers from Dresden (Upper Saxon) demonstrates the most backward shifted vowel, especially compared to their short /u/ which is centralized. Preliminary results of calculating the polygon size of the VSA for long vowels (Fig. 1) show variances: Trostberg has the largest VSA (3.68), whereas the VSA of the speakers from Alt Duvenstedt implicates the smallest VSA (3.03).

IV. Conclusion

This study attempts to explore methodologies for analyzing VSA variation in German, with a specific emphasis on horizontal and vertical variation. Initial observations are made from a preliminary dataset of seven locations (including standard-intended speech and read speech), each represented by two speakers. As in previous studies, variations in VSA are identified. Using a polygon formula, VSAs are quantified and facilitate statistical comparisons between these areas. By adding dialect-intended speech to our dataset, we will get insights on vertical variation. Further methodological refinements will expand analytical capabilities and enhance the depth of statistical analysis, so an objective metric for comparable analysis will be provided.

References

[1] W. König, *Dtv-Atlas zur deutschen Sprache,* 17th ed. (dtv 3025). München: Deutscher Taschenbuch-Verlag, 2011.

[2] R. Kehrein, "Areale Variation im Deutschen „Vertikal“," in *Deutsch: Sprache und Raum - Ein internationales Handbuch der Sprachvariation* (Handbücher zur Sprach- und Kommunikationswissenschaft Band 30,4), J. Herrgen and J. E. Schmidt, Eds., Berlin, Boston: De Gruyter Mouton, 2019, pp. 121–158.

[3] E. Jacewicz, R. A. Fox, and J. Salmons, "Vowel space areas across dialects and gender," presented at 16th Int. Congr. of Phonetic Sciences. Available: https://api.semanticscholar.org/CorpusID:113403028

[4] M. Gackstatter and O. Niebuhr, "Eine kontrastive phonetische Analyse niederdeutscher Langvokale," in *Linguistik Online*, vol. 53, no. 3, 2012.

[5] S. Kleiner, "F1/F2-Diagramme als Darstellungsmittel bairisch geprägter standardsprachlicher Vokalsysteme," in *Bayerisch-österreichische Varietäten zu Beginn des 21. Jahrhunderts – Dynamik, Struktur, Funktion: 12. Bayerisch-Österreichische Dialektologentagung* (Germanistik Band 167), A. N. Lenz, L. M. Breuer, T. Kallenborn, P. Ernst, M. M. Glauninger, and F. Patocka, Eds., Stuttgart: Franz Steiner Verlag, 2017, pp. 263–284.

[6] B. Siebenhaar, "Instrumentalphonetische Analysen zur Ausgestaltung des Sprechlagenspektrums in Leipzig," in *Zeitschrift für Dialektologie und Linguistik*, pp. 151–190, 2014.

[7] S. Jannedy and M. Weirich, "*Plapper - a smartphone app for selfrecording: Exploration of vowel spaces*," presented at 5th Phonetics and Phonology in Europe conf. (PaPE) 2023.

[8] J. E. Schmidt, J. Herrgen, R. Kehrein, and A. Lameli, Ed. *Regionalsprache.de: Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Forschungszentrum Deutscher Sprachatlas Marburg, 2020ff.

[9] R. Kehrein. (2012). *Regionalsprachliche Spektren im Raum. Zur linguistischen Struktur der Vertikale*. BiblioScout.

[10] F. Schiel. Automatic Phonetic Transcription of Non-Prompted Speech. presented at Int. Congr. of Phonetic Sciences, 1999, Available: https://www.phonetik.uni-muenchen.de/forschung/publikationen/ICPhS99_Schiel.pdf

[11] P. Boersma and D. Weenink, *Praat: doing phonetics by computer* [Computer program]. Version 6.1.38, retrieved 2 January 2021, Available: http://www.praat.org/

[12] R Core Team (2021). *R: A language and environment for statistical computing* [Computer program]. R Foundation for Statistical Computing. [Online]. Available: https://www.r-project.org/

The learnability of segmentals and suprasegmentals – evidence from Bulgarian learners of Modern Greek

Milena Milenova [a,b]
[a] *Sofia University “St. Kliment Ohridski”, Bulgaria*
[b] *Aristotle University of Thessaloniki, Greece*

Keywords — Bulgarian L2 Greek, new L2 segments, L2 suprasegmentals

I. Introduction

This paper explores two salient features of the interlanguage phonology of Bulgarian speakers of Modern Greek: the stopping of the interdentals /θ, ð/ and the realisation of /o/ as [u] in unstressed syllables. The interdentals /θ, ð/ are absent from the phonemic inventory of Contemporary Standard Bulgarian (CSB) and their accurate production entails the creation of a new phonetic category [1], [2]. In CSB unstressed /o/ undergoes phonological vowel reduction: its vowel quality is neutralized and it merges with /u/ [3]. In Standard Modern Greek (SMG) unstressed /o/ undergoes phonetic reduction, namely displacement in the vowel space with retention of vowel quality [4]. The accurate production of /o/ in unstressed syllables requires the implementation of a new stressed – unstressed distinction which retains the quality of the target vowel. The aim of the present study is to explore the relative degree of learning difficulty related to the accurate production of the target sounds. To compare their relative learnability a pre-test post-test experiment was conducted.

II. Method

A. Participants

The participants in the experimental group were 10 beginner Bulgarian learners of Modern Greek (8 females and 2 males, M_{age}=19.6). Their productions of the target segments were recorded two times: prior to pronunciation instruction (Time 1 / T1) and after 15 pronunciation-training sessions (Time 2 / T2). The productions of a control group of 12 native Modern Greek speakers (8 females and 4 males, M_{age}= 25.3) were recorded to provide baseline data.

B. Stimuli

The production data were collected by means of a reading task. The stimuli for the interdentals comprised real words containing the target segments in initial and medial position in stressed and unstressed syllables with all five Modern Greek vowels /i e a o u/.

For the vowel, the stimuli were the symmetrical disyllables [ˈpopo] and [poˈpo].

The elicitation words were embedded in a carrier phrase. The data were hand-annotated and analysed using the Praat software [5].

C. Measurements

The phonetic learning was measured by examination of the T1 and T2 realisations of the target sounds. The assessment was based on auditory cues, spectral evidence from the respective waveforms and wide-band spectrograms, as well as measurements of acoustic parametres. For the interdentals the first spectral moment (M1) was measured. For the vowel, measurements of duration, F1 and F2 were taken. The productions of the learners were compared to the productions of the native speakers to evaluate approximation to the target norms. To evaluate the statistical significance of the results separate mixed-design ANOVAs were carried out for each acoustic parametre.

III. Results and discussion

A. Results

The results revealed that at T2 the production of the target fricatives was significantly improved, whereas no improvement was observed for the target vowel. Specifically, the phonetic learning of /θ, ð/ was demonstrated by a decrease in the stopping of the target fricatives and an increase in the interdental fricative realisations. Moreover, improvement in the spectral properties was registered with approximation of the respective M1 values of the control group. As for the target vowel, both T1 and T2 productions of unstressed /o/ differed significantly from the L1 norm in the temporal, F1 and F2 dimensions.

B. Discussion

The present findings suggest that the creation of a new L2 phonetic category is easier than the unlearning of an L1 phonological rule. Considering that the unlearning of categorical vowel reduction entails the application of a novel way for the signaling of the stressed – unstressed distinction, the results imply differential learnability of L2 segmentals and L2 suprasegmentals. These results are in line with the existing evidence suggesting that the acquisition of L2 suprasegmental features requires more time than the attainment of L2 segmentals [6].

References

[1] Flege, J. E. (2002). Interactions between the native and second-language phonetic systems. In P. Burmeister, Th. Piske & A. Rohde (Eds.), *An integrated view of language development: Papers in honour of Henning Wode* (pp. 217–244). Trier: Wissenschaftlicher.

[2] Flege, J. E. & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge: Cambridge University Press.

[3] Andreeva, B., Barry, W. and Koreman, J. (2013). The Bulgarian stressed and unstressed vowel system. A corpus study. *14th annual conference of the international speech communication association, Interspeech 2013*, August 26-29, Lyon, France, 345-348.

[4] Arvaniti, A. (2007). Greek phonetics: the state of the art. *Journal of Greek Linguistics* 8, 97-208.

[5] Boersma, D., & Weenink, P. (2024). Praat: Doing phonetics by computer. Version 6.4.107. Retrieved from http://www.praat.org

[6] Saito, K. (2018). Advanced segmental and suprsegmental acquisition. In P. Malovrh & A. Benati (Eds.). *The handbook of advanced proficiency in second language acquisition* (pp. 282-303). Wiley-Blackwell.

Acquiring tongue shape complexity in consonants with multiple articulations

Claire Nance and Sam Kirkham; Lancaster University

Keywords — ultrasound tongue imaging, tongue complexity, speech articulation, palatalisation, child speech

I. Introduction

This presentation reports the findings of a pilot study comparing the complexity of midsagittal tongue shapes used by adult and child speakers. When children learn to speak, they learn to coordinate different speech articulators precisely in time. For sounds which require multiple lingual gestures, children need time to acquire 'lingual differentiation' i.e. control and coordination of multiple parts of the tongue [1], [2]. Previous articulatory work has shown that children use less complex tongue shapes than adults, indicating lesser mastery of lingual differentiation [3], though there are very few studies which directly compare complexity in adult and child productions [4]. The aim of this study is to investigate acquisition of lingual differentiation in consonants with phonemic secondary articulations. These consonants require precise coordination of multiple tongue gestures so are predicted to become adult-like later in child speech development, but have not yet been considered in studies analysing child tongue shape complexity [3], [4], [5], [6], [7].

The study investigates tongue shape complexity in Scottish Gaelic, a Celtic language spoken by ~70,000 people in Scotland [8]. Gaelic is an endangered language and family transmission is limited [9]. In the Outer Hebrides, where the data for this study were collected, children are now automatically enrolled in Gaelic Medium Education (GME), a form of immersion schooling. This means that although the majority of children in this part of Scotland acquire Gaelic, though often do so relying on the school system for language transmission and mainly use English at home and in social contexts [10]. This study compares palatalised and velarised laterals and nasals i.e., /l̪ʲ l̪ˠ n̪ʲ n̪ˠ/. Here, we investigate the following questions: 1) Do adults and children differ in tongue shape complexity? 2) Do palatalised and velarised consonants differ in tongue complexity? 3) Do consonants of different place and manner vary in tongue shape complexity? 4) Does tongue shape vary according to home language background and gender?

II. Methods

Synchronised audio and ultrasound data were collected from two groups of speakers on the Isle of Lewis, north-west Scotland: 1) eight adults who use Gaelic in professional settings, and 2) sixteen children acquiring Gaelic at primary school. The adults were aged 21-60 (3f, 4m) and represent a community target for Gaelic acquisition. The children were aged 4-11 years (5f, 11m). Two children came from a Gaelic-only home, six from an English-only home, and eight from a bilingual home.

All data were recorded in Articulate Assistant Advanced [11] using a Telemed Micrus ultrasound machine at ~90Hz frame rate and a stabilisation headset [12]. Audio data were recorded with a headset microphone connected to an audio interface. The data reported here consider word-initial palatalised and velarised laterals and nasals presented to participants as individual words in AAA. The word list consisted of eight words (2 per phoneme) and was repeated 2-3 times per participant (total 445 tokens) alongside other distractors and before a short English word list. Splines were fitted to the ultrasound data in AAA using the DeepLabCut plugin [13] and tongue coordinates were extracted rotated to each speaker's occlusal plane. In this initial pilot study, we consider data from the sonorant acoustic midpoint (labelled in Praat [14]). Future analyses will consider dynamic tongue shape across the word-initial sonorant and following vowel. Data were exported from AAA and further analysis carried out in Python and R. We first used the Python script from [5] to calculate Maximum Curvature Index (MCI) at sonorant midpoint. This method calculates a number representing how curled up or stretched out the tongue shape is. We also calculated Number of Inflection Points (NINFL) [7] in AAA. NINFL measures the number of times a tongue spline changes from convex to concave.

A linear mixed effects model was fitted to the MCI data, and an ordinal mixed effects model to the NINFL data. Models contained participant age (child/adult), gender (male/female), and an interaction of consonant and secondary articulation. Word and speaker were included as random intercepts. Significance testing was carried out by model comparison. Due to the small numbers of children at each age, age differences within the child sample were explored qualitatively, as well as home language differences.

III. Results

MCI values: Children had higher significantly higher Modified Curvature Index values than adults, indicating a more curled up tongue shape (Fig1.a). In nasal consonants, palatalised consonants had higher MCI values than velarised. There were no significant differences between palatalised and velarised laterals. Among the children, younger children have higher MCI values, with children approaching adult norms by age 8;4 (Fig1.b). There are some differences according to home language background: children from Gaelic-only homes have lower (more adult-like) MCI values (Fig 1.c). However, there are only two children (aged 6 and 8 years) in this sample currently. There were no significant differences for speaker gender. **NINFL values:** This analysis was the same results as the MCI analysis, except the other way round i.e. children have lower NINFL values than adults etc. These results are not detailed fully here due to space constraints.

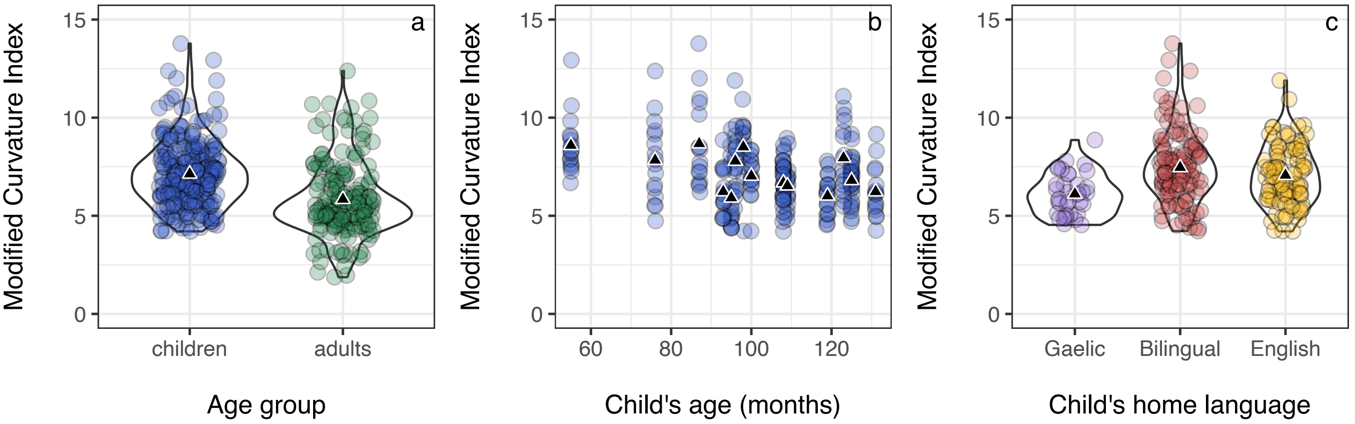


Fig. 1. Panel a: MCI values for children and adults. Panel b: MCI values according to age of chidlren in months. Panel c: MCI values comparing home langauge background among the children. Group means (panels a, c) or speaker means (panel b) are shown as black triangles.

IV. Discussion and Conclusions

The results show children have significantly higher MCI values than adults. This indicates a more curled up tongue shape and lesser mastery of lingual differentiation. These results echo previous results from clinically-focussed studies using MCI [3], [6], and support the idea that children are more likely to move the tongue a more closely linked unit than adults [15]. Qualitative analysis of the children's data indicates that MCI values are higher in the youngest children, and adult-like in children over ~8 years. This supports auditory research funding protracted acquisition of consonants requiring multiple gestures [16]. NINFL results obtained here also indicate lesser lingual differentiation in children (lower NINFL scores). MCI and NINFL were able to distinguish between palatalised and velarised nasals. The higher values in palatalised nasals are likely due to the advanced tongue root gesture in palatalisation [17]. There is some indication of home language differences among the child sample, where children from Gaelic-only homes have more adult-like MCI values. However, we currently only have data from two children who come from a Gaelic-only home so more data are needed to confirm this result. Future work will extend the sample size, as well as considering tongue shape across the consonant and surrounding vowel in a wider range of words and consonants.

References

[1] F. E. Gibbon, 'Undifferentiated Lingual Gestures in Children With Articulation/Phonological Disorders', *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 2, pp. 382–397, 1999, doi: 10.1044/jslhr.4202.382.

[2] D. Abakarova, S. Fuchs, and A. Noiray, 'Developmental Changes in Coarticulation Degree Relate to Differences in Articulatory Patterns: An Empirically Grounded Modeling Approach', *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 9, pp. 3276–3299, 2022, doi: 10.1044/2022_JSLHR-21-00212.

[3] M. Dokovova, E. Sugden, G. Cartney, S. Schaeffler, and J. Cleland, 'Tongue Shape Complexity in Children With and Without Speech Sound Disorders', *Journal of Speech, Language, and Hearing Research*, vol. 66, no. 7, pp. 2164–2183, 2023, doi: 10.1044/2023_JSLHR-22-00472.

[4] H. Kabakoff, S. P. Beames, M. Tiede, D. H. Whalen, J. L. Preston, and T. McAllister, 'Comparing metrics for quantification of children's tongue shape complexity using ultrasound imaging', *Clinical Linguistics & Phonetics*, vol. 37, no. 2, pp. 169–195, 2023, doi: 10.1080/02699206.2022.2039300.

[5] K. M. Dawson, M. K. Tiede, and D. H. Whalen, 'Methods for quantifying tongue shape and complexity using ultrasound imaging', *Clinical Linguistics & Phonetics*, vol. 30, no. 3–5, pp. 328–344, May 2016, doi: 10.3109/02699206.2015.1099164.

[6] H. Kabakoff, D. Harel, M. Tiede, D. H. Whalen, and T. McAllister, 'Extending Ultrasound Tongue Shape Complexity Measures to Speech Development and Disorders', *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 7, pp. 2557–2574, 2021, doi: 10.1044/2021_JSLHR-20-00537.

[7] J. L. Preston, P. McCabe, M. Tiede, and D. H. Whalen, 'Tongue shapes for rhotics in school-age children with and without residual speech errors', *Clinical Linguistics & Phonetics*, vol. 33, no. 4, pp. 334–348, Apr. 2019, doi: 10.1080/02699206.2018.1517190.

[8] National Records of Scotland, 'Scotland's Census 2022'. 2022. Accessed: Jun. 07, 2024. [Online]. Available: https://www.scotlandscensus.gov.uk

[9] C. Moseley, *Atlas of the world's languages in danger*, 2nd ed. Paris: UNESCO, 2010. [Online]. Available: http://www.unesco.org/languages-atlas/

[10] C. Nance, 'Bilingual language exposure and the peer group: Acquiring phonetics and phonology in Gaelic Medium Education', *International Journal of Bilingualism*, vol. 24, no. 2, pp. 360–375, 2020.

[11] A. Wrench, *Articulate Assistant Advanced (Version 221.2)*. Edinburgh: Articulate Instruments, 2023.

[12] L. Spreafico, M. Pucher, and A. Matosova, 'UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science', in *Interspeech 2018*, 2018. doi: 10.21437/interspeech.2018-995.

[13] A. Wrench and J. Balch-Tomes, 'Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut', *Sensors*, vol. 22, no. 3, p. 1133, 2022, doi: 10.3390/s22031133.

[14] P. Boersma and D. Weenik, 'Praat: doing phonetics by computer [Computer program]. Version 6.4.01'. Accessed: Nov. 30, 2023. [Online]. Available: http://www.praat.org/

[15] E. Rubertus and A. Noiray, 'Children's coarticulation patterns as a window to the phonology-phonetics interface', in *Proceedings of the International Seminar on Speech Production*, C. Fougeron and P. Perrier, Eds., Autrans, France, 2024.

[16] S. McLeod and K. Crowe, 'Children's Consonant Acquisition in 27 Languages: A Cross-Linguistic Review', *American Journal of Speech-Language Pathology*, vol. 27, no. 4, pp. 1546–1571, 2018, doi: 10.1044/2018_AJSLP-17-0100.

[17] R. Bennett, M. Ní Chiosáin, J. Padgett, and G. McGuire, 'An ultrasound study of Connemara Irish palatalization and velarization', *Journal of the International Phonetic Association*, pp. 1–44, 2018.

Prosody Manipulation Training: Can Both Intermediate and Advanced Learners Benefit?

Radek Skarnitzl [a], Tomáš Bořil[a] [1]
[a] *Institute of Phonetics, Faculty of Arts, Charles University, Czech Republic*

Keywords — phrasing, melody, rhythm, prosody manipulation, learner proficiency

I. Introduction

Prosodic aspects of speech have been shown to be essential for successful communication and for the comprehensibility of L2 speakers (e.g., [1]). At the same time, studies examining the effects of prosodic instruction (e.g., [2]) have demonstrated that prosodic aspects are teachable and learnable. However, the cleft between the results of empirical research and their implementation into teaching practice has been lamented for a long time (see [3] for a recent example). Prosody is targeted only rarely in English lessons in the Czech context, despite the considerable differences in the prosodic makeup of English and Czech [4], [5], [6].

The current study is conceptually based on [7] who used short stretches of L1 French university teachers' L2 English speech with modified prosodic patterns as feedback in a training session and demonstrated the usefulness of this approach. In contrast with [7], this study focuses on Czech speakers of English, and the instruction targeted predominantly phrase-level prosody. One of the main aims of the present research is to reveal whether both intermediate and advanced learners of English may benefit from a relatively short prosody-focused training where their own modified speech served as a model. On the one hand, research shows that prosody training can be successful even with learners who have been using the target language for a long time and seem to be fossilized [8]. On the other hand, however, highly advanced learners have rarely been targeted and, crucially, learners at this stage are likely to have progressed from explicit to implicit acquisition processes [9], possibly weakening effects of explicit instruction.

II. Method

Six intermediate (B1/B2 as per the CEFR) and six advanced (C1/C2) speakers participated in the individualized prosody-focused training. A baseline recording (a diagnostic reading text) was acquired during the first session. Portions of the recording were modified using PSOLA [10] in Praat [11], based on the judgement of the first author, who has been teaching the phonetics of English for over twenty years. Nine or ten such stimuli were used for each participant, with one stimulus comprising between one and three prosodic phrases. The features which were targeted during the manipulations include the following:

- phrasing (prosodic phrases are longer in Czech than English) and the corresponding phrase boundary signals (i.e., deceleration and marked nuclear pitch movement when a phrase boundary was added into the participants' speech)
- pitch range (Czech is very flat as compared with English)
- temporal characteristics of syllables (specifically, shortening of unstressed syllables)
- prominence (when it was suitable to shift the nuclear-accented word)

During the first training session (ca. 75 minutes), the learners first received information about the main prosodic aspects of English, their importance in communication, as well as the reasoning behind using their own voice for individualized feedback. The prosodic phrase was emphasized as the central unit of connected speech – not only in terms of intonation [12], but also rhythm [13]. The participants then listened repeatedly to the modified stimuli and, after a 20-second silent period (cf. [7]) repeated them. The second training session (ten days later, on average; ca. 50 minutes) was based on the same stimuli, with two additions. First, visual displays of the stimuli were provided (pitch contours and temporal information corresponding to portions which were lengthened and shortened, similarly to what appears in Praat's *Manipulation* window). Second, the respondents were encouraged to use any form of gesticulation to reinforce the prosodic patterns. A delayed post-test recording was obtained at least six weeks after the second training session. During this last session, the participants were asked to record the initial diagnostic text, as well as a completely new text; they were encouraged to make any notes that would help them with the reading. The analyses below are based on the comparison of the first diagnostic recording and the two post-test recordings.

Fundamental frequency was analyzed using default settings for filtered autocorrelation in Praat (note that this is the new, more robust Praat's method of extracting f_0), only with f_0 ceiling changed to 500 Hz. The Pitch object was converted to PitchTier, and portions containing creaky voice were manually removed, so as not to affect subsequent analyses. The rPraat package [14] was then used to convert f_0 to semitones (ST) re 50 Hz and to stylize the contour using either one point corresponding to mean f_0 at the mean time point for vowels shorter than 100 ms, or two points (the first and last f_0 point within the vowel) fitted with linear regression for longer vowels (see [15]). The stylized f_0 contours were used to compute 1) f_0 range (note that the stylization avoids reporting outlier values); and 2) cumulative slope index (CSI [16]) to capture overall melodic variability, always for the entire phrase and separately for its nuclear and prenuclear portion. In addition to f_0 properties, the suitability of prosodic boundary placement was determined.

III. Results and Discussion

The results, at this point for five intermediate and five advanced speakers since analyses are ongoing, indicate considerable improvements in the prosodic patterning between the two timepoints (i.e., before and six weeks after the training). Only f_0 range is reported here, but the CSI results suggest a similar improvement. Fig. 1 shows that f_0 range, extracted from the stylized contours, is markedly larger in the post(-training) conditions than it was in the pre-training condition, and more so in the nuclear portions of prosodic phrases (i.e., where the nuclear pitch pattern is realized). Importantly, the improvement is similar in both the intermediate and advanced speakers. In addition, there appears to be an effect of text, with the text which was familiar to the participants (having been asked to read it also during each training session; condition post1) manifesting a lower f_0 range than the new text (condition post2). On the other hand, phrase score (which reflects a compound index of the suitability of prosodic boundary placement) turns out to be better in most speakers between the pre1 and post1 conditions, but markedly worse in the post2 condition.

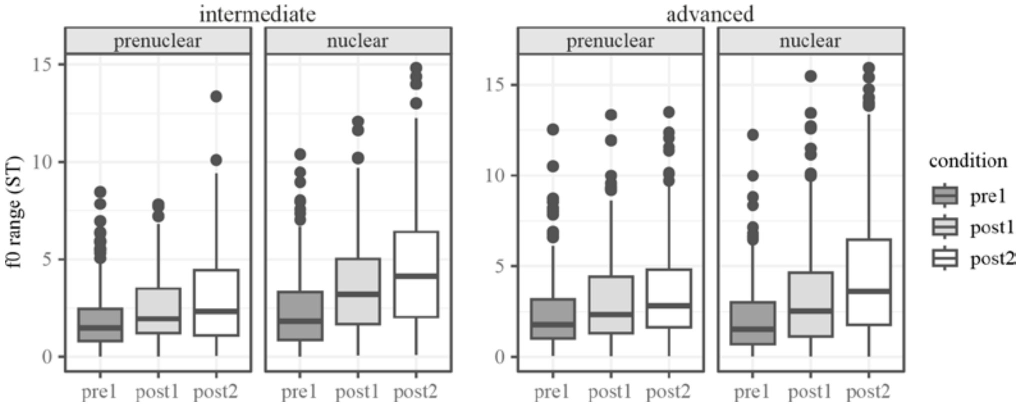


Fig. 1. Boxplots of f_0 range (ST) in prenuclear and nuclear portions of prosodic phrases before the training (pre1) and after the training in the same text (post1) and in a new text (post2) in intermediate and advanced speakers.

The results of our study indicate that a short prosody-focused training, comprising an awareness-raising session and audio-only and audio-visual sessions with the speakers' voices modified using PSOLA, yields marked improvements in the melodic variability of prosodic phrases. Both intermediate and advanced speakers of the target language may benefit from such a short training. Naturally, this concerns a read text, which participants are able to prepare in advance; such a short intervention would not be expected to result in changes in ordinary speech.

References

[1] T. M. Derwing, M. Munro, and G. Wiebe, "Evidence in favor of a broad framework for pronunciation instruction," Lang. Learn., 48, pp. 393–410, 1998.

[2] T. M. Derwing, and M. Rossiter, "The effects of pronunciation instruction on the accuracy, fluency and complexity of L2 accented speech," Appl. Lang. Learn., 13, pp. 1–17, 2003.

[3] T. M. Derwing, J. Levis, S. Sonsaat-Hegelheimer, "Bridging the research-practice gap in L2 pronunciation," in Second language pronunciation: Bridging the gap between research and teaching, T. M. Derwing, J. Levis and S. Sonsaat-Hegelheimer, Eds. Oxford: Wiley Blackwell, pp. 1–18, 2022.

[4] R. Skarnitzl, and H. Hledíková, "Prosodic phrasing of good speakers in English and Czech," Frontiers in Psych., 13, paper 857647, 2022.

[5] R. Skarnitzl, and J. Rumlová, "Phonetic aspects of strongly-accented Czech speakers of English," Acta Univ. Carolinae – Philo., pp. 109–128, 2019.

[6] J. Volín, K. Poesová, and L. Weingartová, "Speech melody properties in English, Czech and Czech English: Reference and interference," Res. in Lang., 13, pp. 107–123, 2015.

[7] A. J. Henderson, and R. Skarnitzl, "'A better me': Using acoustically modified learner voices as models," Lang. Learn. Technol., 26(1), pp. 1–21, 2022.

[8] T. M. Derwing, M. J. Munro, J. A. Foote, E. Waugh, and J. Fleming, "Opening the window on comprehensible pronunciation after 19 years: A workplace training study," Lang. Learn., 64, pp. 526–548, 2014.

[9] R. DeKeyser, "Skill Acquisition Theory," in Theories in second language acquisition: An introduction, B. VanPatten and J. Williams, Eds. Lawrence Erlbaum Associates Publishers, 2007, pp. 97–113.

[10] E. Moulines, and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun., 9, pp. 453–467, 1990.

[11] P. Boersma, and D. Weenink, Praat: doing phonetics by computer, v. 6.4.06 Retrieved February 25, 2024 from http://www.praat.org/

[12] J. C. Wells, English intonation: An introduction. Cambridge University Press.

[13] W. Dickerson, "The ripples of rhythm: Implications for ESL instruction," in Proceedings of the 10th Pronunciation in Second Language Learning and Teaching conference, J. Levis, C. Nagle and E. Todey, Eds., 2019, pp. 36–54.

[14] T. Bořil, and R. Skarnitzl, "Tools rPraat and mPraat: Interfacing phonetic analyses with signal processing," in Proceedings of the 19th International Conference on Text, Speech and Dialogue, P. Sojka, A. Horák, I. Kopeček and K. Pala, Eds. Cham: Springer International Publishing, pp. 367–374, 2016.

[15] D. J. Hermes, "Stylization of pitch contours," in Methods in Empirical Prosody Research, S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzk, I. Mleinek, N. Richter, J. Schließer, Eds. Berlin: De Gruyter, pp. 29–62, 2006.

[16] R. Hruška, and T. Bořil, "Temporal variability of fundamental frequency contours," Acta Univ. Carolinae – Philo., pp. 35–44, 2017.

[1] This work was supported by the European Regional Development Fund project "Beyond Security: Role of Conflict in Resilience-Building" (reg. no.: CZ.02.01.01/00/22_008/0004595).

Non-native stress perception by Hungarian listeners: revisiting 'stress deafness’

Gabriela Tavares [a], Andrea Deme [b] and Susana Correia [a]

[a] *CLUNL - Linguistics Research Centre of NOVA University Lisbon, Portugal*

[b] *Department of Applied Linguistics and Phonetics of ELTE Eötvös Loránd University Budapest, Hungary*

Keywords — non-native speech perception, word stress, stress ‘deafness’, European Portuguese, Hungarian

Although it is traditionally assumed that speakers of languages with non-contrastive word stress display ‘deafness’ to perceive stress contrasts, this effect is not homogenously observed among those languages [1]. Furthermore, speakers of languages with lexically contrastive word stress can also display ‘deafness’ [2]. Word stress results from the effect of suprasegmental acoustic cues – pitch, duration, intensity – used to signal the relative prominence of a syllable, and it may interact with segments, as some languages display vowel reduction in unstressed position [3]. The relative weight of each acoustic cue for stress is language-specific [4]. For example, European Portuguese (EP) has lexically contrastive word stress (e.g., túnel [ˈtunɛɫ] 'tunnel' vs. tonel [tuˈnɛɫ] 'barrel'), and stress can fall on one of the last three syllables of the prosodic word [5]. Vowel reduction has been attested as the main cue to stress and, in its absence, EP listeners display stress ‘deafness’ [2], [6], [7]. At a suprasegmental level, word prominence in EP is acoustically marked by duration, but in paroxytonic words – the regular, unmarked stress pattern in EP –, no systematic acoustic cues occur [6]. Additionally, f0 does not play a role in EP word stress, operating rather at sentence level [6]. Despite the importance of word stress to intelligibility [8], [9], to date, few cross-linguistic studies investigated how language-specific stress properties relate to stress perception, and even more so in the case of non-native speech perception [4]. The present study aims at contributing to fill in this gap, by investigating how language-specific suprasegmental acoustic cues relate to stress ‘deafness’. To this purpose, we conducted a perceptual study with Hungarian listeners presented with EP stimuli. Hungarian is a syllable-timed language, with non-contrastive word stress, fixed on the first syllable [10]. Previous studies have attested that speakers of this language display robust ‘deafness’ to stress contrasts [1],[11],[12].

We recruited a group of Hungarian native speakers aged 18-45 with no previous contact with EP (n = 65). An additional sample of EP native speakers (n = 12) was used as a control group. Participants were presented with an oddity discrimination task with catch trials, in which they were asked to listen to a sequence of three stimulus, determine if they belong to the same category or, if not, identify the odd one. Previous studies conducted with Hungarian listeners used dissyllabic pseudowords and only contrasts between stimuli stressed in the 1st syllable and stimuli stressed in the 2nd syllable were investigated. Considering that in EP stress can fall on last, penultimate or antepenultimate syllable of a word, trisyllabic pseudowords were used in this experiment, following Correia et al. [2]. Stimuli complied with a CVCVCV structure and included only the [i] and [u] vowels, since these are the sole EP vowels that do not undergo vowel reduction in unstressed position. To promote categorical processing rather than phonetic discrimination, stimuli was produced by three EP native speakers, two female and one male. We tested five conditions: stress on the 1st syllable (e.g., [ˈtikuɾu]), stress on the 2nd syllable (e.g., [tiˈkuɾu]), stress on the 3rd syllable (e.g., [tikuˈɾu]), 1st vs. 2nd syllable stress contrast (e.g., [ˈtikuɾu]-[tiˈkuɾu]), and 2nd vs. 3rd syllable stress contrast (e.g., [tiˈkuɾu]-[tikuˈɾu]). Similar to previous studies [1], [2], [11], [12], we added consonantal contrasts (e.g., [tiˈbuɾi]/[tiˈʃuɾi]), as a baseline condition. Each participant completed 123 trials, 75 trials for stress contrast perception, and 48 with consonantal contrasts, presented in two separated blocks. All trials were randomized across participants. The experiment was built and hosted online in Gorilla Experimental Builder [13].

A linear mixed-effects logistic regression revealed a significant effect of *L1* (χ^2(3, N = 1) = 24.03, *p* < .001), with Hungarian participants displaying lower accuracy in discrimination than the Portuguese listeners. Additionally, we found an interaction of *L1*type of contrast* (stress vs. consonantal): χ^2(6, N = 2) = 42.689, *p* < .001. Pairwise comparisons revealed that both Hungarian and Portuguese listeners’ results for stress contrast were significantly less accurate than consonantal contrasts (*p* > .001 and *p* = .002, respectively). However, a significant difference was found between Hungarian and Portuguese results in stress contrasts (p < .001), but not in consonantal contrasts. These results confirm that EP speakers display stress ‘deafness’ in the absence of vowel reduction [2] and that Hungarian listeners display a robust stress ‘deafness’, as observed in previous studies [1], [11], [12]. Regarding perception as a function of stress position, we also find a significant *L1*stress position* interaction (χ^2(11, N = 5) = 32.72, *p* < .001; Fig. 1).

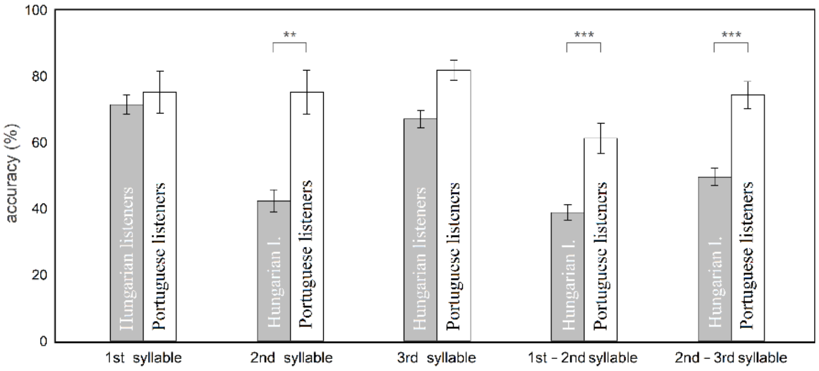
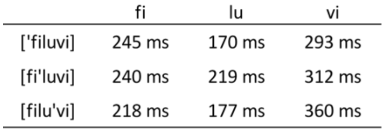


Fig. 1. Accuracy for stress perception, for Hungarian (grey) and Portuguese (white) listeners, as a function of stress position

Pairwise comparisons revealed that Hungarian participants' accuracy was significantly lower than that of Portuguese listeners in the conditions *stress on the 2nd syllable, 1st vs. 2nd syllable stress contrast*¸ and *2nd vs. 3rd syllable stress contrast*, but not in the conditions *stress on the 1st syllable* and *stress on the 3rd syllable.* These results can be reasoned in light of the EP specific stress system. As mentioned, in EP, penultimate stressed syllables are not significantly longer than unstressed syllables and, therefore, do not carry any additional acoustic cues, as exemplified in Table 1.

TABLE 1. Syllable duration for the tokens produced by one of the female EP speakers, for the pseudoword /filuvi/



	fi	lu	vi
['filuvi]	245 ms	170 ms	293 ms
[fi'luvi]	240 ms	219 ms	312 ms
[filu'vi]	218 ms	177 ms	360 ms

Duration values shown in Table 1 further explain the lower accuracy in the results for the Hungarian listeners in trials that included paroxytonic stimuli. Improved performance in trials with *stress on the 1st syllable* may be explained by the regular stress pattern in Hungarian, which is in the first syllable of a word. Also, enhanced performance of Hungarian listeners in the condition *stress on the 3rd syllable* may have resulted from the fact the last syllable of a word in citation format is usually longer, as shown in Table 1. Interestingly, vowel length is contrastive in Hungarian and, consequently, speakers of this language are sensitive to durational contrasts. Low accuracy in stress contrasts (*1st vs. 2nd syllable* or *2nd vs. 3rd syllable*) suggests that Hungarian participants overall fail to discriminate contrasting stress patterns.

The results collected in the experiment suggest that Hungarian listeners are able to perceive stress when stimuli comply with the L1 pattern (stress fixed on the 1st syllable) or are acoustically enhanced with duration (stress on the 3rd syllable). This outcome is in line with the findings from Salsignac [14]: listeners display a bottom-up strategy (attention to acoustic signals) when non-native stimuli have a salient acoustic cue to stress, otherwise, they rely on the L1 default stress position. Additionally, our results show that Hungarian speakers fail to perceive stress contrasts (*1st vs. 2nd syllable* or *2nd vs. 3rd syllable*), even when a native acoustic cue (duration) is present, which suggests a categorization problem, that is, a robust stress 'deafness'. In sum, our findings support the idea that stress 'deafness' does not occur across-the-board, but rather depends on language-specific properties, and, in non-native speech perception, on the interaction between L1 and L2 properties. Moreover, this study emphasizes the importance of further research in non-native word stress perception.

REFERENCES

[1] S. Peperkamp, I. Vendelin, E. Dupoux, "Perception of predictable stress: A cross-linguistic investigation", *Journal of Phonetics*, vol. 38, no 3, pp. 422–430, 2010.

[2] S. Correia, J. Butler, M. Vigário, S. Frota, "A stress "deafness" effect in European Portuguese", *Language and Speech*, vol. *58,* no. 1, pp. 48–67, 2015.

[3] F. Ramus, M. Nespor, J. Mehler, "Correlates of linguistic rhythm in the speech signal", *Cognition*, vol. *73,* no. 3, pp. 265–292, 1999.

[4] A. Chrabaszcz, M. Winn, C. Y. Lin, W. J. Idsardi, "Acoustic Cues to Perception of Word Stress by English, Mandarin, and Russian Speakers", *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1468–1479, 2014.

[5] I. Pereira, "Acento de Palavra" in *Gramática do Português - Volume III*, E. B. P. Raposo *et al*. Eds. Lisboa, Portugal: Fundação Calouste Gulbenkian, 2020, pp. 3397–3425.

[6] M. R. Delgado-Martins, "Sept etudes sur la perception". Lisboa, Portugal: Instituto Nacional de Investigação Científica, 1986.

[7] A. Castelo and E. N. Santos, "As vogais do português entre aprendentes chineses e suas implicações no desenvolvimento de um programa de português", in *Português como língua estrangeira, de herança e materna: abordagens, contextos e práticas*. Boavista Press: pp. 123–136, 2017.

[8] L. D. Hahn, "Primary Stress and Intelligibility: Research to Motivate the Teaching of Suprasegmentals", *Tesol Quarterly*, vol. 38, no. 2, pp. 201–223, 2004.

[9] J. Field, "Intelligibility and the Listener: The Role of Lexical Stress", *TESOL Quarterly*, vol. 39, no. 3, pp. 399–424, 2005.

[10] A. Markó, "Hangtan", in *Nyelvtan,* A. Imrényi *et al*., Eds. Budapest, Hungary: Osiris Kiadó, pp. 75–203, 2017.

[11] S. Peperkamp and E. Dupoux, "A typological study of stress 'deafness'", in *Laboratory Phonology 7,* C. Gussenhoven and N. Warner, Eds. pp. 203–240, 2002.

[12] F. Honbolygó, A. Kóbor, V. Csépe, "Cognitive components of foreign word stress processing difficulty in speakers of a native language with non-contrastive stress", *International Journal of Bilingualism*, vol. 23, no. 2, pp. 1–15, 2017.

[13] A. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, J. Evershed, "Gorilla in our Midst: An online behavioral experiment builder", *Behavior Research Methods*, pp. 1–20, 2018.

[14] J. Salsignac, "Une étude sur la perception de l'accent primaire de langues étrangères", *Letras de Hoje*, vol. 33, no. 4, pp. 81–107, 1998.

Perception of Tashlhiyt Singleton-Geminate Contrasts by Speakers from Korean, Mandarin, and Mongolian Backgrounds

Kimiko Tsukada [a, b], Jeong-Im Han [c], Yurong [d], Pierre Hallé [e], Rachid Ridouane [e]
[a] Macquarie University, Australia, [b] The University of Melbourne, Australia, [c] Konkuk University, Korea,
[d] Inner Mongolia University, China, [e] LPP (CNRS & Sorbonne Nouvelle), France

Keywords — singleton-geminate contrasts, within-word position, Tashlhiyt, cross-linguistic speech perception

I. Introduction

Tashlhiyt, a variety of Amazigh language spoken in Morocco, is unique in that it uses durational variation contrastively for all consonants in all word positions [1]. Cross-linguistic perception of its singleton-geminate contrasts is still under-studied with a few exceptions [2, 3]. In our own work, we showed that word-medial and -final (but not word-initial) contrasts were surprisingly easy for naïve non-native listeners, regardless of whether their native language (L1) has geminates or not (Japanese, French, and Taiwan Mandarin listeners).

In this study, we examined the performance of native listeners of three additional L1 backgrounds (Korean, mainland China Mandarin, and Mongolian) that lack true gemination but differ with respect to the incidence of fake gemination (as in English *one nail, unnoticed* [4]) on the same Tashlhiyt contrasts as used in our previous work. Unlike Tashlhiyt, consonant length is not contrastive at the level of words in Mandarin or Mongolian. While Mongolian uses vowel length contrastively and closed syllables frequently, Mandarin has a limited incidence of even fake gemination. Korean fortis stops have a longer closure duration than lenis or aspirated stops. But whether or not Korean fortis obstruents should be viewed as geminates remains controversial. Thus, our primary aim was to add to the current understanding of non-native perception of singleton-geminate contrasts. We hypothesize that non-initial and absolute (or utterance) initial gemination are perceived differently. In particular, due to the absence of acoustic cues to duration in absolute initial position, singleton-geminate contrasts should be generally difficult for voiceless stops [2].

II. Methods

A. Speech materials

Twelve (4 consonants x 3 positions) Tashlhiyt gemination contrasts were used (/d/: *diʁ-ddiʁ, tidi-tiddi, fad-fadd;* /t/: *tili-ttili, juti-jutti, jufat-jufatt;* /s/: *sir-ssir, tisi-tissi, ifis-ifiss;* /n/: *niʁ-nniʁ, inas-innas, imun-imunn*). This allowed for a comparison between the three within-word positions: absolute-initial, medial, and final. The stimuli were produced in isolation by a male native speaker.

B. Participants

We recruited three groups of participants: Korean (n = 20, M_{age} = 21.4), Mandarin (n = 10, M_{age} = 19.6) and Mongolian (n = 23, M_{age} = 21.8)). The Korean group participated in the study in Seoul, Korea, and the Mandarin and Mongolian groups participated in the study in Hohhot, Inner Mongolia Autonomous Region, China. None of the participants had ever been exposed to Tashlhiyt.

C. Procedure

Participants' discrimination of the singleton-geminate contrasts was tested using an AXB procedure. Following an 8-trial practice (with feedback), the participants received 192 test trials (with no feedback). They could take a break after 96 trials if they wished. The experiment was run online using PsyToolkit [5, 6].

III. Results and Discussion

The overall mean discrimination accuracy averaged across all positions was 76%, 67% and 77% for the Korean, Mandarin, and Mongolian groups, respectively (Fig. 1). The Mandarin group was less accurate than both Korean (t(18.6) = 3.2, p < .01) and Mongolian (t(18.5) = -3.5, p < .01) groups. The Korean and Mongolian groups did not differ. Table I shows discrimination accuracy for each group according to within-word position (Absolute Initial, Medial, Final) and consonant type (d n s t). Overall, accuracy was lowest in the absolute initial position for all three groups. However, their performance varied widely depending on the consonant type. In particular, in the absolute initial position, the /s-ss/ contrast was discriminated the *most* accurately by all groups, but the same contrast was discriminated the *least* accurately in other positions (shaded gray in Table I). Further, despite the expectation that the /t-tt/ contrast, without the closure duration information, may be more difficult than other contrasts in absolute initial position, it was not the hardest contrast for the Korean and Mandarin groups. For these two groups, the /d-dd/ and /n-nn/ contrasts were the hardest to discriminate in absolute initial position as shown in Table I.

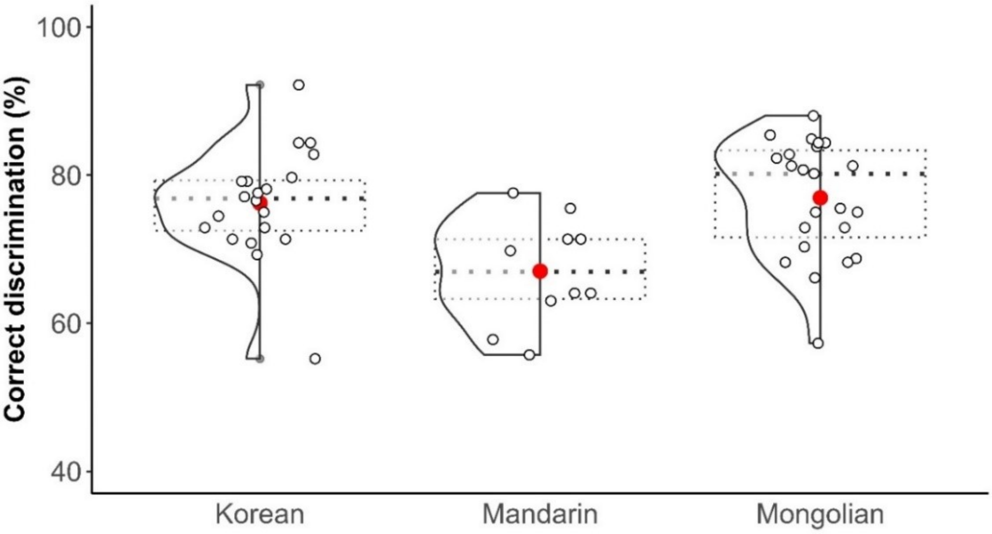


Fig. 1. Overall discrimination accuracy (%) for three groups of participants. The horizontal line and the red circle in each box indicate the median and mean, respectively. The bottom and top of the box indicate the first and third quartiles.

TABLE I. MEAN LENGTH DISCRIMINATION ACCURACY (%) BY THREE GROUPS OF PARTICIPANTS FOR TRIALS DIFFERING IN THE WITHIN-WORD POSITION (ABSOLUTE-INITIAL, MEDIAL, FINAL) AND CONSONANT TYPE (D N S T) OF THE TARGET TOKEN. STANDARD DEVIATIONS ARE IN PARENTHESES.

	Absolute-Initial				**Medial**				**Final**			
Group	d	n	s	t	d	n	s	t	d	n	s	t
Korean	48 (13)	48 (14)	75 (14)	58 (12)	90 (10)	90 (9)	84 (11)	85 (12)	92 (9)	86 (9)	80 (13)	81 (15)
Mandarin	44 (10)	44 (13)	67 (14)	51 (11)	81 (16)	74 (18)	70 (14)	70 (14)	81 (18)	81 (12)	71 (9)	71 (13)
Mongolian	52 (14)	54 (14)	70 (13)	51 (12)	91 (15)	89 (12)	84 (12)	87 (14)	93 (9)	85 (12)	83 (11)	84 (13)

A three-way Analysis of Variance with group, position (absolute-initial, medial, final) and consonant (d n s t) reached significance for the main effects of group [$F(2, 50) = 6.7, p < .01, \eta_G^2 = .09$], position [$F(2, 100) = 387.7, p < .001, \eta_G^2 = .54$] and consonant [$F(3, 150) = 7.9, p < .001, \eta_G^2 = .03$]. A two-way interaction between position and consonant also reached significance [$F(6, 300) = 34.1, p < .001, \eta_G^2 = .18$] as a result of the differing influence of consonants at each position within words.

These results are in good agreement with what was reported in our previous study, which compared the perception of native and non-native (French, Japanese, Taiwan Mandarin) groups differing in the use of singleton-geminate contrasts in their L1. Taken together, the positional asymmetry in discrimination accuracy (particularly for the /s-ss/ contrast) observed in multiple L1 groups may support the hypothesis that non-initial and absolute-initial gemination engage different perceptual processes.

REFERENCES

[1] R. Ridouane, “Tashlhiyt Berber,” J. Int. Phon. Assoc., vol. 44, pp. 207–221, 2014.

[2] P. Hallé, P. Buech, Y.-C. Chang, F.-F. Hsieh, J. Gao, and R. Ridouane, “Perception of Tashlhiyt consonant quantity contrasts by native vs. nonnative listeners from three languages,” Proc. of the 20th ICPhS, pp. 2130–2134, August 2023.

[3] P. A. Hallé, R. Ridouane, and C. T. Best, “Differential difficulties in perception of Tashlhiyt Berber consonant quantity contrasts by native Tashlhiyt listeners vs. Berber-naïve French listeners,” Frontiers in Psychology, vol. 7, Article 209, 2016.

[4] G. E. Oh and M. A. Redford. “The productionandphoneticrepresentationoffakegeminatesinEnglish,” J. Phon., vol. 40, pp. 82–91, 2012.

[5] G. Stoet, “PsyToolkit: A software package for programming psychological experiments using Linux,” Behavior Res. Methods, vol. 42, pp. 1096–1104, 2010.

[6] G. Stoet, “PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments,” Teaching of Psychology, vol. 44, pp. 24–31, 2017.

Perception of Japanese Consonant Length by Vietnamese Speakers Differing in Japanese Language Experience

Kimiko Tsukada[a, b], Đích Đào[c], Trang Le[d]

[a] *Macquarie University, Australia,* [b] *The University of Melbourne, Australia,* [c] *University of Social Sciences and Humanities, Vietnam National University - Ho Chi Minh City, Vietnam,* [d] *Waseda University, Japan*

Keywords — singleton-geminate contrasts, Japanese, Vietnamese, cross-linguistic speech perception

I. Introduction

Japanese uses durational variation contrastively for both vowels and consonants. For example, *yoka* 'leisure' contrasts with *yooka* 'eight days' on the one hand and with *yokka* 'four days' on the other hand. It is widely acknowledged that length contrast is difficult for non-native speakers from diverse L1 (first language) backgrounds including Vietnamese, which is the target language of this study. Unlike Japanese, consonant length (i.e., short/singleton vs long/geminate) is not contrastive in Vietnamese. This may pose learning difficulties for native Vietnamese speakers.

Vietnam is currently ranked within the top 6 countries/regions of the world in terms of the number of learners (174,521) of Japanese (The Japan Foundation [1]). Within Japan, Vietnam (31,643 or 14.4%) is the second largest country of origin of non-native learners of Japanese after China (67,027 or 30.5%) as of 2022 (Agency for Cultural Affairs, Government of Japan). However, empirical research focusing on the acquisition of the above-mentioned singleton-geminate contrasts by an emergent group of Vietnamese speakers is still limited. A better understanding of how native Vietnamese speakers process these Japanese speech sounds is beneficial for improving communicative efficiency. Thus, our aims were 1) to add to the prior literature examining the effects of L1 and Japanese learning by comparing native Vietnamese and native Japanese speakers, and 2) to determine how they may differ in the perception of Japanese singleton-geminate contrasts on account of their experience with Japanese.

II. Methods

A. Speakers and speech materials

Twelve Japanese word pairs (/t/: *heta-hetta, kato-katto, mate-matte, oto-otto, sate-satte, wata-watta;* /k/: *ake-akke, haka-hakka, ika-ikka, kako-kakko, saka-sakka, shike-shikke*) were audio-recorded by six (3 males, 3 females) L1 Japanese speakers and used as stimuli. The speech materials (/(C)V**C(C)**V/ tokens) were arranged in 200 unique triads. Excluding the first 8 trials for practice, the triads contained 96 singleton or 96 geminate tokens intervocalically (underlined and bolded). Only tokens with stops were considered in this study. As voiced geminates are limited in Japanese ([2, 3]), only voiceless stops (/t, k/) were used. On average, the closure durations were 96 ms and 262 ms for singletons and geminates, respectively. The geminate-to-singleton ratios were 2.7 for alveolars (/t/-/tː/) and 2.8 for velars (/k/-/kː/), respectively.

B. Participants

Participants consisted of five groups of native Vietnamese speakers and a control group of native Japanese speakers. Four groups were learners of Japanese covering a wide range of proficiency levels, i.e., N1 to N5 of the Japanese Language Proficiency Test (JLPT according to which, the easiest level is N5 and the most difficult level is N1). Two groups of learners of Japanese (N3 (n = 17, M_{age} = 21.5), N5 (n = 15, M_{age} = 19.1)) and a group of native Vietnamese speakers (n = 12, M_{age} = 21.0) without any Japanese learning experience participated in the study in Ho Chi Minh City, Vietnam. The other two groups of learners (N1 (n = 13, M_{age} = 28.6), N3 (n = 12, M_{age} = 33.5)) participated in the study in Tokyo, Japan. Both N1 and N3 groups had a mean length of residence (LOR) of 8.4 years in Japan. The N1 and N3 groups started learning Japanese at the mean age of 17.4 years (sd = 4.1) and 22.8 (sd = 6.6), respectively. A control group of native Japanese speakers (n = 10, M_{age} = 21.0) participated in the study in Eugene, Oregon. All native Japanese speakers were born and spent the majority of their life in Japan. Their mean LOR in the US was 0.4 years (sd = 0.22) at the time of participation. None of the native Japanese speakers participated in the audio-recording sessions. According to self-report, all six groups of participants had normal hearing.

C. Procedure

The participants responded to 200 trials via a two-alternative forced-choice AXB discrimination task. The presentation of the stimuli and the collection of perception data were controlled by the PRAAT program ([4]). The participants were given two ('A', 'B') response choices on the computer screen. They were asked to select the option 'A' if they thought that the first two tokens in the AXB sequence were the same (e.g., ‘yoka$_2$’-‘yoka$_1$’-‘yokka$_3$’) and to select the option ‘B’ if they thought that the last two tokens were the same (e.g., ‘soto$_3$’-‘sotto$_1$’-‘sotto$_2$’; where the subscripts indicate different speakers). No feedback was provided during the experimental sessions. The participants could take a break after every 50 trials if they wished. The participants were required to respond to each trial, and they were told to guess if uncertain.

III. Results and Discussion

The overall mean discrimination accuracy in *d*-prime was 1.0, 1.7, 1.9, 2.0, 3.1 and 4.5 for the non-learner, N5, N3 (Vietnam), N3 (Japan), N1 and the native Japanese groups, respectively (Figure 1). One-way analysis of variance with Group (non-learner, N5 (Vietnam), N3 (Vietnam), N3 (Japan), N1 (Japan), native Japanese) reached significance [$F(5, 73) = 25.7, p < .001, \eta_G^2 = .64$].

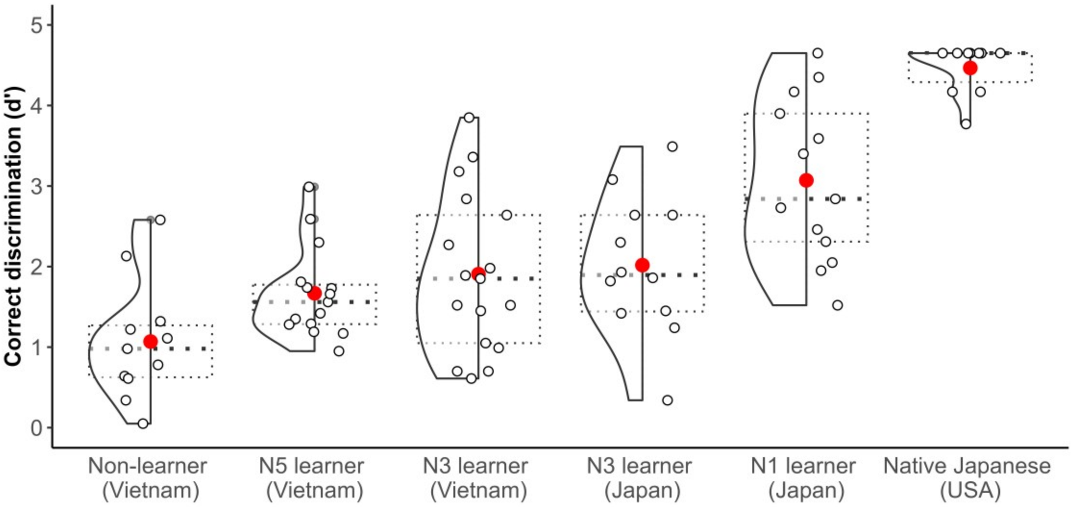


Fig. 1. Overall discrimination accuracy (*d*-prime) for six groups of participants. The horizontal line and the red circle in each box indicate the median and mean, respectively. The bottom and top of the box indicate the first and third quartiles.

Bonferroni-adjusted *t*-tests were used for post-hoc pairwise comparisons. Not surprisingly, the native Japanese group was at near ceiling with little individual variation and outperformed all other groups including the N1 group. On the other end of the spectrum, despite not having experience in Japanese, the non-learner group was significantly less accurate than the N1 and native Japanese groups only. In other words, the N3 (Japan), N3 (Vietnam) and N5 groups, who did not differ from one another, did not outperform the non-learner group. Only the highly experienced (N1, native Japanese) groups outperformed the non-learner group. Of the four groups of learners, the only difference that reached statistical significance was the one between N1 and N5 groups. Regardless of the difference in the countries of residence (Japan, Vietnam), the two N3 groups did not significantly differ from each other. Somewhat unexpectedly, the difference between the N1 and two N3 (Japan, Vietnam) groups did not reach statistical significance, either.

In summary, even with substantial learning and living experience in Japan, the N1 learners were still not native-like. This demonstrates genuine difficulty of Japanese consonant length, which has a pedagogical implication. However, only the N1 (but not N3) learners differentiated themselves from the non-learners and N5 learners in Vietnam. This clear difference is a testament to learnability of Japanese consonant length beyond early childhood in a naturalistic setting. On the other hand, while the difference between the N1 and N3 (Vietnam) approached reaching significance ($p = 0.06$), the lack of difference between the N1/N3 groups in Japan (both with LOR of 8.4 years) and the N3 group in Vietnam suggests that simply living in the target language country may not be sufficient to acquire skills to perceive Japanese singleton-geminate contrasts efficiently. As seen in Figure 1, a large spread and overlap for these three groups may have masked potential between-group differences. In fact, a majority of N3 learners (67% in Japan and 65% in Vietnam) had *d*-prime scores within the range of the N1 group (1.52 – 4.65). Given this, it may be necessary to use a greater variety of measures in addition to JLPT to accurately characterize learners’ proficiency and how it is related to their cross-language speech perception.

References

[1] The Japan Foundation (https://www.jpf.go.jp/j/project/japanese/survey/area/country/2020/vietnam.html)

[2] Q. Hussein and S. Shinohara, “Partial devoicing of voiced geminate stops in Tokyo Japanese,” JASA., vol. 145, pp. 149–163, 2019.

[3] S. Kawahara, “The phonetics of sokuon, or geminate obstruents,” In H. Kubozono, Ed., Handbook of Japanese Phonetics and Phonology. Walter de Gruyter, pp. 43–78, 2015.

[4] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” Computer program, 2016.

Pronunciation in Finnish speaking proficiency assessment - evidence from five L1s

Riikka Ullakonoja [a], Reeta Neittaanmäki [a], Tuija Hirvelä [a,]
[a] *Centre for Applied Language Studies, University of Jyväskylä, Finland*

Keywords — pronunciation, speaking proficiency, assessment, language testing

I. BACKGROUND AND RATIONALE

The aim of the paper is to investigate the role of pronunciation assessment in speaking proficiency assessment. The focus of the paper is on pronunciation, as it is very likely the main contributor to the perception of foreign accent and thus identification of the L1 of the speaker [1]. The data comes from the project *'Broken Finnish': Accent perceptions in societal gatekeeping* (2008-2024 Research Council of Finland) where the intermediate level Finnish examination in the National Certificates of Language Proficiency (NCLP) was studied in terms of foreign accent perceptions in language assessment. Sound samples from 49 speakers of five different first languages (L1s) were evaluated by experienced NCLP raters (n=44). [2] The NCLP Finnish examination has a gate-keeping role in the society, as most participants take it to demonstrate the language proficiency required for Finnish citizenship. The speakers' L1s were Arabic, Thai, Russian, Estonian and Finland-Swedish. They were chosen for the study, because most of them are big migrant groups in Finland and the Finland-Swedish population is the old minority group in Finland. All these groups also face negative attitudes. This was deliberately chosen in the research design to find out any possible bias in the rating process in the NCLP; if there were any, these would be the groups in which they would most probably show.

In language assessment, especially in a high-stakes language test, the raters are expected to rely on the official assessment criteria as a basis for their rating. However, recent findings [2, 3, 4] suggest that the raters can be influenced by factors external to the rating criteria, for example, in oral proficiency assessment attitudes are created by a certain foreign accent and this creates a real bias. There is evidence that professional raters in NCLP seem to be influenced by foreign accent in a way that if they recognized the accent, they were stricter in their pronunciation ratings [3]. This is not surprising as attitudes towards a speaker group are known to affect perception of accent [5].

II. MATERIAL AND METHODS

A. Material

The recordings were made during a real-life test taking situation, mostly in a language laboratory using a headset microphone, and saved in .wav format. The test takers (n=49), on the intermediate level Finnish speaking proficiency examination, were asked to speak for 90 seconds about an every-day situation. For the purposes of the project *'Broken Finnish': Accent perceptions in societal gatekeeping* 44 trained NCLP raters rerated the samples. Most raters reported to be female, have Finnish as L1 and to work Finnish as L2 teachers. Their age and experience in teaching and rating varied much. In the rating task, the samples were played to the raters in a randomized order, and they were asked to respond to several questions per sample. They were asked to rate the sample on the NCLP scale from 1-6 (corresponding to CEFR scale A1-C2). In addition to the general criterion usually given in the NCLP assessment, they were also asked to rate the six analytical criteria (fluency, flexibility, coherence/cohesion, propositional precision/range/idiomaticity, pronunciation/phonological control/grammatical accuracy) as well as to comment on their rating or the speaker. In addition, the raters were asked in an open-ended question "What do you assume the speaker's L1 to be?" and they were asked to justify their response in an open-ended question. For a more thorough description of the rating platform and questionnaire, see [2].

B. Methods

First, averages were calculated to get an overview of the data. Second the data was analysed with the Many-Facets Rasch model (MFRM) [6] using three-facets rating scale model (raters, speech samples, criteria) and bias analysis was used to investigate the interactions between the analytical criteria and L1s.

III. RESULTS

Fig 1 shows an overview of the data in terms of the rating averages by criterion in relation to L1. The five speaker groups differ in speaking proficiency rating: Estonian and Finland-Swedish L1 speakers are the most proficient (C1-B2) according to all criteria, Arabic and Thai L1 speakers are the least proficient (A2-B1) while Russian L1 speakers are in the middle (B1-B2). The comparison of the six analytical criteria (fluency, flexibility, coherence, vocabulary, pronunciation and grammar) shows, that pronunciation stands out as having the highest rating for all languages except for Estonian. For Estonian pronunciation and fluency have the same rating (4.3) and overall, the criteria differ very little from each other.

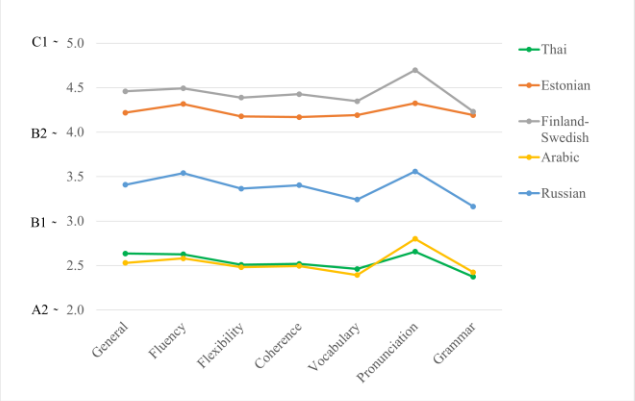


Fig. 1. Rating averages by criterion in relation to L1

As Fig. 2 shows the bias analysis of the criteria and L1 (derived from MFRM analysis), shows that pronunciation (in brown) indeed differs from other criteria and that it differs significantly depending on the L1.

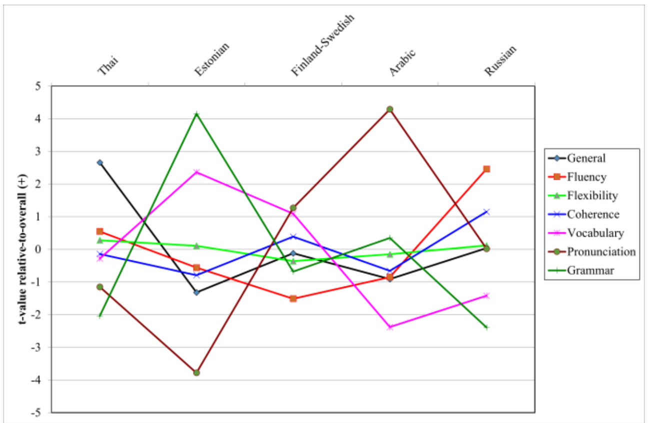


Fig. 2. The results from criterion-by-L1 Bias analysis

IV. Discussion and conclusions

The results show the peculiar role of pronunciation in the speaking proficiency assessment as compared to the other criteria that we will further discuss in the presentation. It also behaves very differently in the L1 speaker groups. The findings contribute in developing the assessment in ethical sense by providing detailed information on both phonetic and social features that may influence the assessment.

References

[1] Ullakonoja, R. Venäläisen aksentin tunnistaminen suomen suullisen kielitaidon arvioinnissa. In: Toivola, M., Lintunen, P., Heikkola L.M. (eds.) Puheen tutkimuksen uusia suuntia – Aineistona vapaasti tuotettu puhe. New directions in speech research – freely produced speech as data. AFinLA-teema/n:o 17, 2024, pp. 49–76. https://doi.org/10.30660/afinla.137869

[2] Halonen, M., A. Huhta, S. Ahola, T. Hirvelä, R. Neittaanmäki, S. Ohranen & R. Ullakonoja. Ensikielen tunnistamisen merkityksestä suullisen kielitaidon arvioinnissa Yleisissä kielitutkinnoissa. In: S. Grasz, T. Keisanen, F. Oloff, M. Rauniomaa, I. Rautiainen & M. Siromaa (eds.) Menetelmällisiä käänteitä soveltavassa kielentutkimuksessa – Methodological turns in applied language studies. AFinLAn vuosikirja 2020. Jyväskylä: Suomen soveltavan kielitieteen yhdistys AFinLA, 2020, pp. 56–70. https://doi.org/10.30661/afinlavk.89453

[3] Ahola, S. Rimaa hipoen selviää tilanteesta. Yleisten kielitutkintojen suomen kielen arvioijien käsityksiä kieli-taidon arvioinnista ja suullisesta kielitaidosta. [Barely passing the test task – NCLP Finnish raters' beliefs about language assessment and spoken language skills]. University of Jyväskylä. 2022

[4] Halonen, M., Ahola, S., Hirvelä, T., Neittaanmäki, R., Ohranen, S. & Ullakonoja, R. Kielitaidon arviointi kansalaisuuden portinvartijana. In: Renvik, T.A., & Säävälä, M. (eds.). Kotoutumisen kokonaiskatsaus 2023: Näkökulmana väestösuhteet. TEM oppaat ja muut julkaisut 2024:1 Helsinki: Työ- ja elinkeinoministeriö. 2024, pp. 77–86.

[5] Moyer, A. Foreign accent – the phenomenon of non-native speech. Cambridge: Cambridge University Press. 2013.

[6] Linacre J.M. Many-Facet Rasch Measurement, 2nd Ed. Chicago: MESA Press. 1994.

Coping with reverberant acoustics in singing by extending the plosive closures in vowel-plosive-vowel sequences

Allan Vurma [a], Einar Meister [b], Lya Meister [b], Jaan Ross[a], Marju Raju[a], Veeda Kala[a], Tuuri Dede[a]
[a] *Estonian Academy of Music and Theatre, Estonia,*
[b] *Tallinn University of Technology, Estonia*

Keywords — singing, text intelligibility, plosive closures, room acoustics, reverberation

I. INTRODUCTION

The intelligibility of both singing and speaking can be affected by factors such as masking by room reverberation and sounds produced by accompanying instruments and ensemble partners [1]. Achieving good text intelligibility in singing is often particularly challenging. For example, in one study [2], text intelligibility in sung phrases decreased by 76% compared to spoken phrases. This may mean that sometimes sung words can be completely lost for listeners, and the singer may be faced the dilemma of whether or not to seek additional vocal technical means to improve the intelligibility of the text even if it affects, for example, the good cantilena.

In this study, the hypothesis is posited that extending the closure duration of voiceless plosives in vowel–voiceless plosive–vowel (VCV) sequences sung in reverberant acoustics (e.g., in a concert hall) enhances the recognition of plosives. In such cases, the masking effect of the reverberation on plosives is reduced because the level of the reverberation field starts to decay during the plosive's closure. The longer the closure, the lower the masking by reverberation at the start of the plosive burst and the transition of formants. Some previous studies have suggested that adding extra pauses between words or at prosodic boundaries in synthetic speech can improve the intelligibility of text spoken in reverberant acoustics [3][4][5]. However, the impact of elongation of the plosive closures on sung text intelligibility is inconclusive.

II. MATERIALS AND METHOD

The core of this study comprises perceptual experiments utilizing VCV stimuli with modified closure durations in various acoustic environments. To select the closure durations for the stimuli, we initially analyzed performances of Italian-language Classical or Romantic period opera arias from their repertoire by 11 professional classically trained singers. The singers were also requested to read aloud the text of the aria (1) spoken in normal conversation, and (2) in an oratorical style. The recordings were conducted in a studio with negligible reverberation (T30 = 0.2 s), enabling us to gain insights into the behavior of singers in rooms where there is no disturbing masking by room reverberation.

The research material was segmented and closure durations were measured with Praat [6]. Closure durations were the longest (mean = 93.8 ms, sd = 48.3 ms) in the oratorical reading. Sung performances exhibited the shortest closure durations (mean = 74.7 ms, sd = 48.2 ms) while conversational speech fell in between (mean = 84.7 ms, sd = 38.6 ms). In all performance styles, especially in singing, duration of outliers reached up to 300 ms. Based on these data, we opted to employ three values for the plosive closures in VCV stimuli, with durations as follows: 60 ms, slightly below the median observed in the analyzed sung performances; 260 ms, aligning more with the closure durations of outliers; and an intermediate duration of 150 ms, positioned between these two extremes.

For the perceptual experiments, the recordings of /a-k-a/, /a-p-a/, and /a-t-a/ sequences sung by two classically trained professional opera singers – a mezzo-soprano and a tenor – served as the basis for stimuli. The closure durations of plosives in VCV stimuli were manipulated in Praat, while burst durations and transitions to the following vowel were left unchanged to preserve naturalness. The Praat Vocal Toolkit [7] was used to create stimulus sets with simulated reverberation (Church (Ch) or Big Room (BR)) and with or without Brown Noise (BN) to mimic the masking effect of accompaniment. The final stimulus set included three series I: based on tenor recordings at pitch G3, II and III: based on mezzo-soprano recordings at pitch G4 and F5, respectively. Consequently, the paradigms of the test series I and II each included 90 stimuli (3 plosives x 3 closure durations x 2 burst intensities x 5 acoustic conditions). In series III, the contrast of burst intensities was omitted while the acoustic condition Clear (Cl) was added, resulting in 54 stimuli (3 plosives x 3 closure durations x 6 acoustic conditions).

In the first experiment, 34 listeners (11 males, 23 females) participated in Praat-administered perception tests in a soundproof booth using the same setup (a laptop, calibrated external audio card, Sennheiser HD 560s headphones). In the second experiment, 33 listeners were seated in a real concert hall, and the stimuli (without simulated reverberation) were played from a loudspeaker on stage. In the concert hall, Brown Noise was played from a separate loudspeaker above the stage. Listeners were required to identify the plosive they heard in the stimuli.

To analyze the results of the perception tests, a Generalized Linear Mixed Effect Model (GLMM) for each pitch series was fitted with *Duration, Acoustic condition, Consonant, Burst* as fixed effects, and intercepts for *Age group* and *Gender* as random effects, using the lme4 package [8] in R [9]. Post hoc tests for fixed effects were performed with the *emmeans* package [10].

III. RESULTS

GLMMs confirmed a significant increase in correct responses for stimuli with longer closure durations in most acoustic conditions, observed in both listening tests: with artificially added reverberation (Figure 1, left) and in the real concert hall setting (Figure 1, right). In artificially added BR acoustics, the improvement was 24 percentage points, in Ch acoustics 11 percentage points (see left panel), and in a real concert hall 10 percentage points (see Cl_ALL in right panel). However, there was no improvement in recognition when BN was added to the stimuli. Recognition of plosives was also better when the pitch of the stimuli was lower (compare curves G3, G4, and F5 in the right panel). Additional analysis of the results showed that correct responses increased for stimuli with a stronger plosive burst. The recognition was also better among younger listeners and when the listeners were seated in the front rows (as opposed to the back rows) of the concert hall.

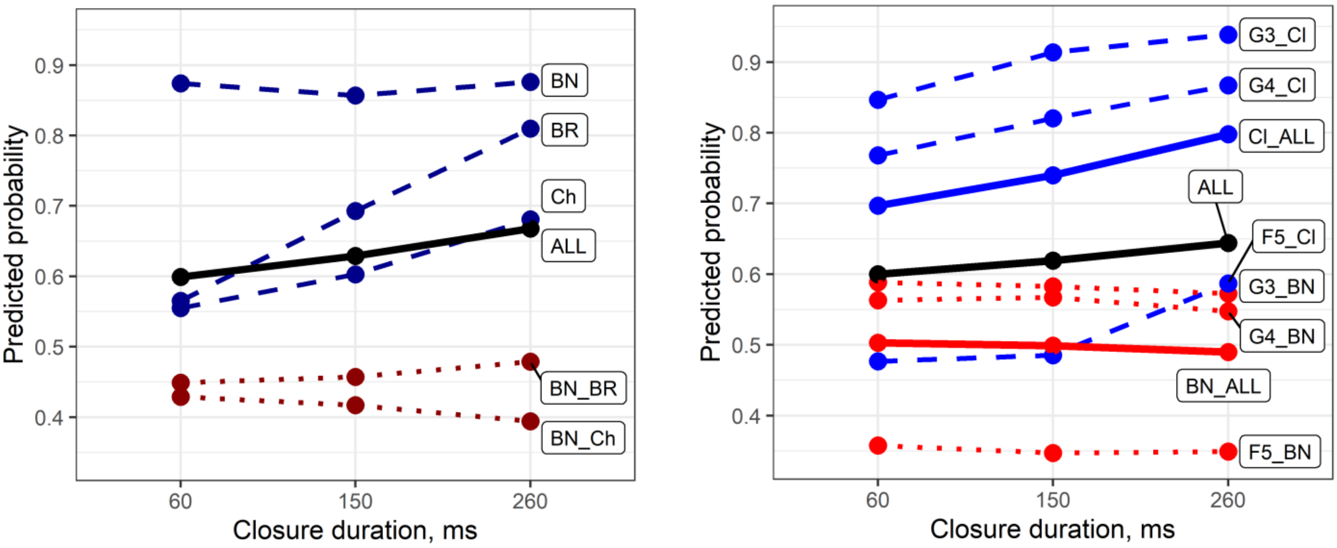


Figure 1. GLMM-predicted probability of correct responses in tests with artificial reverberation and noise (left), and in a real concert hall (right). BN – with added Brown Noise, BR – with added Big Room reverberation, Ch – with added Church reverberation, Cl – stimuli as recorded in a studio; G3, G4, F5 – pitches of the stimuli.

IV. DISCUSSION AND CONCLUSIONS

The results of the perception test showed that elongating the plosive closure in sung VCV sequences in the acoustics of typical concert halls can improve the recognition of plosives and, therefore, the overall intelligibility of the sung text. The plosive closure duration and the intensity of the plosive burst are the only factors that singers can adjust. The other studied factors – the acoustics of the performance venue, accompaniment, type of plosive, and pitch – are either prescribed by the composer or are beyond the singer's control. However, further investigations are needed to assess whether singers actually utilize these features and whether elongating the plosive closure could cause other issues, such as worsening legato or disrupting the naturalness of perceived prosody.

V. REFERENCES

[1] J. Meyer, Acoustics and the Performance of Music. Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instrument Makers. 5th edition, Springer, New York, 2009.

[2] L.B. Collister and D. Huron, "Comparison of Word Intelligibility in Spoken and Sung Phrases", Empirical Musicology Review, vol. 3, no. 3, pp. 109–125, 2008.3.

[3] P. N. Petkov, N. Braunschweiler, and Y. Stylianou, "Automated pause insertion for improved intelligibility under reverberation". in Proceedings INTERSPEECH 2016, September 8–12, 2016, San Francisco, USA.

[4] P. J. Scharpff and V. van Heuven, "Effects of pause insertion on the intelligibility of low-quality speech," in Proceedings FASE/Speech-88 Symposium, 1988.

[5] V. Best, C. R. Mason, J. Swaminathan, E. Roverud, and G. Kidd, Jr., "Does providing more processing time improve speech intelligibility in hearing-impaired listeners?" 169th Meeting of the Acoustical Society of America, Pittsburgh, Pennsylvania, 18–22 May 2015, Psychological and Physiological Acoustics: Paper 1aPPb28.

[6] R P. Boersma and D. Weenink, 2024. *Praat: doing phonetics by computer* [Computer program]. Version 6.4.03, retrieved 4 January 2024, from http://www.praat.org/

[7] R. Correte, *Praat vocal toolkit*, [computer program], 2021–2022 https://www.praatvocaltoolkit.com/index.html (Last viewed Dec 08, 2022).

[8] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Usinglme4", *Journal of Statistical Software,* 67(1), 1–48. 2015.

[9] R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[10] Lenth R (2024). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.10.2, https://rvlenth.github.io/emmeans/.

Laryngeal processes in multilingual speech

Zsuzsanna Bárkányi & Zoltán G. Kiss

The Open University and Hungarian Research Centre for Linguistics & Eötvös Loránd University

***Keywords* — L3 phonology, L2 speech acquisition, postlexical processes, regressive voicing assimilation, presonorant voicing, speech perception**

Introduction and Hypotheses

This paper is aimed at advancing the field of multilingual speech research by exploring individual learner differences in the perception and production of voicing related phonological processes. The most influential second language (L2) speech learning models (SLM-r – [6], 2021; PAM – [2]; L2LP – [5]) emphasise the importance of learners' perception of similarities and differences between sounds in their languages for the acquisition of non-native speech. These models mostly focus on segmental contrasts, or on sound learning (SLM-r), but there is a significant lack of studies on the acquisition of "dynamic" allophonic alternations. Our current understanding of third language (L3) speech learning is primarily based on production studies. Less is known about the role of perception in the complex interplay between native and non-native sound systems (e.g., [7]). To fill this gap in research, we will take a closer look at the following laryngeal phonological processes in the speech of trilingual L1 Hungarian, L2 English and L3 Spanish speakers.

1. Regressive Voicing Assimilation (RVA) (between adjacent obstruents), which is present in participants' L1 Hungarian and in most true voice languages including Spanish.
2. Pre-sonorant Voicing (PSV), which is not present in participants' L1 and is a typologically uncommon process, but is present in their L3 Spanish but both Hungarian and English lack it.

   Both 1. and 2. are dynamic phonological alternations that apply within the word and across the word boundary (sandhi), and neither of them creates new segments.
3. Aspiration of voiceless stops in English (their Voice Onset Time), in which case a new phonetic category must be learned for L1 Hungarians (unlike English, Hungarian as well as Spanish are voicing languages, where there is no aspiration of stops).
4. Intervocalic voiced stop spirantisation in Spanish (SPIR), an allophonic alternation that creates new segments. This is a process that is absent in both Hungarian and English.

The following hypotheses are tested in the study:

1. As participants are highly competent speakers (at least B2 of CEFR) all the mentioned processes are perceived.
2. The dynamic alternations (PSV and SPIR) are more likely to be produced within the word than across word-boundary.
3. RVA is likely to be unlearned as its lack is the default scenario for aspirating languages.
4. PSV is less likely to be learned as it is a marked process, and furthermore, because Spanish input is highly variable.
5. Aspiration is less likely to be learned as it requires phonetic fine-tuning of the existing sound categories.
6. SPIR is likely to be learned as it creates a new segment, which is the more frequent allophone, and also, it is fairly salient.

Methods

Participants were fourteen young adults who acquired their non-native languages in non-immersion context but are highly competent in both L2 and L3 with no clear knowledge or usage dominance between the two.

The production experiments investigated voicing in /s/ before voiced obstruents (RVA) and sonorants (PSV) in the English and Spanish speech of these participants. VOT was measured in sentence-initial position, while SPIR was examined both within the word and across word boundary. All these acoustic analyses were carried out in Praat (v. 6.2.23; [3]). Voicing was measured as a percentage of the presence of voicing oscillation relative to the duration of /s/ using manual segmentation and visual inspection of the waveforms and spectrograms. VOT was measured as the time between the *last* release burst until the appearance of the following vowel's voicing/formants. This method made it possible to somewhat normalize the effect of the place of articulation on of the stops on VOT. The SPIR measurements used the technique in [1], [4]: the degree of spirantisation was determined by the difference between the intensity minimum of the consonant and the intensity maximum of the following vowel: the more lenited the consonant, the smaller the intensity difference is expected to be with respect to the following vowel; we also measured the median harmonics-to-noise ratio in the consonant to supplement the intensity-difference measurements.

For the perception experiments, we opted for a novel holistic approach, namely, a short story was recorded in both English and Spanish by two phonetically trained bilingual female speakers Then, the same short story was recorded, with RVA in English and no PSV in Spanish to mirror the L1 voicing patterns of listeners. Then voiceless stops were not aspirated in English, and in the Spanish text voiced stops were not spirantised. Participants ranked the recordings on a scale of five: from "native-like" to "non-native-like".

Results and Discussion

The production data indicate that L1 Hungarians applied RVA in both their English and Spanish speech: /s/ was predominantly voiced before voiced stops; this is a non-target like application of RVA in English, as English lacks RVA. However, neither their English, and more importantly, nor their Spanish speech contained PSV: /s/ remained predominantly voiceless before sonorants. The VOT data shows variation, some the participants applied the Hungarian (and Spanish) VOT setting for initial voiceless stops: mean VOT for these speakers was around or below 35 ms, which is usually the amount in non-aspirating languages, although certain speakers did produce English-like aspiration (or in between values) with a mean VOT of well over 35 ms. SPIR of intervocalic voiced stops was mostly absent in the Spanish speech of the participants.

Our results show that perception does not closely mirror production, but production – as proposed by SLM-r – is dependent on perception. Although both the lack of aspiration and the non-target-like application of RVA in English were perceived by most participants, they were not consistently attested in their speech. The lack of PSV and SPIR in Spanish remained mostly unnoticed, and as a result participants did not produce these processes. In order to find out whether PSV and SPIR were perceived at all but regarded as "irrelevant" individual or register-dependent features rather than position-dependent allophones, a follow-up forced choice perception AXB experiment was carried out where the only difference between A and B was SPIR (or the lack of it) or PSV (or the lack of it). Our results show that all the participants perceived SPIR and PSV in nonce words, while the lack of these processes in a Spanish discourse was not perceived. Based on this, Hypothesis 1 has to be rejected as not all of these processes have been perceived. Regarding the other hypotheses,

Hypothesis 2 is partly born out: SPIR is slightly more common within the word than in sandhi, but PSV is not acquired at all.

Hypothesis 3 has to be rejected too as learners keep applying RVA in English.

Hypothesis 4 is born out: PSV has not been acquired.

Hypothesis 5: the acquisition of aspiration gave mixed results, but it is not acquired across the board.

Hypothesis 6 is rejected too, as SPIR was attested in only a few instances.

As for learner groups, we observed three general variation patterns: 1. learners who neither perceived nor produced a process, that is, who did not learn them. This is the predominant pattern for PSV and SPIR in Spanish 2. Learners who perceived the process (or the lack of it), but failed to transfer it into their speech productions. This is the predominant pattern for RVA and a common pattern for VOT; and 3. learners who perceived and produced the process(es) under scrutiny. This has been attested for a few students regarding VOT, even if productions were often not entirely target-like. The fourth possible scenario, i.e. production is target-like, but is not perceived has only marginally been attested in our data.

All this suggests that the acquisition of new phonetic categories, if perceived, is easier than that of dynamic phonological processes where for sequential learners' L1 laryngeal patterns seem to be dominant. Perceptual salience and typological markedness seem to have a limited effect.

References

[1] Amengual, M. 2019. Type of early bilingualism and its effect on the acoustic realization of allophonic variants: early sequential and simultaneous bilinguals. Int. Journal of Bilingualism, 23(5), 954–970.

[2] Best, C.T. and Tyler, M.D. (2007) 'Commonalities and complementarities: Nonnative and second-language speech perception', in O.-S. Bohn and M.J. Munro (eds) Language Experience in Second Language Speech Learning: In honor of James Emil Flege. John Benjamins Publishing Company (Language Learning & Language Teaching), pp. 13–34. https://doi.org/10.1075/lllt.17.07bes.

[3] Boersma, and Weenink, D. (2022). Praat: Doing Phonetics by Computer [Computer Program]. Version 6.2.23. Retrieved from https://www.praat.org (May 2, 2023).

[4] Eddington, D. 2011. What are the contextual phonetic variants of in colloquial Spanish? Probus, 23(1), 1–19.

[5] Escudero, P. (2005) 'Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization'.

[6] Flege, J. and Bohn, O.-S. (2021) 'The Revised Speech Learning Model (SLM-r)', in, pp. 3–83. https://doi.org/10.1017/9781108886901.002.

[7] Gut, U., Kopečková, R., & Nelson, C. (2023). *Phonetics and Phonology in Multilingual Language Development.* Elements in Phonetics, Cambridge: Cambridge University Press.

Phonetic Adaptation/Realization of the '-ed' Past Morpheme in Native and Non-Native Productions

Lucas Da Silva, Joaquín Romero
Universitat Rovira i Virgili, Spain

Keywords — '-ed' past morpheme, voicing, place of articulation, clusters

I. Introduction

This study investigates the phonetic realization of the English '-ed' past morpheme in various phonetic contexts among native and non-native speakers of the language. One of the most common problems encountered by Spanish speakers of English is the pronunciation of the past tense of regular verbs (-ed) [2]. The reason behind it is that in Spanish there are no words ending in final /t/, and words that do end in a final /d/ are either realized as a voiceless fricative /θ/ (e.g., 'Madrid' pronounced as /maˈðɾiθ/), or as a voiced dental fricative /ð/ (e.g., 'Madrid' pronounced as /maˈðɾið/), or simply not pronounced (e.g., 'Madrid' pronounced as /maˈðɾi/). Additionally, some consonant clusters might pose difficulties for Spanish speakers and speakers of other first languages like Catalan. That is because post-vocalic consonant clusters are not typical at all in Spanish, and much less typical in Catalan than in English. English syllable structure can be complex, with examples like "strengths" demonstrating CCCVCCCC, while Spanish has a much simpler syllable structure like C(C)V(C).

Following this pattern, consonant sequences can become very complex in English across word boundaries. Even for native speakers of English, pronouncing certain sequences of two or more consonants can pose difficulties. To ease this challenge, individuals employ various strategies, one of which is cluster reduction. This strategy involves omitting one of the consonants to simplify pronunciation. In final clusters consisting of three or four consonants, it is typically the middle consonant that is omitted (e.g., asked /æskt/ becomes /æst/) [4]. If the sequence is across word boundaries, the past '-ed' morpheme is likely to be deleted, especially when followed by a consonant sound in a consonant cluster rather than a vowel sound [2], [3], [5]. Another strategy employed by native speakers to simplify consonant clusters is resyllabification, which entails breaking up a final consonant cluster when it precedes a word beginning with a vowel sound. In such instances, the final consonant of the cluster is shifted to the next syllable. For example, in the phrase 'She moved it.', a native speaker typically resyllabifies it as /ˈʃiˈmuvˈdɪt/ rather than /ˈʃiˈmuvdˈɪt/ [1], [4].

Given that only a few studies have investigated the production of past '-ed' morpheme clusters by non-native speakers [3], the purpose of this paper is to analyze the strategies that L1 Spanish/Catalan speakers employ to deal with clusters and consonant sequences that violate syllable structure rules of their native language(s).

II. Methodology

The study involved obtaining recordings of four participants: two native English speakers and two non-native speakers proficient in Spanish and Catalan. They were asked to read sentences containing consonant clusters with the '-ed' past morpheme, while being recorded in a sound booth at Universitat Rovira i Virgili in Tarragona, Spain. Both groups were instructed to read continuously, without pauses, to emulate natural speech and facilitate connected speech processes. The sentences, encompassing all possible combinations of English clusters with the '-ed' morpheme, were randomized and limited to a maximum of 10 words to ensure coherence. Native speakers were given 4 seconds per sentence, while non-native speakers were given 5 seconds per sentence, allowing the latter enough time for producing the sentences completely. The sentences for this particular paper contain the combinations 'stressed Sam', 'forced Sam', 'chained Noah', 'burned Noah', 'dumped Pam', 'disturbed Ben', 'kicked/shocked Kate', and 'begged Gaby'.

The sequences were analyzed using Praat, where the clusters were measured, isolating each sound to examine the presence or absence of the '-ed' past morpheme, the voicing of the alveolar sound when present, and the length of the assimilated sounds when the alveolar sound was absent to determine if this indicated potential blending of the articulatory gestures associated with the past marker.

III. Results

The analysis of this study focuses on three distinct phonetic contexts where the past '-ed' morpheme is followed by a consonant sound: homorganic environments where sounds share the same place of articulation (e.g., /ndn/ and /sts/), environments where sounds do not share the place of articulation (e.g., /ptp/ and /bdb/), and environments where sounds share the same general articulator (tongue) but at different places of articulation (e.g., /ktk/ and /gdg/).

In contexts characterized by homorganic environments (e.g., /ndn/ and /sts/), focusing on nasals and sibilants, preliminary qualitative results indicate that one native speaker consistently includes the alveolar sound for the '-ed' past morpheme, whereas it is frequently omitted by the second native speaker and all non-native speakers. Voicing patterns indicate that the alveolar stop aligns with the surrounding phonetic context, with voiceless articulation prevailing in voiceless environments and voiced articulation in voiced ones. As observed in Fig. 1, in the cases where one native speaker and the two non-native speakers omit the alveolar stop in 'stressed Sam' or forced Sam', pronouncing it as /ˈstrɛˈsæm/ or /ˈfɔrˈsæm/, the assimilated /s/ is longer, presumably indicating an intention to somehow mark the absent past morpheme. In the same way, when speakers delete the voiced alveolar stop in 'chained or burned Noah', pronouncing it as /ˈtʃeˈnoə/ or /ˈbɝˈnoə/, the assimilated /n/ is longer than a corresponding single /n/, suggesting again an intention to mark the absent past morpheme. In environments with different places of articulation (e.g., /ptp/ and /bdb/), focusing on voiceless and voiced bilabials, native speakers consistently include the alveolar sound for the '-ed' past morpheme, whereas non-native speakers do not. Voicing patterns reveal nuanced articulatory adjustments, with native speakers producing voiceless or voiced alveolar stops contingent upon the preceding phonetic context. This observation underscores the need for further investigation into the intricate dynamics of phonetic adaptation in diverse phonetic contexts. In these cases, when the alveolar stops were absent, there was not a noticeable longer bilabial closure in the acoustic signal. Lastly, in environments with different places of articulation but both involving the tongue (e.g., /ktk/ and /gdg), the alveolar sound for the 'ed' past morpheme is consistently present, with voiceless articulation maintained even when the preceding segments are voiced, with one exception for one non-native speaker, where the sound was voiced when the velar was voiced.

In summary, this research offers valuable insights into how the articulatory and phonological processes, as well as syllable structure differences across languages, influence the pronunciation of the '-ed' past morpheme in diverse phonetic contexts and among various speaker groups. These findings emphasize the significance of examining individual characteristics and cross-linguistic influences when studying the acquisition of second language phonology. Understanding how phonetic patterns are different in native and non-native speakers can provide valuable understanding into the complexities of L2 acquisition processes.

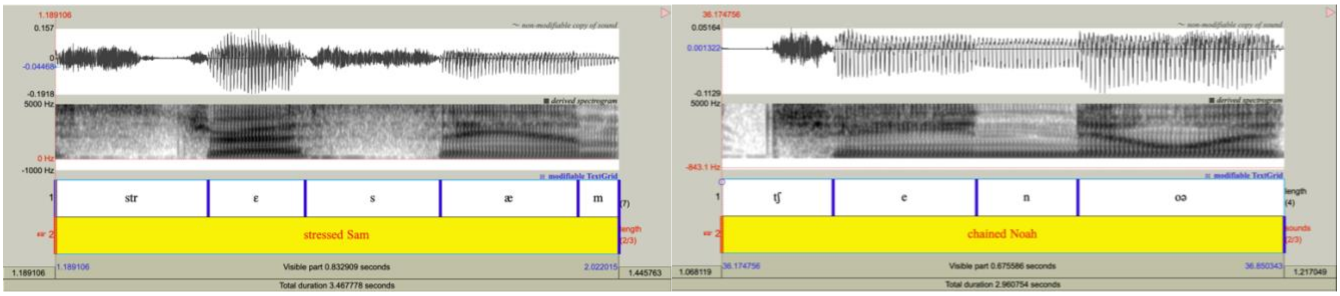


Fig. 1. Illustrations of native speaker omitting the past '-ed' morpheme in 'stressed Sam' (left) and 'chained Noah' (right)

References

[1] G. Alameen, "The effectiveness of linking instruction on NNS speech perception and production" (Doctoral dissertation, Iowa State University), 2014.

[2] P. Bell, P. Trofimovich and L. Collins, "Kick the ball or kicked the ball? Perception of the past morpheme–ed by second language learners." Canadian Modern Language Review, vol. 71, no. 1, pp. 26-51, 2015.

[3] D. R. Caballero and N. Rosado, "Neurolinguistic programming and regular verbs past tense pronunciation teaching." English Language Teaching, vol. 11, no. 11, pp. 1-18, 2018.

[4] M. Celce-Murcia, D. M. Brinton, D. M. and J. M. Goodwin, Teaching Pronunciation: A Course Book and Reference Guide. 2nd ed. Cambridge University Press, 2010.

[5] Cruttenden, A. (2014). *Gimson's Pronunciation of English*. Routledge.

Durational Aspects of fast speech in German

Reinhold Greisbach[a], Kornélia Juhász[b,c,d], Zsuzsa Szánthó[c,d], Szabina Zsoldos[c], Andrea Deme[c,d]
[a]*University of Cologne, Cologne, Germany*
[b]*HUN–REN Hungarian Research Centre for Linguistics, Hungary*
[c]*ELTE Eötvös Loránd University, Hungary*
[d]*MTA–HUN–REN NYTK Lendület "Momentum" Neurophonetics Research Group, Hungary*

Keywords — fast speech, consonants, duration

I. Introduction

Phonetic and phonological descriptions of the sound inventory and sound structure of a language are based on precise pronunciation by speakers of that language or (mechanically speaking) articulation at a slow speech rate. Words that are transcribed in pronunciation dictionaries are in most case only the slowly and clearly spoken version of a word. If the speaker speeds up, articulators have to move faster in order to produce more sounds in a shorter period of time. However, there are mechanical constraints on the speed and acceleration required to move an object, known as inertia. The inertia of the articulators limits the achievable speech speed. If the speaker tries to increase speed even further, it can only be achieved by reduction: lenition, and/or omission of some articulatiory movements that may also lead to missing sounds in the flow of speech (elision).

Classical methods aiming to analyse these reduction phenomena focus on spontaneous speech, which is assumed to be faster than "normal" speech. For German it is known that spontaneous speech contains many sound elisions in unaccented syllables. Reference [2] gives a set of rules observed in spontaneous speech of German, while [1] describes even more drastic reduction phenomena in the German word <eigentlich> ('actually'). Reference [3] used a method to accelerate read speech and study reduction phenomena under controlled conditions and found comparable results for faster speech rates. In the present study, we analysed how some German vowels, and consonants in phrase accented syllable onset position are realized in fast speech, which we elicited using the controlled read aloud method of [3].

II. Methods

The corpus used in this study was primarily designed for the comparative investigation of the process of speeding up on vowel length and quality in German and Hungarian. In this study, however, we are focusing only on vowel durations and syllable onset consonants of the German material. To avoid elisions and assimilations, we decided to use CVC-words in phrase accented positions. As we wanted to use this corpus for EMA-investigations as well, we included only labial and apical consonants in the onset and coda slots. The words ("X") were produced in the carrier phrase "Mehr X!" ("more X!") for the German corpus. The speakers were told to produce the first phrase at what they felt normal speech rate and then repeat it several times with ever increasing speed. The fastest sample was chosen to represent "fast speech" the first sample as "normal speech". 14 native female speakers of German (average 22.5 years) produced 12 words 6 times in random order resulting in 72 tokens for every speaker.

As mentioned, in this abstract we focus primarily on the effects of speeding up on the vowels, and onset consonants of the CVC syllables/words; effects on vowel length contrast and on Hungarian segments are discussed in [5]. The consonants were not balanced in the German material, because the corpus of onset consonants differed in Hungarian, and German and balancing the consonants would have inflated the number of tokens for the speakers to produce (while the task of producing 72 tokens in this experimental setting was already very demanding for them). We decided to use a semi balanced corpus of consonants containing only the labial [b], [f] and the apical [n], [z] in the onset, and thus we had /ba/, /baː/, /biː/, /bʊ/, /fuː/, /fɪ/ (labial), and /na/, /zaː/, /zɪ/, /ziː/, /nʊ/, /nuː/ (apical) as the onset-nucleus sequence. Data presented on vowels come also from these sequences. We measured speech rate (gauged as the duration of words), reduction rate (as duration ratio of fast and slow segment variants in percentage points where the smaller the ratio, the greater the reduction is), and voice offset time (VOffT; of prevocalic voicing) (only in /b/) at the beginning of the closure. Data were analysed using Pearson's correlation test, Welch test, and paired t-test.

III. Results

A. Phonemic Phenomena

Every speaker was able to speed up. Sound deletions, assimilations and lenitions (e.g. spirantization of plosives) did almost never occur, as the words under consideration were produced in phrase accented position. There was a positive strong correlation between speaking rate in normal speech [in ms] and reduction rate (across conditions) [in %] ($r = 0.63$; $p < .005$). This indicates that speakers who were speaking relatively fast in the normal speed condition, could not speed up as much as speakers who were speaking at a moderate rate in normal speed. At a certain speed, physiological and linguistic restrictions limit further reductions (of this type of phrase accented CVC-words).

Overall evaluation showed a stronger reduction of vowels (irrespective of phonological length) compared to syllable onset consonants. More specifically, in normal speech, vowels (136 ms) were longer than consonants (108 ms) on average. This difference was neutralized in fast speech where the average duration of the consonants (68 ms) equalled that of the vowels (67 ms). Furthermore, there was a strong positive correlation between reduction rate of the consonants [average reduction to 63%] and the vowels [average reduction to 49%] of $r = 0.68$ ($p < .05$) (inside every word), which indicates that some words (or to be more precise: some the combination of syllable onsets and nuclei) were reduced more than others.

Naturally, durations were strongly dependent on the (inherent) duration of the consonant type (see table 1.), while the reduction rate seemed to be similar across these types (except for /b/ which was reduced the smallest extent, see table 1).

TABLE I. CONSONANT REDUCTION RATE

Onset consonant	**Speaking rate**		
	Normal speed	***Fast speed***	***Reduction rate***
/n/ (N=252)	87 ms	52 ms	60%
/b/ (N=336)	99 ms	70 ms	71%
/z/ (N=252)	117 ms	69 ms	59%
/f/ (N=168)	150 ms	92 ms	61%

B. *Consonant to Vowel interactions*

The corpus contained some occasional quasi minimal pairs of the same consonant in front of a short-long vowel pair. In /b/ durations that preceded long /aː/ in <Bahn> (M = 74 ms), and /a/ in <Bass> (M = 67 ms), there was a statistically significant difference in fast speech (Welch t-Test $t = 2.92$, $df = 92.87$, $p = .004$), but not in normal speech ($M_{\text{before long /aː/}} = 90$ ms, $M_{\text{before short /a/}} = 87$ ms; Welch t-Test $t = 0.80$, $df = 93.79$, $p = .43$). Duration of /z/ before /iː/ ($M_{\text{before long /iː/}} = 131$ ms), and /ɪ/ ($M_{\text{before short /ɪ/}} = 115$ ms) (in the quasi-minimal pairs <sieht> and <Sinn>), differed at normal speed significantly (Welch t-Test $t = 3.51$; $df = 166$; $p = .0006$) but reduced in fast speech to 69 ms and 71 ms (Welch t-Test n.s.). Since these differences show divergent results, it must be assumed that reduction in fast speech is not systematic as a function of vowel length.

C. *Subphonemic Phenomena*

Our corpus also allowed us to get an insight into some subphonemic phenomena in fast speech. Specifically, we could investigate the phenomenon of voicing in German plosives. Although there is a phonemic distinction between voiced and unvoiced plosives in German, voiced plosives are not always completely voiced in syllable onsets. In our corpus, /b/ occurred before /a/ and /aː/ in intervocalic syllable onset position: /meːɐ **b**{a aː}/ Data showed that these /b/ realisations were only occasionally fully voiced. The average VOffT at the beginning of the closure of the plosive was 54 ms and thus 63% of the total duration of the /b/ on average in normal speech. In fast speech, however, average VOffT did not change in absolute time (56 ms) but the voicing ratio rose to 81% of the /b/, thus making /b/ realizations in fast speech relatively more voiced than /b/ realizations in normal speech (Paired t-test for the relative durations: $t = 5.99$, $df = 95$, $p < .0001$). This finding may reflect that the timing of the function of the laryngeal system remained stable, while the supralaryngeal articulators' movements increased in speed.

IV. DISCUSSION

We found that fast speech influenced both consonants and vowels, but vowels were affected more than consonants. For consonants, we observed different intrinsic compression rates as a function of manner of articulation, as plosives were more resistant to compression (about 70%) than continuants (about 60%). However, we found no systematic effect of the following vowel's length on the onset consonant's duration. We also observed that the supralaryngeal system (articulation) was able to speed up to about 50% in untrained speakers, while the laryngeal system (phonation) proved more resistant to tempo changes, resulting in the fact that in fast speech, the laryngeal system may still be in voicing modus while the supralaryngeal articulators (lips, tongue tip) are already in the configuration of the following (voiceless) sound.

REFERENCES

[1] O. Niebuhr and K. Kohler, "Perception of phonetic detail in the identification of highly reduced words" Journal of Phonetics, 39, pp. 319-329, 2011.

[2] K. Kohler, Einführung in die Phonetik des Deutschen, 2nd ed, Berlin: Erich Schmidt, 1995.

[3] R. Greisbach, "Reading aloud at maximal speed" Speech Communication*,* 11, pp. 469-473, 1992.

[4] B. Lindblom, "Explaining phonetic variation: a sketch of the H & H theory" in Speech production and speech modeling, W. J. Hardcastle and A. Marchal, Eds., Dordrecht, Netherlands: Kluwer Academic Publishers 1990, pp. 403–439.

[5] A. Deme, K. Juhász, Z. Szánthó, S. Zsoldos, R. Greisbach, "Duration and quality of German and Hungarian short and long vowels in fast speech", Abstract for the ISAPh 2024 congress.

Acknowledgements
The research was supported by the TKA–DAAD grant No. 177375., the NKFIH grant No. FK128814, and the ÚNKP-23-3-I grant (Zs. Sz.).

Exploring Pronunciation Pedagogy:
Form-Focused Instruction in Tunisian EFL Education

Author: Yosra Jaoua

Linguistics, English Department, Faculty of Arts and Humanities of Sfax, Tunisia

Keywords* — *pronunciation teaching; form-focused instruction; corrective feedback; segmental features; prosodic features; speech technology; human raters; intelligibility principle

I. Introduction

Nowadays, pronunciation is de-emphasized in Tunisia as it has become a casualty of Communicative Language Teaching (CLT) since its appearance (in the 1980s) where meaning and function are prioritized over form instruction under the assumption that it would improve through exposure [1]. This underlies the assumption that the primary focus of L2 pedagogy should be on meaning rather than on the form [2]. Despite this, non-native speakers' enunciations still exhibit pronunciation difficulties which may put them at professional and social disadvantages whether in EFL or ESL settings [3][4][5][6].

The impact of directing the attention of L2 /FL language learners toward phonological forms during meaningful communication has become a recent point of interest in research on pronunciation. One approach gaining attention is focus-on-form instruction (FFI). In this method, L2 learners engage in communicative tasks to practice and observe pronunciation features, as opposed to engaging in exercises and drills that are disconnected from context (i.e., focus on forms). FFI seeks to avoid excessive theoretical dogmatization in instruction. Accordingly, in FFI, language Ls are not expected to be aware of the complex anatomical, phonetic and phonological aspects of the speech sounds but rather to attend to them while performing communicative tasks [7].

In light of this, the present study delved into examining the distinct outcomes of FFI concerning the enhancement of pronunciation skills in Tunisian secondary school EFL learners. The research specifically focused on addressing the challenges associated with certain English pronunciation features that were deemed problematic through what Rod Ellis refers to as the remedial approach [8]. In doing this, we relied on the results of an action-research that was conducted in Tunisia in the year 2000. For the analysis of students' productions of the segmental features, we relied on an expert judgement approach. As for the suprasegmental features, we made use of machine-based acoustic analyses using PRAAT.

II. TAREGT FEATURES

As such, a total of sixty-three Tunisian secondary school learners of English partook in the study. The primary aim was to assess the effectiveness of FFI, both with and without corrective feedback (CF), in fostering the development of learners' prosodic as well as segmental pronunciation. All participants, except those in the control group, received a fifteen-hour FFI treatment designed to make them attend to and practice the target features. The latter included the falling and rising tones inherent in open versus check questions in interrogative sentences as well as demanding versus confirmation question tags. The study also investigated students' improvement in two graphemes that are problematic due to their inconsistencies viewing their correspondence to a high number of phonemes. These were the digraph <th> as representing both the voiced dental fricative /ð/ (as in ***th****ough)* and the voiceless dental fricative /θ/ (as in ***th****ought)* and the digraph <gh> as being silent as in the previous words or corresponding to /g/ as in the word 'spaghetti' or 'ghost' or /f/ as in the word 'laugh' or 'enough'. The control group received instruction through the regular course of the year without FFI. In this vein, it had a free conversation class without any feedback on the target features. During FFI, the instructor delivered CF only to students in the FFI + CF group by recasting their erroneous pronunciation of the target features mentioned above during focused tasks and through a dubbing project. On the other hand, no CF was delivered to the participants in the FFI-only group. The rationale behind choosing an experimental group with CF and another without was to investigate whether the effects of FFI vary according to whether or not learners receive CF (i.e. to see the extent to which the teacher's use of recasts improves pronunciation accuracy as an element of good communicative competence).

III. INSTRUMENTS

The participants were recorded reading eight sentences containing words that have the target segmental features. To analyze these speech tokens, two L1 English teachers provided the phonemic transcriptions of learners' productions of the target features. Based on these transcriptions, we assigned scores to learners' productions. For the analysis of these segmental features, the human judgments of phonemic accuracy rather than instrumental analysis were decided sufficient because the focus at this level was on phonemic errors that could affect comprehensibility and not on detailed phonetic errors reflecting allophonic variations [9]. Accordingly, the two raters are teachers who have just come to teach at a local American primary school and were assigned as they are still unfamiliar with the Tunisian English accent and thus were considered reliable in judging the degree of intelligibility, comprehensibility and accentedness of the participants' speech tokens. If the relevant consonant was pronounced correctly in each word, 1 point was assigned, and if it was pronounced incorrectly (i.e., substituted with another consonant or not pronounced), a score of 0 was assigned. This yielded two sets of scores for each participant for each target consonant. The mean scores between the two transcribers were calculated and regarded as each learner's score based on the results of intraclass correlation coefficient (ICC) for interrater reliability. As for the suprasegmental features, instrumental acoustic analyses of participants' dialogue readings were conducted on (F0) fundamental frequency values of English falling and rising boundary tones and compared to those of Native speakers.

REFERENCES

[1] T.M., Derwing, M.J., Munro, J.A., Foote, E. Waugh, and J. Fleming, "Opening the Window on Comprehensible Pronunciation After 19 Years: A Workplace Training Study". 2014. Language Learning, 64: 526-548. https://doi.org/10.1111/lang.12053

[2] T. M., Derwing & M. J., Munro . "Second language accent and pronunciation teaching: A research-based approach". 2005. TESOL Quarterly, 39, 379-397 http://dx.doi.org/10.2307/3588486

[3] J. Morley, "How many languages do you speak?" Perspectives on pronunciation-speech-communication in EFL/ESL. Nagoya Gakuin University Roundtable on Linguistics and Literature Journal, 1988. 19, 1-35. Nagoya, Japan: Nagoya Gakuin University Press.

[4] M. Celcia-Murcia, D. Brinton, & J. Goodwin, Teaching pronunciation: A reference for learners of English to speakers of other languages. 1996. Cambridge: Cambridge University Press.

[5] R. Ellis, "The importance of focus on form in communicative language teaching. Eurasian Journal of Applied Linguistics". 2015. 1. 1-12. 10.32601/ejal.460611.

[6] R. Ellis, "Rod Ellis's essential bookshelf: Focus on form," *Language Teaching*, vol. 57, no. 2, pp. 246–261, 2024. doi:10.1017/S026144482200012X

[7] Lan, Yizhou & Wu, Mengjie. Application of Form-Focused Instruction in English Pronunciation: Examples from Mandarin Learners. 2013. Creative Education. 04. 29-34. 10.4236/ce.2013.49B007.

[8] R. Ellis, "Researching the effects of form-focused instruction on L2 acquisition". 2006. Applied Linguistics, 19 (1), 18-41

[9] P. Ladefoged & K, Johnson A Course in Phonetics. 2015. Hampshire. Cengage Learning Publications.

Reduction of unstressed English vowels by EFL speakers with different language backgrounds

Heini Kallio [a], Kamil Kaźmierski [b]
[a] *Tampere University, Finland,*
[b] *Adam Mickiewicz University, Poland*

Keywords — EFL production, stress realizations, vowel reduction, vowel duration

I. Introduction

English is a language with varying word stress and relatively strong stress contrasts that are produced by controlling prosodic features as well as segmental quality of vowels [1]. Compared to stressed ones, vowels in unstressed syllables in English generally undergo centralization and are considerably shorter in duration. Languages with fixed word stress, in turn, have weaker contrasts between stressed and unstressed segments than English. EFL learners with fixed stress first language (L1) may thus have difficulties in both the positioning of English stress [2,3] and the reduction of unstressed vowels [4], which may hinder the intelligibility of their speech [5].

This study investigates the reduction of unstressed monophthong vowels by EFL speakers from three L1 backgrounds: Hungarian, Polish, and Slovak. In comparison to English, these languages have fixed word stress and acoustic realizations of stress contrasts are relatively weak [6,7,8]. Little or no reduction is considered to occur in unstressed vowels in Hungarian, Polish, or Slovak. However, some centralization effect has been found for all the three languages, but centralization seems to depend on the vowel category [9,10,11]. As for prosodic stress markers, the phonetic quantity distinction in Hungarian and Slovak may further hinder the use of duration as a cue for signalling prominence in these languages [7,11]. In this pilot study, we scrutinize the production of stressed vs. unstressed EFL vowels by measuring vowel space areas (VSA) and segment durations. Native English speakers are expected to produce higher level of reduction in unstressed vowels than EFL learners, but some differences can occur between EFL learner groups depending on their L1. Although EFL learners' stress production is widely studied, the current study is, to the best of our knowledge, the first one to use VSAs in comparing EFL stress productions of Hungarian, Polish, and Slovak speakers. Our objective is to provide novel insights into L1-specific characteristics that could prove beneficial in developing EFL teaching.

II. Material and Methods

A. Speech data

The speech data analysed here is part of a larger corpus that consists of 61 speech samples from EFL learners with either Slovak, Czech, Hungarian, or Polish as their L1 and seven speech samples from native British English speakers [12]. In this pilot study we use speech samples from 16 speakers: four speakers from each of the following L1s: Slovak, Hungarian, Polish, and English. The EFL speakers selected for this study were assessed at CEFR level B1 in their prosodic proficiency in English [12]. Each participant read aloud a narrative text consisting of 16 sentences. The recordings were done one participant at a time using either a recording studio or some other quiet space with a portable recording device. All participants were given two minutes to get acquainted with the text and were then asked to read the text aloud as if they were telling the story to someone. No pronunciation instructions were given. Each recording was approximately 1.5 minutes long.

B. Methods

Stressed and unstressed monophthong vowel segments were annotated in the speech data using a forced aligner [13] and the onsets and offsets of vowel segments were corrected manually in Praat [14]. Vowels were defined as stressed/unstressed based on the native productions and standard British English transcription of the narrative text. Nasalized vowels, vowels in context with rhotics and approximants, and vowels shorter than 15 ms were discarded from the analysis. Due to the limited amount of data per speaker, the consonantal context was not further controlled for. The final number of stressed vowels selected for analysis varied between 448 and 567 per speaker, and the number of unstressed vowels varied between 686 and 819 per speaker.

Formant measurements were extracted at 7 equidistant time points in vowel intervals along with segment durations using a Praat script. Here we use the midpoint measures of F1 and F2 to determine the stressed and unstressed vowel space areas. Formant measurements were normalized using the vowel-extrinsic version of the Nearey method [15], as it has recently been shown to perform best at preserving linguistically meaningful differences and normalizing out purely physiology-driven differences [16]. The VSAs and relative duration differences of stressed and unstressed vowels were compared between native speakers of English and EFL speakers with respect to their L1s.

III. RESULTS

The L1-specific areas of 80% probability density estimates for stressed vs. unstressed vowel spaces are shown in Fig. 1. As expected, the degree of VSA shrinkage differs between native and non-native English speakers. For the native English speakers, the unstressed VSA is 55% smaller than the stressed. The degree of shrinkage for the Polish EFL speakers is 35%, for the Slovak EFL speakers 31%, and for the Hungarian EFL speakers 37%. The differences between stressed vs. unstressed VSAs were statistically significant for the native speakers, for the Hungarian EFL speakers, and for the Polish EFL speakers ($p < 0.05$). As for stressed vs. unstressed vowel durations, the unstressed vowels produced by the native speakers were on average 30% shorter than stressed vowels, while the average segment shortening was only 4.5% for Hungarian EFL learners, 7.5% for Polish EFL learners, and 7.2% for Slovak EFL learners.

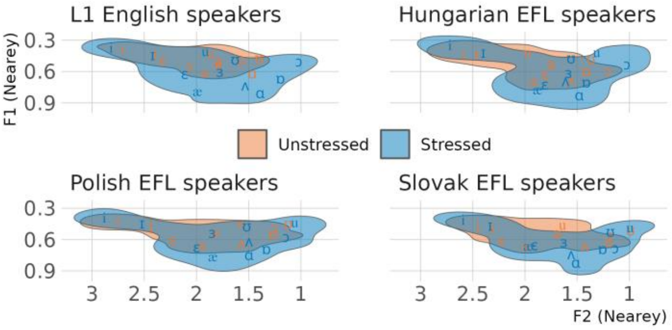


Fig. 1. Stressed vs. unstressed vowel space areas per speaker group.

IV. DISCUSSION

Results from the pilot study indicate that intermediate EFL learners differ from native speakers in unstressed vowel reduction in terms of both vowel centralization and duration, but the results also differ with respect to speaker L1. Our results support previous findings on the use of duration in producing English stress by speakers of Hungarian, Polish, and Slovak [12]: segment durations differ between natives and EFL speakers, but there also seems to be differences stemming from speakers' L1. It should be noted that lesser vowel reduction in EFL speakers compared to natives can stem from weak realizations of stress contrasts in EFL learners' L1s as well as misplaced word stress: here stress was defined based on the native productions and standard British English transcription of the text, and mistakes in stress placement likely affect the VSA and duration measures. It is also possible that some unstressed vowels undergo more coarticulation than centralization in both native and nonnative speakers [17]. In the future, we will analyse more speakers to further scrutinize the effect of speaker L1 on the centralization and duration of unstressed vowels by vowel category. We will also compare dynamic formant movements of vowels between speaker groups [18].

REFERENCES

[1] M. Halle and S. J. Keyser, "English stress", Harper and Row New York, 1971.

[2] J. Field, "Intelligibility and the listener: The role of lexical stress" TESOL Quarterly, 39 (3), 399–423, 2005.

[3] L. D. Hahn, "Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals" TESOL Quarterly, 38 (2), 201–223, 2004.

[4] L. Kalvodová and R. Skarnitzl, "Production and perception aspects of weak forms in Czech-accented English", In Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices, 108–121, 2023.

[5] A. Lepage and M. G. Busà, "Intelligibility of English L2: The effects of incorrect word stress placement and incorrect vowel reduction in the speech of French and Italian learners of English", In Proceedings of the International Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics, vol. 5, 387–400, 2014.

[6] S. Beňuš, U. D. Reichel and K. Mády, "Modeling Accentual Phrase Intonation in Slovak and Hungarian", In L. Veselovská & M. Janebová (Eds.), Complex Visibles Out There, Olomouc, Czech Republic: Palacký University, vol. 4, pp. 677–689, 2014.

[7] I. Vogel, A. Athanasopoulou, and N. Pincus, "Acoustic properties of prominence in Hungarian and the Functional Load Hypothesis", In K. É. Kiss, B. Surányi, & É. Dékány (Eds.), Approaches to Hungarian: Papers from the Piliscsaba Conference, vol. 14, pp. 267–292, John Benjamins, 2013.

[8] A. Cwiek and P. Wagner, "The acoustic realization of prosodic prominence in Polish: Word-level stress and phrase-level accent", In Proceedings of 9th International Conference on Speech Prosody 2018, Poznań, Poland.

[9] A. Rojczyk, "Quality and duration of unstressed vowels in Polish", Lingua, vol. 217, 80–89, 2019.

[10] A. Markó, M. Bartók, T. E. Gráczi, A. Deme and T. G. Csapó, "Prominence effects on Hungarian vowels: A pilot study", In Proceedings of 9th International Conference on Speech Prosody 2018, Poznań, Poland.

[11] S. Beňuš and K. Mády, "Effects of lexical stress and speech rate on the quantity and quality of Slovak vowels", In Proceedings of 5th International Conference on Speech Prosody 2010, Chicago, Chicago, USA.

[12] H. Kallio, A. Suni and J. Šimko, J., "Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of English with different language backgrounds", Language and Speech, 65(3), 571–597, 2023.

[13] T. Kisler, U. Reichel and F. Schiel, "Multilingual processing of speech via web services", Computer Speech & Language, vol. 45 , 326–347, 2017.

[14] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer", Version 6.1.54. Online: http://www. praat.org, 2016.

[15] T. M. Nearey, "Phonetic Feature Systems for Vowels", Dissertation, University of Alberta, 1977. Reprinted 1978 by the Indiana University Linguistics Club.

[16] S. Barreda, "Perceptual validation of vowel normalization methods for variationist research", Language Variation and Change, 33 (1), pp. 27–53, 2021. doi:10.1017/S0954394521000016

[17] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory". In Hardcastle, William; Marchal, Alain (eds.). Speech production and speech modelling. Dordrecht: Kluwer, 403–439, 1990.

[18] G. Schwartz and K. Kaźmierski, "Vowel dynamics in the acquisition of L2 English–an acoustic study of L1 Polish learners", Language Acquisition, 27 (3), pp.227-254, 2020.

Perceptual Evaluation of Attitudinal Expressions

Hansjörg Mixdorff[a], Albert Rilliard[b], Navneet Nayan[c]
[a] *Berliner Hochschule für Technik, Germany*
[b] *Université Paris Saclay, CNRS, LISN, France*
[c]*IIT Roorkee, India*

Keywords — ***social attitudes, Hindi, German, auditory-visual speech, free labeling***

I. Introduction

Attitudes significantly shape speech communication by guiding dialogue navigation and understanding. Our beliefs and cultural norms heavily influence how we express and interpret attitudes, crucial for effective dialogue management. During face-to-face interactions, we rely on linguistic and nonverbal cues to comprehend attitudes, but misunderstandings can occur, especially across cultures. Research comparing attitudinal expressions in German and Cantonese speakers revealed both parallels and differences. Utilizing methods outlined by X et al. [1], we identified sixteen distinct attitudes, such as arrogance and politeness, each linked to specific communication goals. Participants assessed stimuli representing these attitudes, providing single-word descriptions. Analysis in emotional space highlighted disparities between the two language groups, with Cantonese speakers better understanding German utterances [2]. While valence judgments aligned, agreement on activation and dominance varied. Expanding our research, we collected attitudinal expressions in Hindi[3]. Our study compares these with the earlier experiments involving German stimuli. Interestingly, many Hindi-speaking participants used English to assess attitudes, prompting us to draw on emotional analysis of English words by Warriner et al. [4]. Despite differences in terminology, their rating scale provided finer granularity than ours. While valence values correlated highly between the systems, activation/arousal and dominance correlations were weaker. Our reference system, however, showed a moderate correlation between activation and dominance. Although numerical values aren't directly comparable to earlier studies, mapping responses onto the same reference system mitigated these limitations. This system, grounded in perceptual tests with a large subject pool, offers higher granularity than expert judgments. Multidimensional analysis revealed a primary dimension akin to "unpredictability," but also showed that prosodic expressions clustered around the dichotomy of assertive/interrogative.

II. Stimuli and Experimental Design

In prior publications [1][2], we outlined our methodologies for capturing attitudinal expressions, a summary of which follows. We employed scripted dialogues to elicit 16 distinct attitudes, each portrayed through exchanges between presenter and experimenter. These dialogues generated target phrases—either "a banana" or "Mary was dancing"—in the presenter's language, which were then recorded as short video clips for subsequent perceptual tests. For the experiment involving the Hindi corpus, we selected the top-rated male and female presenters from previous studies, resulting in 128 auditory-visual stimuli. Additionally, a subset of 32 stimuli was created in reduced modalities: audio-only and silent video. These stimuli were presented to participants via an online survey on the PsyToolkit server [5][6]. After an introductory explanation in either Hindi or German, participants viewed the stimuli in randomized order, providing verbal judgments via text input. Each stimulus could be replayed once, and a progress bar aided orientation. In the experiment with German-speaking subjects, 160 stimuli were presented randomly from the total of 192. Thirty-four participants, mostly students aged 19-34, completed the task in 40 to 60 minutes. Responses were collected and categorized based on the participants' native languages and the modalities of stimuli. Similarly, in the experiment with Hindi-speaking participants, the number of stimuli presented in one session was reduced to 80, with participants required to complete the experiment twice to cover all 160 stimuli. Participants, mostly students from IIT Roorkee and IIT Bombay, completed the task in 15 to 30 minutes. Responses, predominantly in English, underwent translation and verification for accuracy and relevance to the intended attitude. The replies in German were translated to English. After corrections, unique terms were extracted and mapped onto a scale for analysis. Attitude distributions were analyzed using Multiple Factor Analysis[7][8], revealing five main dimensions that explained approximately 65% of the total variance. A hierarchical clustering algorithm further grouped attitudes into five clusters based on these dimensions.

III. Results and Conclusions

We computed mean values of valence, arousal, and dominance for response terms, mapping the attitudes in a three-dimensional emotional space. Table 1 displays average values for all 16 attitudes in the AV condition for both German and Hindi raters in the Hindi corpus. Notably, attitudes like ADMI rank highest in valence, while CONT and IRRI are perceived as most negative, consistent with previous findings. Arousal tends to be low overall, with attitudes like DECL and SINC eliciting the least arousal. Conversely, attitudes like DOUB and IRRI rank higher on the arousal scale. Perceivers feel most in control with positive attitudes like POLI and ADMI, while attitudes like DOUB and UNCE imply insecurity. Hindi speakers tend to rate positive attitudes higher in valence compared to German speakers (red in Table 1). Group differences are significant across all emotional dimensions (p <

.001). We will now examine the similarities and differences between the two rater groups. To that end we performed a multivariate GLM-based analysis of the dependent variables *valence*, *arousal* and *dominance*. As independent factors we introduced the type of *attitude*, the *language of the rater*, and the *modality.*

Table 1: Valence/Arousal/Dominance values for sixteen attitudes for German and Hindi raters.

attitude	abbrev-	V	A	D	V	A	D
	iation	Ger	Ger	Ger	Hin	Hin	Hin
admiration	ADMI	.33	-.02	.21	.51	.13	.30
arrogance	ARRO	-.28	-.08	-.06	-.29	-.04	-.01
authority	AUTH	-.05	-.18	.07	-.00	-.19	.18
contempt	CONT	-.28	-.10	-.10	-.25	-.10	-.03
neutral statement	DECL	.04	-.25	.11	.06	-.25	.22
doubt	DOUB	-.06	-.02	-.05	.07	.00	-.02
irony	IRON	-.05	-.11	.01	-.04	-.08	.03
irritation	IRRI	-.24	-.05	-.06	-.26	.01	-.03
obviousness	OBVI	.00	-.18	.08	.07	-.18	.18
politeness	POLI	.32	-.24	.30	.40	-.14	.38
neutral question	QUES	.00	-.11	.02	.10	-.13	.07
seductiveness	SEDU	.19	-.09	.14	.35	.04	.23
sincerity	SINC	.20	-.31	.24	.20	-.24	.30
surprise	SURP	.12	.04	-.00	.27	.12	.05
uncertainty	UNCE	-.21	-.11	-.13	-.22	-.12	-.13
walking-on-eggs	WOEG	-.22	-.16	-.12	-.24	-.10	-.11
	total	-.01	-.12	.04	.04	-.08	.10

Results show that all three factors and some of their combinations have a highly significant influence on the three emotional dimensions, though for space reasons we omit the details here. We also examined the agreement between rater groups by comparing the stimulus-wise results. To that end we step away from the originally intended attitudes and evaluate the three emotional dimensions associated with how the perceivers interpreted those stimuli. We calculated means and standard deviations of valence, arousal and dominance for each stimulus in our experiments as a function of the rater group and yielded Pearson's r of the stimulus-wise judgements to examine the agreement between the two rater groups. For comparison, we also calculated split-correlations inside the German and Hindi speaking groups on the Hindi stimuli. The results can be seen in Table 2. The agreement is obviously higher on Hindi data than on German data. One reason for the difference may be the smaller number of presenters in the Hindi experiment (N=8) as compared to the German one (N=15). Kruskal-Wallis test of independent samples reveals significant differences between audio-only (AU) presentation and AV and VI modalities, valence tends towards neutrality when the face is not visible. Arousal is slightly reduced in AU condition for most attitudes, while dominance increases. MFA-based clustering analysis reveals distinct groups of attitudes. Cluster #1: DOUB and SURP, characterized by terms like shock, surprise, and doubtful. Cluster #2: QUES, UNCE, and WOEG, characterized by terms like unsure, hesitant, and worried. Cluster #3: ADMI, IRON, SEDU, characterized by terms like praise, happy, and delighted. Cluster #4: ARRO, CONT, IRRI, characterized by terms like irritation, arrogance, and rude. Cluster #5: AUTH, DECL, OBVI, POLI, SINC, characterized by terms like confident, calm, honest.

Table 2: Stimulus-wise intra- and inter-rater group correlations (Pearson's r), top: Hindi stimuli, bottom: German stimuli.

rater groups compared	valence	arousal	dominance
German split	0.860	0.777	0.875
Hindi split	0.803	0.758	0.747
German:Hindi	0.828	0.802	0.834
German split	0.799	0.684	0.765
Hindi:German	0.671	0.570	0.683

This shows that in fact the sixteen attitudes overlap and many cannot be differentiated without additional information such as specific wording or even familiarity with the speaker. The analysis based on lemmas highlights a significant distinction mirroring the main axis of the MFA analysis. Clusters #1 and #2 contrast with clusters #4 and #5, representing expressions of interrogation, uncertainty, or doubt versus politeness, assertion, and authority. This parallels a linguistic and communicative function where interrogative and assertive utterances oppose each other. This dimension reflects the appraisal of novelty and unpredictability versus expectedness or familiarity, emphasizing the intertwining of attitudes in both emotional and linguistic systems. Using a lemma base created from text, not audio-visual stimuli, has limitations, especially with translation. However, we verified, that even after translation, 85% of the variance was preserved. While direct polling of valence/arousal/dominance values from perceivers could avoid term translation, it would not allow for the semantic analysis presented here and would increase experimental costs with multiple tasks for subjects.

REFERENCES

[1] A. Rilliard, D. Erickson, T. Shochi, and J.A. de Moraes,, "Social face to face communication - American English attitudinal prosody", INTERSPEECH 2013. 1648-1652.

[2] H. Mixdorff, A. Rilliard, T. Lee, M.K.H. Ma, A. Hönemann: Cross-cultural (A)symmetries in Audio-visual Attitude Perception. Proceedings of Interspeech 2018, Hyderabad, India.

[3] Mixdorff, H., Nayan, N., Rilliard, A., Rao, P. and Ghosh, D. 2023. Developing a Corpus of Audio-visual Attitudinal Expressions in Hindi. Proceedings of ICPhS 2023, Prague, Czech Republic.

[4] A.B.Warriner, V. Kuperman, M.Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas, Behavior Research Methods, 2013, Volume 45, Number 4, Page 1191.

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

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

[7] F. Husson, S. Lê, and J. Pagès, Exploratory multivariate analysis by example using R, Second edition. Boca Raton: CRC Press, 2017.

[8] S. Lê, J. Josse, and F. Husson, "FactoMineR: An R Package for Multivariate Analysis," J. Stat. Soft., vol. 25, no. 1, 2008, doi: 10.18637/jss.v025.i01.

Pronunciation peer-teaching revisited: a pilot study

Kristýna Červinková Poesová
Faculty of Education, Charles University in Prague

Keywords* — *primary teaching, teacher training, pronunciation peer-teaching, stimulated recall interview

One of the key factors contributing to EFL/ESL teachers' uncertainty and/or unwillingness to address pronunciation in their classes has been a lack of teacher training in phonology and/or pronunciation pedagogy [1], [2], [3], [4]. Even in contexts where pronunciation training occurs, there are still some unresolved aspects. For instance, teacher training in pronunciation pedagogy in the Australian context proved effective in the short run, particularly in increasing teacher trainees' confidence. However, in the long run, not all changes were sustained in teachers' classroom practices [5]. Another research study from Canada showed that certain aspects of pronunciation teaching tended to be avoided, such as the transition from more to less controlled pronunciation practice [6].

The aim of the current study was to investigate how teacher training received at the beginning of student teachers' education paths influenced their prospective teaching careers. The respondents were primary English teachers who successfully completed four semesters of English phonetics and phonology at the Faculty of Education in Prague, including half a semester dedicated to pronunciation peer-teaching [7]. The pronunciation activities conducted in front of their peers followed an identical structure. The obligatory parts included a lead-in, where the topic had to be introduced interactively with pupils' active involvement, and pronunciation practice, which could be oriented either perceptually, productively, or both. The submission of a detailed activity plan was required. The peer-teaching sessions were recorded in the spring of 2014 and were scrutinized and analysed ten years later. Seven participants agreed to watch and critically reflect on their own micro-teachings. Through stimulated recall interviews, insights into the teachers' rationale, feelings underlying their actions, and suggestions for improvement were captured.

The preliminary findings suggest that although the respondents consider pronunciation teaching at the primary level important and attempt to integrate it into their classes, they face constraints on the amount of time available for it. Their reflections primarily emphasized general didactic principles, such as the importance of clear instructions, appropriate timing or language use, rather than focusing on specific features of the English sound system. Further interviews are planned to gain more data.

References

[1] Henderson, A., Frost, D., Tergujeff, E., Kautzsch, A., Murphy, D., Kirkova-Naskova, A., Waniek-Klimczak, E., Levey, D., Cunnigham, U., & Curnick, L. (2012). The English Pronunciation Teaching in Europe Survey: Selected Results. *Research in Language*, *10*(1), 5–27. https://doi.org/10.2478/v10015-011-0047-4

[2] Baker, A. A. (2014). Exploring teachers' knowledge of L2 pronunciation techniques: Teacher cognitions, observed classroom practices and student perceptions. *TESOL Quarterly*, 48(1), 136–163. doi: 10.1002/tesq.99

[3] Murphy, J. (2014). Teacher training programs provide adequate preparation in how to teach pronunciation. In L. Grant (Ed.), *Pronunciation myths: Applying second language research to classroom teaching* (pp. 188–224). Ann Arbor, MI: The University of Michigan Press.

[4] Darcy, I. (2018). Powerful and Effective Pronunciation Instruction: How Can We Achieve It? *The CATESOL Journal, 30*(1), 13–45.

[5] Burri, M. & Baker. A. (2021). 'I Feel… Slightly out of Touch': a Longitudinal Study of Teachers Learning to Teach English Pronunciation over a Six-Year Period. *Applied Linguistics, 42*(4), 791–809.

[6] Buss, L. (2017). The role of training in shaping pre-service teacher cognition related to L2 pronunciation. Ilha Do Desterro, *70*(3), 201–226. https://doi.org/10.5007/2175- 8026.2017v70n3p201

[7] Červinková Poesová, K. (2023). Peer pronunciation teaching: Initial training of Czech pre-service primary teachers. In A. Henderson & A. Kirkova-Naskova (Eds.), Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices (pp. 23–30). Université Grenoble-Alpes. https://doi.org/10.5281/zenodo.8137816

Timbre quality of vowel-like hesitation as function of foreign language anxiety

Zsuzsa Szánthó[a]

[a]*ELTE Eötvös Loránd University, Hungary*

Keywords* — *foreign language anxiety, second language, filled pause, vowel-like sound, timbre

I. INTRODUCTION

The characteristics of spontaneous speech may be influenced by several factors, including the emotional state of the speaker [1], e.g., anxiety [2], together with a special form of anxiety, foreign language anxiety (FLA) [3]. The realization of the vowel-like hesitations, and thus their timbre, differ between languages: in Spanish, the most common vowel-like hesitation is [e]-like sound [4], while in Hungarian it is [ə]-like sound [5]. In Spanish (L2) utterances of speakers showing FLA, compared to speakers not showing FLA, [ə]-like hesitations are more frequent [6], whereas the [e]-like hesitations are absent [7]. The aim of this pilot study is to analyze vowel-like hesitations in Hungarian native speakers' Hungarian and Spanish (L2) utterances. The question is how the timbre of the [e]- and [ə]-like hesitations in Spanish (L2) varies depending on whether the speaker is showing FLA. A further question is to what extent these hesitations in Spanish (L2) utterances differ in timbre from those in Hungarian (L1) and from reference data of [e]-like hesitations of native Spanish speakers [8]. Furthermore, if, unlike speakers showing FLA, speakers not showing FLA hesitate with [e]-like sound in Spanish (L2) [7], are their [ə]-like hesitations in Spanish more [e]-like? Our hypotheses are that (i) in Spanish (L2) utterances, the [ə]-like hesitations of speakers not showing FLA and speakers showing FLA differ in that the [ə]-like hesitations of the former's are closer in timbre to the [e]-like hesitations, while the speakers showing FLA pronounce [ə]-like hesitations that are more different in timbre from the [e]-like hesitations; (ii) in Hungarian (L1) and Spanish (L2) utterances, the [ə]-like hesitations are more different in timbre for speakers not showing FLA than for speakers showing FLA; (iii) in Spanish (L2) utterances, speakers showing FLA pronounce more vowel-like hesitations than speakers not showing FLA, and the latter's vowel-like hesitations are closer in timbre to [e]-like hesitations than the vowel-like hesitations of speakers showing FLA.

II. METHODS

The study, based on 36 audio recordings in Hungarian (L1) and Spanish (L2), included 9-9 women showing and not showing FLA. Informants were selected and grouped using an online questionnaire about their age, sex, level of language proficiency in Spanish, their possession of exam certificate and its level according to the Common European Framework of Reference (CEFR). In the questionnaire their level of FLA was also estimated through a test adapted from [3]. The speakers' Spanish language proficiency was level B2 or higher (CEFR) and although not all of them lived in Spanish-speaking territories, all of them reported to use Spanish daily. In their spontaneous speech in Hungarian (L1) and in Spanish (L2), the timbre of the vowel-like hesitations, and its two forms, i.e., [e]- and [ə]-like hesitations were analyzed by measuring the frequency values of the first and second formants. The vowel-like hesitations were categorized perceptually. A total of 582 vowel-like hesitations were analyzed. The measured frequency values were also compared with reference data [8] of the [e]-like hesitations of native Spanish speakers.

III. RESULTS

The number of elements of each form of vowel-like hesitations was uneven both in groups and in languages (TABLE 1). In Spanish, only speakers not showing FLA pronounced the [e]-like hesitation. However, vowel-like hesitations were the most frequent in the Spanish utterances of speakers showing FLA. According to the qualitative analysis, in Spanish (L2), there was a difference in the formant values of the [ə]-like hesitations in function of groups. The [ə]-like hesitations were more open acoustically in both speakers showing and speakers not showing FLA in Spanish (L2) than in Hungarian (L1), but, in addition, in speakers not showing FLA it was more palatal acoustically as represented in area of the vowel space (Fig. 1). Furthermore, the [e]-like hesitations in Spanish (L2) were closer in timbre to the reference data of Hungarian (L1) [ɛ] sound than to the one of the Spanish (L1) [e]-like hesitation or [e] sound in area of the vowel space. Statistical analysis showed that language had a significant effect on F_1 ($F(1, 18) = 9.6$; $p < 0.05$), while different forms of hesitation and FLA did not. Accordingly, vowel-like hesitations were acoustically more open in Spanish (L2) than in Hungarian (L1). Further analysis showed that language and anxiety interacted to affect F_2 ($F(1,18) = 7.4$; $p < 0.05$), while only the language main effect was significant ($F(1, 18) = 7.9$; $p < 0.05$). Due to the unbalanced number of elements, the data was not allowed to be reliably subjected to statistical analysis.

TABLE I. NUMBER AND MEAN FREQUENCY (COUNT PER SYLLABLE) OF VOWEL-LIKE HESITATIONS

		[ə]-like		**[e]-like**		**vowel-like**	
language	**FLA**	**number**	**frequency**	**number**	**frequency**	**number**	**frequency**
Spanish	not showing FLA	63	0,009	63	0,009	126	0,017
	showing FLA	300	0,056	0	0	300	0,056
Hungarian	not showing FLA	158	0,023	0	0	158	0,023

	showing FLA	111	0,014	0	0	111	0,014
total		632	-	63	-	695	-

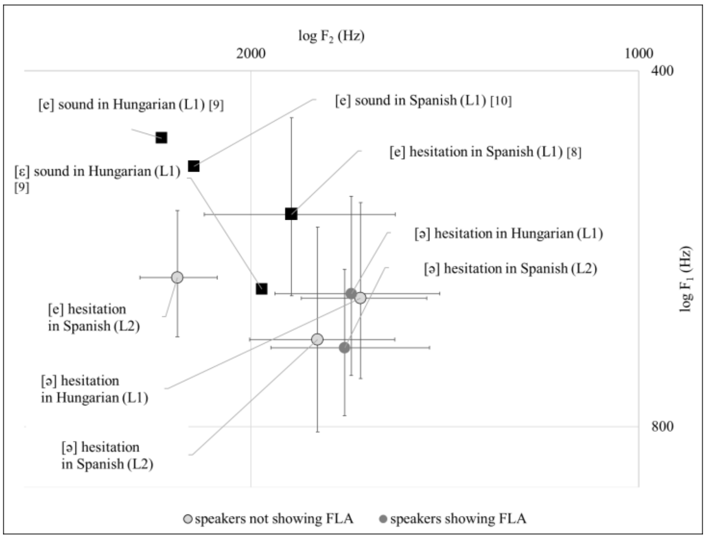


Fig. 1. Timbre of vowel-like hesitations and reference data (Hz)

IV. Conclusions

The results of the research corroborated our first hypothesis: the [ə]-like hesitations in Spanish (L2) utterances of speakers not showing FLA and speakers showing FLA differed in that the [ə]-like hesitations of speakers not showing FLA were closer in timbre to the [e]-like hesitations of native Spanish speakers [8] than those of speakers showing FLA. In contrast, the results did not corroborate out second hypothesis: the [ə]-like hesitations in Hungarian (L1) and Spanish (L2) utterances did not differ in timbre more for speakers not showing FLA than for speakers showing FLA. However, it is important to point out that the differences in the two groups were of a different nature: in Spanish (L2), the [ə]-like hesitations of speakers not showing FLA were not only more open acoustically, but also more palatal. The results also corroborated our third hypothesis: in Spanish (L2) utterances, speakers showing FLA not only pronounced more vowel-like hesitations than speakers not showing FLA, but the vowel-like hesitations of speakers not showing FLA were also closer in timbre to [e]-like hesitations than the vowel-like hesitations of speakers showing FLA. In summary, in Spanish (L2) utterances, the vowel-like hesitations of speakers not showing FLA are less “[ə]-like”, with a timbre closer to native Spanish speakers’ reference data of [e]-like hesitation than the vowel-like hesitations of speakers showing FLA. However, it is important to point out that when examined in Hungarian (L1) area of the vowel space, the [e]-like hesitations pronounced in Spanish (L2) are closest to the Hungarian (L1) [ɛ] sound.

References

[1] J.-A. Bachorowski and M. J. Owren, “Vocal Expressions of Emotion,” in *Handbook of Emotions,* H.-J. M. Lewis and J. M. L. Feldman Barrett, Eds. New York: The Guilford Press, 2008, pp. 196–210.

[2] M. Gósy, *Pszicholingvisztika*. Budapest: Osiris Kiadó, 2005.

[3] P. D. MacIntyre and R. C. Gardner, “The subtle effects of language anxiety on cognitive processing in the second language,”. *Language Learning*, vol. 44, pp. 283–305, 1994.

[4] M. J. Machuca and J. Llisterri and A. Ríos, “Las pausas sonoras y los alargamientos en español: Un estudio preliminar,” *Normas*, vol. 5, pp. 81–96, 2015.

[5] M. Gósy, “A semleges magánhangzó: a funkció, a kiejtés és az akusztikum összefüggései,” *Magyar Tudomány, Beszéd és beszédtudomány*, 2007. http://www.matud.iif.hu/07jan/13.html (Downloaded 17/11/2023)

[6] Zs. Szánthó, “Foreign language anxiety and filled pauses in spontaneous L2 speech,” *Proc. 3rd International Symposium on Applied Phonetics (ISAPh 2021)*, pp. 67–70, 2021.

[7] Zs. Szánthó, “Többet "őzünk", ha szorongunk?” in *Bonsai-tanulmányok,* P. Balázs-Piri and L. Miklós, Eds. Budapest: ELTE Eötvös Kiadó, 2022, pp. 91–100.

[8] M. J. Machuca, “An acoustic study on the use of fillers in Spanish as a foreign language acquisition”. *Second Language Acquisition. Learning Theories and Recent Approaches,* 2022 (in publication)

[9] A. Deme and T. E. Gráczi and V. Horváth and A. Markó, “Magánhangzó-realizációk az olvasásban és a spontán beszédben,” (conference presentation), *Beszédkutatás 2011. konferencia*. Budapest: MTA Nyelvtudományi Intézet, 27–28/10/2011.

[10] A. Quilis and M. Esgueva, „Realización de los fonemas vocálicos españoles en posición fonética normal,” In *Estudios de fonética I,* M. Esgueva and M. Cantarero, Eds. Madrid: Consejo Superior de Investigaciones Científicas, 1983, pp. 137–252.

Acknowledgements

I thank Andrea Deme for her tireless help and support. The research was supported by the TKA–DAAD grant No. 177375 and the ÚNKP-23-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

Quantifying sociolinguistic change: Effects of age and gender on dialectal variation in a large corpus of spontaneous Finnish speech

Tuukka Törö, Antti Suni, Juraj Šimko
University of Helsinki, Finland,

Keywords — sociophonetics, linear discriminant analysis, self-supervised learning, XLS-R

I. Introduction

Socio-economic group identities are indexed through all modes of verbal communication, and sociolinguistic variables are intermingled. Same forms may index various identities or meanings depending on both the social context as well as what other forms are present [1]. Studying such variability has its challenges such as finding representative speakers, creating a setting that elicits 'natural' speech and controlling variables stemming from the backgrounds and interpersonal dynamics between interlocutors. Self-supervised large-scale models trained on audio only data can remedy some of this: we can extract embeddings from a large audio dataset without the need for transcriptions and filter out variation that is not of interest for a specific question. The embedding spaces of these models have been shown to include various kinds of information about the audio signal regarding prosody and local phonetic features [2].

We present an approach for analyzing sociolinguistic variables in a large corpus of colloquial speech using latent space embeddings of a self-supervised speech model. Our aim was to investigate if our methodology corroborates previously posited claims (and rebuttals) of Finnish dialectal differences diminishing [3], and of sociolinguistic change being driven by young females [4]. Using dimensionality reduction and clustering techniques, we investigate sociolinguistic variables, not only in insulation but in how they interact. Our results show promise in quantifying sociolinguistic change in 'wild' data, complementing established methodology.

II. Data and Method

Lahjoita Puhetta (Donate Speech) [5] is a large corpus of spontaneous colloquial Finnish speech gathered online for AI research. It consists of ~3,200 hours of self-recorded speech from over 20,000 speakers including information about the speaker's age group, gender, and a subjective judgement of the dialect they speak. We used a balanced subset of the corpus with male and female speakers aged 21-40, 41-60 and 61-100 from 16 dialect regions resulting in 108 speakers per dialect.

We divided every speaker's recordings into utterances at silence intervals and extracted 2048-dimensional embeddings for every speaker from a pretrained XLS-R model fine-tuned for language identification [6]. We used the original embeddings as well as linear discriminant analysis (LDA) to investigate the space. LDA utilizes class labels to transform the original space and yields a projection that minimizes within-class variance and maximizes between-class variance [7], filtering out information that does not contribute to variance between the given classes. We transformed the original space with LDA using dialects as classes.

To examine interclass relationships in the latent spaces, we applied a hierarchical agglomerative clustering algorithm using mean vectors for each class. We also examined the within-dialect and between-dialect variance across age and gender categories to quantify sociolinguistic change.

III. Results and Discussion

Clustering the dialects on the original XLS-R embeddings largely followed geographical relationships but the Southeastern dialects were clustered together with Southwestern dialects which goes against prior knowledge of Finnish dialect groups [8]. Transforming the space with LDA to three components showed clusters (Figure 1.) that perfectly follow geographical relationships.

Quantifying variance on the LDA embeddings (see Figure 2.) points towards younger females having a more uniform way of speaking and on the other end, old males having most distinct dialectal differences.

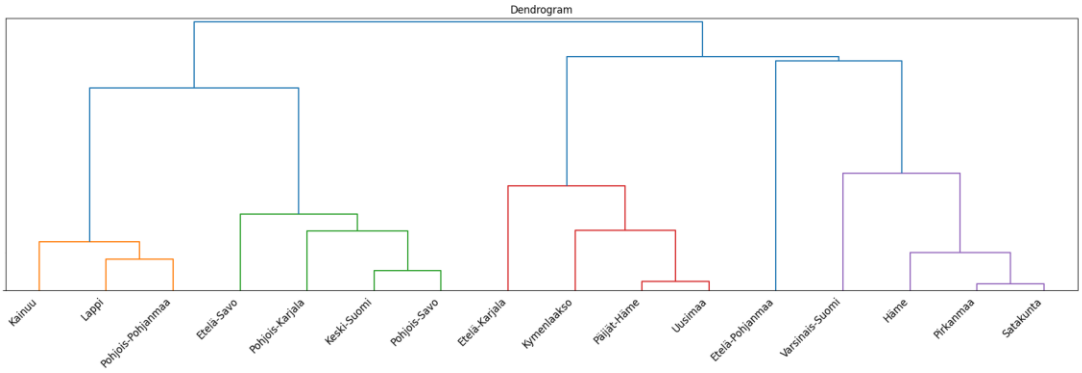


Fig 1. Dialects clustered with 3 LDA components shows clear groupings on various levels that reflect geography (log distance).

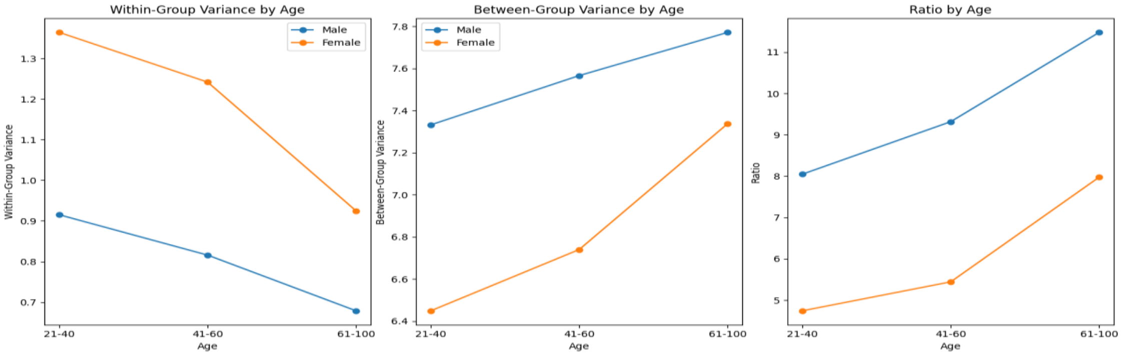


Fig 2. Within-dialect and between-dialect variances and their ratio by age and gender with three dimensional LDA embeddings using dialect labels

Our results show that the self-supervised model's latent space contains information about dialect, gender and age and that using LDA can reveal information about how gender and age influence dialectal change on a large scale. There seems to be a tendency for younger and female speakers having a more ambiguous distinction between dialects and older and male speakers showing a clearer separation. This is in line with earlier claims of dialects converging and of young females being at the forefront of sociolinguistic change. However, the model lacks interpretability and while its training data is audio only – and we averaged over multiple utterances and speakers – it is possible that there are traces of distinctive lexical information left on the embeddings.

The current method shows promise for efficiently quantifying sociolinguistic change, for example, in low-resource languages where transcribed data might not be available. Utilizing LDA could also help when it is not possible to acquire data from several speakers of different genders or ages, as it can filter out, for example, gender differences in the space.

When analyzing large datasets of noisy data with a model whose internal workings are opaque, there are any number of underlying variables that may affect our findings. In the future, investigating correlation between the XLS-R embeddings and acoustic and textual information will help us investigate what linguistic-phonetic sources of variation that reflect sociolinguistic change are present in the speaker embeddings.

References

[1] L. Hall-Lew, E. Moore, and R. J. Podesva, Social meaning and linguistic variation: theorizing the third wave. Cambridge University Press, 2021.

[2] G. -T. Lin et al., "On the Utility of Self-Supervised Models for Prosody-Related Tasks," 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023, pp. 1104-1111, doi: 10.1109/SLT54892.2023.10023234.

[3] P. Nuolijärvi and M. Sorjonen, "Miten kuvata muutosta?: puhutun kielen tutkimuksen lähtökohtia murteenseuruuhankkeen pohjalta. Kotimaisten kielten tutkimuskeskus." 2005.

[4] W. Labov, "Driving forces in linguistic change," in Proceedings of the 2002 International Conference on Korean Linguistics, August 2002, pp. 1-24.

[5] A. Moisio, D. Porjazovski, A. Rouhe, Y. Getman, A. Virkkunen, R. AlGhezi, et al., "Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks," Language Resources and Evaluation, vol. 57, no. 3, pp. 1295-1327, 2023.

[6] J. Valk and T. Alumäe, "VoxLingua107: a Dataset for Spoken Language Recognition," in Proc. IEEE SLT Workshop, 2021.

[7] P. Xanthopoulos et al., "Linear discriminant analysis," in Robust Data Mining, 2013, pp. 27-33.

[8] Institute of the Languages of Finland (Kotus), "Suomen murteet," n.d. [Online]. Available: https://www.kotus.fi/kielitieto/murteet/suomen_murteet. [Accessed: Apr 7, 2024].

Phonetic Interference of L3 Japanese on L2 English Word-Initial Stop Production in Mandarin Trilinguals

Min Zeng
Waseda University, Japan

Keywords — composite L2-L3 category, L2 production, multilingualism, voice onset time

I. Introduction

While word-initial stops in Mandarin, English and Japanese are primarily cued by voice onset time (VOT) [1, p. 60], [2], [3, p. EL96], word-initial voiceless stops in these languages differ in lag patterns. Word-initial stops can be classified into three categories: voicing lead (−125 to −75 ms), short lag (0 to 25 ms), and long lag (60 to 100 ms) [4, p. 403]. Mandarin [p^h, t^h, k^h] and English /p, t, k/ fall under the long lag category, and are produced by monolinguals with VOT values centering at 100 ms in Mandarin and 80 ms in English [5, p. 20]. In contrast, Japanese /p, t, k/, weakly aspirated, are produced by monolinguals with VOT values that fall between those of prototypical short lag and long lag stops [6, p. 75]. We observe a hierarchy of these voiceless stops: Mandarin [p^h, t^h, k^h] exhibit longer VOT values than those of English /p, t, k/, which in turn exhibit longer VOT values than those of Japanese /p, t, k/.

Previous studies on the production of word-initial stops by native Mandarin-speaking bilinguals and trilinguals have revealed two important findings. First, despite experiencing native language (L1) transfer, Mandarin bilinguals, even those with low proficiency in their second language (L2) English, appear to produce English voiceless stops with VOT values akin to those of native English speakers [7, p. 560]. Second, Mandarin trilinguals experience cross-linguistic phonetic interactions [8, p. 101]. To the best of our knowledge, no study has focused on the acquisition of L2 stops by Mandarin trilinguals who are advanced in both their L2 and L3. This study aims to fill this gap. Our research question is, how do Mandarin trilinguals, who are advanced in both their L2 English and L3 Japanese, produce English word-initial voiceless stops? We hypothesize that the performance of the Mandarin trilinguals will reflect the phonetic interference from both their L1 Mandarin and the later acquired L3 Japanese.

II. Method

Thirty-one Mandarin trilinguals, 34 Mandarin bilinguals and 34 Japanese bilinguals formed the MT, MB and JB groups, respectively, participating in the production experiment. The MT group, consisting of international students at a Japanese university, had an average 3.69-year residency in Japan. Their average age of acquisition was 6.9 years for English and 18.39 years for Japanese. Regarding language proficiency, they scored higher than 85 on a TOEFL iBT test (CEFR B2-C1 level) and passed the Japanese Language Proficiency Test N1 level. The MB and JB groups, university students from Shanghai and Tokyo, respectively, both had L2 English levels ranging from beginner to intermediate (CEFR A1-B2 level). The production material included nine English words (i.e., *panda*, *Paris*, *parrot*, *taxi*, *tablet*, *tango*, *candy*, *camel*, and *carrot*), where the target stops /p, t, k/ occur at the onsets of their first syllables. The participants produced these words using a carrier sentence “The target word is ___.”

In total, we collected 2673 valid productions (99 participants * 9 words * 3 repetitions). Using Praat to segment the VOT values of the target stops, we selected the zero-crossing points and relied primarily on the waveforms and secondarily on the vowel formants to pinpoint the boundary between the stop and its following vowel. One linear mixed model (LMM) was applied to these VOT values (the dependent variable) in R. The fixed factor of the LMM was Group (three levels: MT, MB, and JB). The random intercepts were Participant and Stimulus. We assessed the main effects of the fixed factor and performed the post-hoc comparison of contrasts.

III. Results

The descriptive statistics are illustrated with boxplots (Fig. 1). The average VOT values of each English stop produced by the MT, JB, and MB groups are as follows: /p/-64.0, /t/-65.3, /k/-74.2 ms by the MT group; /p/-49.5, /t/-52.3, /k/-68.1 ms by the JB group; /p/-75.0, /t/-75.2, /k/-83.4 ms by the MB group. Regarding statistical analysis, the LMM results indicated significant main effects on Group (χ^2 (2) = 35.7, $p < 0.001$), with a large effect size ($\eta^2 = 0.17$). The post-hoc analysis, as detailed in Table 1, revealed a nuanced hierarchy: the Mandarin bilinguals produced L2 English word-initial voiceless stops with significantly longer VOT values than those of the Mandarin trilinguals, which in turn were longer than those of the Japanese bilinguals.

IV. Discussion and conclusion

First, the VOT values of the English /p, t, k/ produced by the Mandarin trilinguals were significantly longer than those of the Japanese bilinguals, suggesting that the trilinguals were influenced by the phonetic system of their L1 Mandarin. Native Mandarin speakers produce the Mandarin [p^h, t^h, k^h] with significantly longer VOT values ([p^h]-104.9, [t^h]-104.4, [k^h]-103.1 ms) [9, p. 772] than those of the Japanese /p, t, k/ produced by native Japanese speakers (/p/-22, /t/-28, /k/-47 ms) [10, p. 70]. Carrying over the Mandarin production patterns into the production of L2 English stops, the Mandarin trilinguals exhibited significantly longer VOT values compared to those of the Japanese bilinguals.

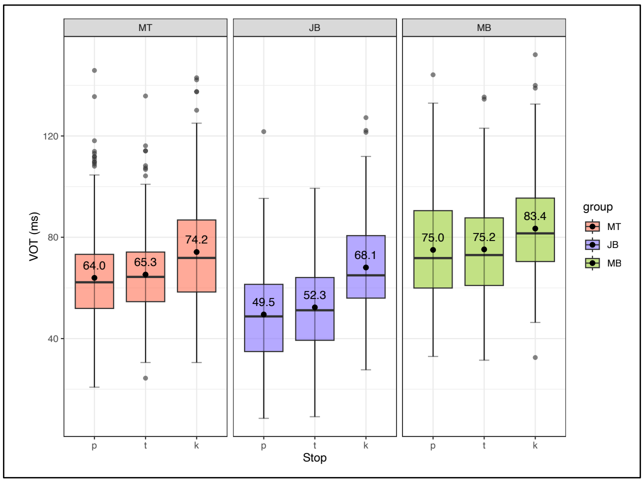


Fig. 1. VOT distributions of the English stops produced by the MT, JB, and MB groups.

TABLE I. SUMMARY OF THE RESULTS OF THE POST-HOC TEST COMPARING THE MT, JB, AND MB GROUPS

Group	**Estimate**	***SE***	***df***	***t* ratio**	***p* value**
MT vs. JB	0.53	0.16	96	3.14	0.0067*
MT vs. MB	−0.45	0.16	96	−2.69	0.0249*
JB vs. MB	−0.98	0.16	96	−5.97	< 0.001*

Notably, the Mandarin trilinguals produced the English /p, t, k/ with significantly shorter VOT values than the Mandarin bilinguals, indicating that the trilinguals experienced phonetic interference from their L3 Japanese. Based on the category formation process proposed by the SLM-r [11, pp. 40-41], we speculate that the trilinguals, with extensive exposure to Japanese, might have discovered the phonetic differences between the voiceless stops in L3 Japanese and L1 Mandarin. Specifically, as the trilinguals modified their realization rules to attune to the new L3 stops they had heard and seen in meaningful conversations, they might have noticed that these stops carry phonetic features, such as the articulatory features of moderately aspirated voiceless stops, resembling their corresponding stops in L2 English, which are markedly different from those in L1 Mandarin. Consequently, the trilinguals experienced the L2 English voiceless stops shifting toward the corresponding stops in L3 Japanese, establishing a composite L2-L3 articulatory category that maintained phonetic contrast with the articulatory category for stops in L1 Mandarin. Therefore, the VOT values of the English voiceless stops produced by the Mandarin trilinguals, 'compromise' VOT values carrying the articulatory features of stops in Japanese, were significantly shorter than those produced by the Mandarin bilinguals.

In summary, the results supported our hypothesis: the performance of the Mandarin trilinguals in L2 English revealed that they experienced phonetic interference from their L1 Mandarin and, more importantly, from their later-acquired L3 Japanese.

REFERENCES

[1] H.-G. Byun, "Acoustic characteristics for Japanese stops in word-initial position: VOT and post-stop fo," *Journal of the Phonetic Society of Japan*, vol. 25, pp. 41–63, 2021.

[2] R. Y.-H. Lo, "The dual role of post-stop fundamental frequency in the production and perception of stops in Mandarin-English bilinguals," *Frontiers in Communication*, vol. 7, p. 864 127, 2022.

[3] A. A. Shultz, A. L. Francis, and F. Llanos, "Differential cue weighting in perception and production of consonant voicing," *The Journal of the Acoustical Society of America*, vol. 132, no. 2, EL95–EL101, 2012.

[4] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[5] J. Yang, "Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults," *Second Language Research*, vol. 37, no. 1, pp. 3–26, 2021.

[6] T. J. Vance, *The sounds of Japanese*. Cambridge University Press, 2008.

[7] L.-m. Chen, K.-Y. Chao, J.-F. Peng, and J.-C. Yang, "A cross-language study of stop aspiration: English and Mandarin Chinese," in *2008 Tenth IEEE International Symposium on Multimedia*, 2008, pp. 556–561.

[8] Y. Sun and S. Profita, "Cross-linguistic study on VOT of Chinese trilingual speakers," *Studies in Literature and Language*, vol. 20, no. 1, pp. 98–102, 2020.

[9] Q. Feng and M. G. Busà, "Acquiring Italian stop consonants: A challenge for Mandarin Chinese-speaking learners," *Second Language Research*, vol. 39, no. 3, pp. 759–783, 2023.

[10] K. Shimizu, " 閉鎖子音の有声性・無声性の音声的特徴に関する考察 [A study on phonetic characteristics of voicing contrasts of stop consonants]," *Journal of the Phonetic Society of Japan*, vol. 22, no. 2, pp. 69–80, 2018.

[11] J. E. Flege and O.-S. Bohn, "The revised speech learning model (SLM-r)," in *Second language speech learning: Theoretical and empirical progress*, R. Wayland, Ed., Cambridge University Press, 2021, ch. 1, pp. 3–83.