LUND
UNIVERSITY

Book of Abstracts

4th International Symposium on Applied Phonetics (ISAPh 2022)

September 14-16 2022

Lund University, Sweden

Sponsored by:

- Riksbankens Jubileumsfond
- Stiftelsen Elisabeth Rausings Minnesfond
- Letterstedtska Föreningen
- Birgit Rausing Language Programme

Index

Invited speakers

Oral sessions

Postersessions

Forensic speech science needs forensic phoneticians

Paul Folkes, Department of Language and Linguistic Science, University of York

KEYNOTE ABSTRACT

Forensic speech science (FSS) is a rapidly evolving field. The most prominent goal of FSS is to establish how individual voices differ from one another. This is of vital importance in forensic speaker comparison cases, which typically require analysis of a voice recorded during a crime (e.g. bomb threats) and comparison with the voice of a suspect (usually recorded in police custody). Comparison of voices is far from straightforward, as anyone trained in phonetics will understand: the voice is highly plastic, affected by many factors relating to the individual (e.g. speech style, health, emotion) and the environment (e.g. background noise, telephone transmission). In short, no two speech samples are ever identical. Voice is therefore an imperfect biometric; voices are certainly distinctive and act as a marker of identity, but, unlike DNA or fingerprints, there is nothing in a voice that is indelible or immutable.

One consequence of this fact is that there is no single method for voice analysis that is universally adopted. Methods of conducting voice analysis have evolved via two largely separate traditions. In Europe FSS is largely grounded on phonetics. Standard analytic methods such as vowel formant and f0 analysis are applied to forensic recordings. In the US, by contrast, the approach is largely informed by engineering and computer science, from which have developed automatic speaker recognition (ASR) systems (similar technology is used for non-forensic purposes e.g. in speech operated tools such as Alexa). ASR technology is making rapid advances, as is evident from its increasing presence in our everyday lives. In controlled experiments ASR systems can now attain almost perfect performance in classifying voice samples by speaker.

However, ASR is not a perfect solution to FSS problems. In my view it never will be. There remain a number of difficulties in using ASR in forensic cases. These include:

- Applicability. ASR is driven by commercial interests, not forensic ones. Even the best systems struggle to handle non-standard speakers and speech, especially the kinds of materials typically available in forensic cases. These are often of poor technical quality, involve speakers who are stressed or emotional, and there may be considerable variability within a speech sample. Recordings may also be very short. These would therefore likely be rejected as unsafe for analysis in ASR systems.
- Transparency. ASR systems remain 'black boxes' – exceptional in performance, but with no one quite sure why. That is, it is not clear how the features extracted via ASR systems map onto the concrete vocal and linguistic features understood by phoneticians. What is it about the voice that is being picked out and classified as similar and unusual? This is an issue of critical importance in the delivery of justice: it is a principle that evidence in a legal case must be fully transparent.
- Evaluation. A key step in forensic voice analysis, both by ASR and phoneticians, is that we must judge the typicality of the observed features against a background population to assess whether the voice is unusual or unremarkable. This background population is defined by the characteristics of the offender, i.e. the person whose identity is in question. We therefore face an inevitable paradox: if the speaker's identity is unknown, so is the population from which he or she is drawn. But ASR systems tend to work with populations defined simply by place and language (e.g. 'Swedish', 'UK English').

1.

The two separate traditions of FSS have so far made only limited moves towards integration. In this talk I hope to show why that integration is beneficial. Drawing on real case materials and empirical research I will argue that phonetic methods offer an essential complement to ASR systems to address these three issues.

- Applicability. Phonetic methods can be used profitably with short or poor quality materials, and to identify the sociolinguistic factors that need to be addressed to better train ASR systems.
- Transparency. Phonetics can help clarify how ASR results map to vocal features. Ongoing research explores the relationship between (i) ASR features and phonetic features, and (ii) relative performance of ASR and phonetic analysis on the same materials.
- Evaluation. Phonetic analysis is crucial in establishing the background population, including uncertainty over how to delimit it.

In summary, I aim to show that, in an era of tremendous technical advance, forensic speech science still needs forensic phoneticians.

Developing prosody in typical and atypical language acquisition

Sónia Frota, School of Arts and Humanities and Center of Linguistics, University of Lisbon

KEYNOTE ABSTRACT

Infants' early sensitivity to the prosodic properties of speech is well documented, and has supported the view that infants are equipped with an input processing mechanism initially tuned to prosodic information. In addition, prosody has been suggested to bootstrap the learning of language. Although the precocious sensitivity to prosody and its potential to facilitate language acquisition seem quite general, recent research has suggested that the early development of prosody appears to be crucially shaped by language experience. Moreover, if infants' perception of prosody is guided by language experience, it is fundamental to determine which and how prosodic patterns/cues are attended to early on in development, and may thus provide useful information to scaffold language learning. Typically and atypically developing infants may vary in their language experience, and it is largely unknown whether the early development of prosody differs in these populations. Crucially, the potential of prosody to facilitate language learning in atypical development is still to be determined.

Infants may utilize the prosodic property of stress to begin developing the ability to segment the speech signal into words and phrases, and for word categorization. Intonation patterns, in turn, usually convey phrase level meanings, while also contributing to speech chunking by signaling prosodic boundaries. I will present findings from a series of speech perception and word segmentation experiments using eye gaze paradigms, focusing on the perception of stress and pitch patterns. The speech perception and word segmentation abilities of monolingual European Portuguese-learning infants with no known risk for language impairments (the typically developing group, TD) are examined and compared to those of infants and toddlers at-risk for language impairments (namely, preterm birth and familial risk for autism or language disorder, the AR group), as well as to infants and toddlers with Down Syndrome (the DS group). The results suggest different developmental paths for early word segmentation across groups. The relation between the perception of stress and pitch patterns, and emerging word segmentation abilities is explored to further our understanding of the role of prosody in typical and atypical language acquisition, with clinical implications for remediation and intervention strategies.

Revisiting second language pronunciation teaching and assessment: Constructs, compatibilities, contradictions, cross-fertilization

Talia Isaacs, IOE—UCL's Faculty of Education and Society, University College London, UK

Keynote Abstract

Second language (L2) English pronunciation teaching and assessment research has undergone major shifts over the past few decades. Pronunciation instructional goals and assessment targets have been revamped and rebranded, resulting in growing prominence and uptake in applied research and classroom and assessment settings. However, with accuracy- and intelligibility-focused constructs persisting, co-existing, and, in some cases, conflated, there are some unresolvable tensions and orientations that it is worthwhile underscoring to raise awareness and as a small step toward fostering respect for, or at least tolerance of, different research traditions within linguistics and applied linguistics and breaking disciplinary silos ([1]).

In the first part of this talk, I will argue for why considerations of construct validity in both low-stakes L2 pronunciation research contexts, and high-stakes L2 speaking/listening assessment settings should be fundamental for all pronunciation research. I will discuss sources of congruence and incongruence between key global constructs that are often used in pronunciation research, also in relation to defining appropriate standards and benchmarks for L2 teaching and assessment. Technological innovations and limitations and how that has shaped the focal construct by default, as compared to human-mediated and human-scored assessments, will also be discussed. In the second part of the talk, from the perspective of a researcher actively collaborating with colleagues researching methodological factors in health intervention research, I will reflect on what the field of L2 pronunciation can learn from conventions and practices in evidence-based medicine to chart bold new research directions. To do this, I will draw on relevant parts of Chalmers and Glasziou's categories ([2]) of avoidable "research waste" in biomedical research as a springboard for discussing ways to promote greater research efficiency and quality in applied L2 pronunciation research. The primary focus is studies designed to investigate which instructional approaches work best and the nature and strength of the existing evidence. I will argue for the need for greater engagement with educational stakeholders throughout the research cycle, including by establishing a stakeholder-relevant applied L2 pronunciation research agenda.

References

[1] T. Isaacs, and H. Rose, "Redressing the balance in the native speaker debate: Assessment standards, standard language, and exposing double standards", TESOL Quarterly, 56(1), 2021, pp. 401-41. https://doi.org/https://doi.org/10.1002/tesq.3041

[2] I. Chalmers and P. Glasziou, P. (2009). "Avoidable waste in the production and reporting of research evidence", The Lancet, 374(9683), 2009, pp. 86-89.

4th International Symposium on Applied Phonetics (ISAPh2022), September 14-16 2022, Lund University, Sweden

Creating pronunciation training content for your language of interest – A hands-on workshop

Jacques Koreman, Department of Language and Literature, Norwegian University of Science and Technology (NTNU)

Keynote Abstract

A. *WHY?*

A foreign accent can lead to lower intelligibility or comprehensibility [1]. This can have negative consequences in communication [2]. Nevertheless, pronunciation is often a neglected area in language courses [3]. For this reason, we have developed online pronunciation training to complement classroom teaching.

B. *WHAT?*

The Computer-Assisted Listening and Speaking Tutor (CALST) is a multilingual platform for pronunciation training. It is multilingual in two ways:

- it offers pronunciation exercises for several target languages – presently: English, Spanish, Italian, Greek, and Norwegian, with Catalan under development; *Maybe you want to be the expert who provides content for a new language?*
- it tailors exercises to the learner's native language comparing the two languages in L1-L2*map* (500+ languages based on UPSID). *This is done automatically, also for new target languages in CALST.*

The exercises in CALST are simple discrimination and identification (listening) exercises, pronunciation exercises, and spelling exercises. These familiarize learners with speech sounds that do not occur in their native language, and the same exercises also provide training for unfamiliar consonant clusters, word stress patterns, and lexical tones. The comprehensive language content ensures that learners can train all linguistic contrasts that may pose a challenge for any learner.

C. *HOW?*

In the workshop, we will first present CALST and L1-L2*map.* Then a short explanation will be given of the principles which we have adopted to devise language content; examples from CALST exercises will be given to demonstrate them. This will be followed by hands-on application of these principles to *your* language of interest:

- Determine the sound inventory of consonants and vowels & create an exercise.
- Determine allophonic variation & create an exercise.
- Determine variable grapheme-to-phoneme mapping & create an exercise.
- Select a consonant cluster, stress pattern, or lexical tone contrast & create an exercise.
- Participants should bring their laptops. We hope to see you there!

References

[1] M.J. Munro, and T.M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners", Language Learning, 45, 1995, pp. 73-97.

[2] S. Lev-Ari, and B. Keysar, "Why don't we believe non-native speakers? The influence of accent on credibility". Journal of Experimental Social Psychology, 46(6), 2010, pp. 1093–1096.

[3] J.A. Foote, P. Trofimovich, L Collins, L., and F.S. Urzúa, "Pronunciation teaching practices in communicative second language classes". *The Language Learning Journal 44*(2), 2016, pp. 181-196.

Speaker’s comfort or listening effort? – On the interaction of the speaker, the classroom’s sound environment and the students’ learning

Viveka Lyberg Åhlander, Speech Langue Pathology, Åbo Academy

KEYNOTE ABSTRACT

Research, as well as experience, tell us that it is harder to concentrate and learn in a noisy environment. It is also known since long that the teaching occupation is voice demanding and that many consider voice problems an occupational hazard that is an inevitable part of the job. Meanwhile, the classroom is often a noisy space affecting both the speaker’s comfort as well as the listeners’ effort in hearing and understanding. The presentation will cover research performed within the crossdisciplinary research environment of Cognition Comprehension and Learning at Lund University. The focus will be the interaction of the speaker’s speech, voice quality, the listeners’ learning and the effect of the room on students and the teacher.

Computational tools for studying speech prosody

Yi Xu Department of Speech, Hearing and Phonetic Sciences, University College London, UK

KEYNOTE ABSTRACT

This workshop will introduce a set of computational tools for studying speech prosody. ProsodyPro and FormantPro are Praat scripts for systematic analysis of large amount of speech data, and they generate detailed pitch, duration, intensity, formant and voice quality measurements. qTA, qTAtrainer and PENTAtrainer are Praat- or Javascript-based modelling tools for analyzing, modelling and manipulating speech prosody. Both sets of tools can be used for research and teaching purposes .

Intelligibility of Swedish foreign accented words

Åsa Abelin, Gothenburg University & Elisabeth Zetterholm, Linköping University

***Keywords* — map task, intelligibility, pronunciation, L2-speech, listener test**

I. Introduction

An intelligible pronunciation is of importance for successful communication, regardless of the interlocutors, and therefore one of the goals in foreign and second language learning and education. Many second language learners produce the target language with a more or less strong foreign accent [1, 2, 3] concerning both pronunciation, grammar and vocabulary. Mispronunciation can cause misunderstandings and create unwanted attitudes, depending on listeners' judgements on how comprehensible and intelligible the pronunciation is [4].

In this project, a map task game is used for conversation between L1-speakers and L2-speakers of Swedish as well as between two L2-speakers of Swedish with different first languages. A map task game is a method for performing relatively spontaneous interactions, but at the same time the task can be directed towards specific linguistic features and words. It can be used for studies on grammar in spoken language, feedback signals, pronunciation including phonetics and phonology as well as human and computer interaction [5].

There are several previous studies about differences between L1- and L2-listeners' perception of accented speech. There has been a discussion whether it is easier for an L2-speaker to understand another L2-speaker, than to understand an L1-speaker. Some studies found that there was no such difference, and that the properties of the speech itself are a potent factor for determining how L2 speech is perceived, even when the listeners are from different language backgrounds [6]. Furthermore, the perception can be influenced by e. g. the learners' progression and the amount of input of the target language (for an overview, see [7]). There is also a word frequency effect, especially strong for L2-learners, often depending on vocabulary size (or lexical proficiency) in the target language [8].

II. Aim and Research Question

The aim of the study is to find out if foreign accented speech is more, or possibly less or equally, intelligible for L1 Swedish listeners compared to L2 Swedish listeners. The research question is:

Is there a difference between L1 Swedish and L2 Swedish listeners' assessments and interpretations of Swedish foreign accented words?

III. Material and Method

For this study, we used recordings from conversations in a map task game. The participants are Swedish L1- or L2-speakers. The L2-speakers are in early stages of learning Swedish at different levels in the national program Swedish for Immigrants (SFI). Recordings were made in ordinary classrooms at the participant's school using a mobile phone or a Sony voice recorder. All recordings were made with a good quality for both auditory and acoustic analyses. Instructions about the test were given by a researcher or a teacher at SFI.

The recordings were made in pairs and an L2-speaker of Swedish was a guide through a printed route in the map consisting of 27 pictures. Another L2-speaker or a Swedish L1-speaker was the follower who has another map were five pictures were missing and no printed route. This design was made in order to give rise to communication obstacles. So far, seven pairs are recorded in this map task game.

For the present intelligibility study a listener test was constructed. 60 participants, who were either L1-speakers or L2-speakers of Swedish, listened to and interpreted twelve words without any context. The words were extracted from the map task recordings, and they are very frequent Swedish words [9]. The extracted words were produced by the direction giving L2-speaker. The listener test was presented on-line and the listeners could listen to the words as many times as they wanted. They were asked to write down what they heard, whether it was a real word or a non-word in Swedish.

IV. Preliminary Results

The answers in the listener test were analyzed with regard to listeners' errors in relation to types of pronunciation errors, i. e. intelligibility of the L2-pronunciations. An in-depth qualitative analysis of how the L1-Swedish and L2-Swedish listeners interpreted the stimuli words was made. By mispronounced words, we mean words pronounced with non-standard vowel or consonant quality or quantity, according to Swedish phonology.

Some of the mispronounced words were easier to interpret while some were much more difficult and there are differences between L1-Swedish and L2-Swedish listeners. Some words were interpreted as another real, sometimes more frequent, Swedish word, while other words were interpreted as non-words. Both qualitative and quantitative errors were found on segmental levels, e.g. the minimal pair glas/glass (glass/ice-cream) where there is a qualitative as well as a quantitative difference in the pronunciation of the a-vowel. One important feature is the rounding of Swedish vowels which are often difficult to produce for L2-learners of Swedish. In this study this feature is represented by the Swedish words cykel (bicycle) and lök (onion). When analyzing the recordings, it is obvious that these vowels are quite hard to pronounce and also difficult to interpret, especially for L2-listeners in the listener test. Overall, the interpretation of the isolated words showed great intelligibility problems for all listeners. The study shows that an L2 Swedish speaker's intelligibility may be relative to whether the listener is an L1 or L2 speaker of Swedish.

V. Summary

Results show differences between L1-Swedish and L2-Swedish listeners concerning the interpretation of words. It might depend on both mispronunciation of speakers and vocabulary size of listeners. The result of this study gives an indication of the importance to practice both on perception and production, and to incorporate knowledge about phonetics in the teaching of Swedish as a second language.

References

[1] J.E. Flege, M. Munro, and I.R.A. MacKay. "Factors affecting strength of perceived foreign accent in a second language", Journal of the Acoustical Society of America 97(5), 1995, pp. 3125–3134.

[2] A. Moyer, Foreign accent. The phenomenon of non-native speech. Cambridge: Cambridge University Press, 2013.

[3] R. Pérez-Ramón, M. L. García Lecumberri, and M. Cooke, "Foreign accent strength and intelligibility at the segmental level", Speech Communication 137, 2002, pp. 70–76.

[4] S. Lev-Ari and B. Keysar. "Why don't. we believe non-native speakers? The influence of accent on credibility", Journal of Experimental Social Psychology 46, 2010, pp. 1093–1096.

[5] A. Anderson, M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. MacAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert, "The HCRC Map Task Corpus", Language and Speech 34 (4), 1991, pp. 351-366.

[6] M.J. Munro, T.M. Derwing, and S.L. Morton, "The mutual intelligibility of L2 speech", Studies in second language acquisition, Cambridge University Press, 2006.

[7] M.L.G. Lecumberri, M. Cooke, and A. Cutler, "Non-native speech perception in adverse conditions: A review", Speech Communication 52, 2010, pp. 864–886.

[8] K. Diependaele, K. Lemhöfer, and M. Brysbaert, "The word frequency effect in first- and second-language word recognition: A lexical entrenchment account", The quarterly journal of experimental psychology, Vol. 66, No. 5, 2013, pp. 834–863.

[9] Språkbanken, Korp 9, https://spraakbanken.gu.se.

4th International Symposium on Applied Phonetics (ISAPh2022), September 14-16 2022, Lund University, Sweden

Practical phonetics in the 21st century

Michael Ashby[1], Patricia Ashby[2]

[1]University College London (UCL), [2]University of Westminster, UK.

[1]m.ashby@ucl.ac.uk, [2]ashbyp@westminster.ac.uk

Keywords — practical phonetics, performance skills, pedagogy, object-based learning, scientism

I. Introduction

This paper is concerned with the teaching, learning and assessment of practical skill in speech sound production which has formed part of general phonetic training in the British tradition for most of the last century, though it is presently in decline. We re-examine the purposes of production training, suggesting that the true purpose is widely misunderstood. We analyse the practices of this training in relation to recent pedagogical thinking and argue that it is student-centred, multi-sensory, and constitutes an unusual example of object-based learning. We scrutinise the arguments used by opponents of the practical approach and relate these to a prevailing scientism. We revisit classic formulations of the methods and benefits of practical training in the published literature, pointing out that even the most recent of these is now more than 20 years old. We indicate ways in which powerful personal computers and mobile devices can now support a reformulated approach to practical training and encourage a reasonable and beneficial integration of practical and laboratory-based approaches to general phonetics.

II. Practical Training in Outline

A. *Teaching*

Typically, practical phonetics training is extended over one or more complete academic years and would involve a minimum of 20–30 hours' small-group contact-time. Graded incremental training materials lead the student from simple exercises gaining mastery of basic articulatory manoeuvres such as control of glottal state or the controlled generation of friction to the full range of human sound-types as set out in the IPA alphabet and chart. Auditory perception skills and production skills are developed in tandem, and students analyse and critique each other's performances as well as receiving interactive feedback from a teacher.

B. *Learning*

The expectation is that the scheduled class-time will be matched or exceeded by private-study time. Students typically work repeatedly over materials, practising daily, and commonly using a mirror to monitor their articulations. Highly motivated students may devise their own materials (such as flash cards), or (where possible) make randomised recordings of their own productions for later listening and self-testing.

C. *Assessment*

The course-end assessment includes an individual oral examination in which the candidates are required to perform various arbitrarily-selected sound types, pitch sequences, etc., both from transcribed materials and in response to spoken directions given by the examiners. A more advanced examination may include the requirement to apply a cumulative succession of specified parametric modifications to a given starting point, potentially resulting in novel sound-types which are not represented by unitary IPA symbols, and which the candidate may never have previously attempted. The assessment rewards insight, not mere imitation.

III. Pedagogy and Aims

It will be clear that the description above is of teaching and learning which is student-centred and multi-sensory—though of course it was devised long before those terms became buzzwords in pedagogical theory. It is also arguably an example of Object-Based Learning, since the learner's own speech organs—together with the associated control and sensory mechanisms—provide the focus for understanding. It is hardly an exaggeration to say that the learner has a phonetics teaching laboratory in their own mouth.

However, the aims of such a practical course are not what they might first appear. There are few applications in which an ability to produce an unlimited range of speech sounds is a requirement, or even a useful accomplishment, so at first sight practical training might seem pointless. But as one statement of the method makes clear, the real motivation is different: "What is not […] obvious, but is undoubtedly the case, is that the acquisition of these 'practical' skills is by far the best way of acquiring a deep understanding of phonetic theory—of the principles underlying the description and classification of the sounds of speech …" [1; page 2].

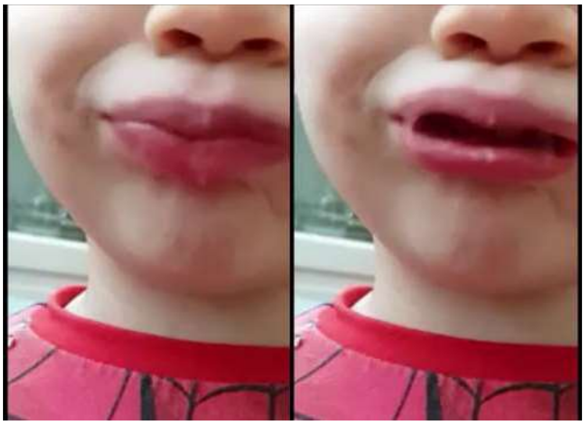
IV. Dissemination and Opposition

Practical training of the kind described has never been universal or widespread. It was chiefly a British practice, associated with figures such as Henry Sweet and Daniel Jones. It was exported in a limited way to the USA, but the chief proponents of it in that context, J. C. Catford and Peter Ladefoged, were themselves British. From the beginning, the legitimacy and value of the practical approach were denied by proponents of "experimental" and "instrumental" phonetics. Over 110 years ago Sweet spoke of "antagonism between the practical linguistic phonetician and the physico-mathematical instrumental phonetician" [2]. In 1935 the psychologist and experimental phonetician E. W. Scripture notoriously claimed that "the investigator might be, and preferably should be, congenitally deaf and totally ignorant of any notions concerning sound or speech" [3]. But it was an unnecessary and manufactured antagonism, fuelled as early as the 1890s by an excessive regard for the methods and trappings of "hard" science [4]. In fact, the development of practical phonetics was partly driven by the failure of the instrumental phonetics of the day to provide any results useful for application in fields such as language teaching, teaching of the deaf, or therapeutic interventions for speech pathologies.

V. Opportunities for re-formulation

The most recent exposition of the merits of practical phonetics training is [1]. The whole landscape of phonetics has changed since Catford's ideas were formed. In [1] "the techniques of instrumental investigation of speech" are still viewed as something separate from the ordinary student's experience, to be encountered "sooner or later" (p. 218). For the last 15 years many tools of the phonetics laboratory have been freely available—even to the solitary learner—on portable devices such as smartphones. Although some texts have attempted an integration of practical and experimental approaches, many otherwise excellent works which continue in wide use perpetuate the anachronistic segregation of descriptive and instrumentally-based approaches [5]. It is time for a re-formulation.

The analysis tools now available to everyone can guide the acquisition of production skills and encourage and foster an even deeper understanding of phonetic theory. The tools range from media applications familiar to all smartphone users (such as photography, video (Fig. 1), and sound recording), through smartphone versions of oscilloscopes, sound-level meters and signal generators (which were once cumbersome and expensive laboratory instruments) to speech-specific tools for spectrum analysis, fundamental frequency extraction, etc. (Fig. 2). At the same time, approaching the tools from a practical standpoint can deepen understanding of the tools themselves. There is no actual conflict between practical and instrumental approaches to phonetics, and neither alone is sufficient. Introduced together and taught in harmony, each can enrich the other.



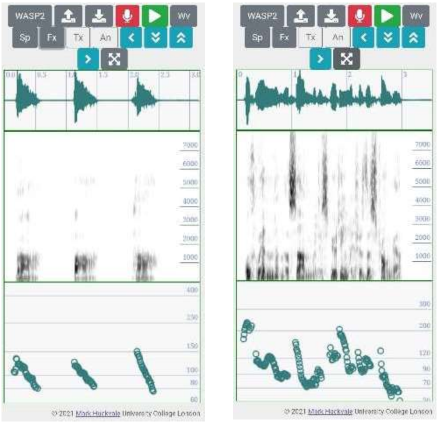


Fig. 1. *(left)* Closed and open phases of a bilabial trill produced by a three-year-old child, captured in slow motion at 920 fps on an Android smartphone. Fig. 2, *(right)* Screen displays from Wasp2 on an Android phone (http://speechandhearing.net/laboratory/wasp/) showing from top, speech waveform, wideband spectrogram, extracted fundamental frequency contour. The left panel shows syllables [ba pa pʰa] used in practising control of VOT, the right panel shows the phrase *The North Wind and the Sun were disputing which was the stronger.*

References

[1] J. C. Catford, A practical introduction to phonetics. 2nd ed. Oxford University Press, 2001.

[2] H. Sweet, The sounds of English: An introduction to English phonetics. Oxford: Clarendon Press, 1908.

[3] E. Fischer-Jørgensen, "Opening address: Some aspects of the 'phonetic sciences', past and present," In M. P. R. van den Broeke & A. Cohen (eds.), Proceedings of the tenth international congress of Phonetic Sciences, Utrecht, 1984, 3–11.

[4] S. Haack, "Six signs of scientism," Logos and Episteme, 3(1), 2001, 75–95.

E. Zsiga, The sounds of language an introduction to phonetics and phonology, Wiley-Blackwell

Gradience and L2 Learning of new phonetic categories vs. recategorization: L2 Spanish stops

Rebeka Campos-Astorkiza, Ohio State University

Keywords —acquisition of phonology, gradient allophonic variation, stops, Spanish

I. Introduction

Models of L2 phonology usually focus on asymmetries in the acquisition of L2 phonemes depending on their L1 status and make reference to (dis)similarity between L1 and L2 sounds to explain L2 acquisition of new sounds [1], [2]. In this study, we explore possible asymmetries in the acquisition of different types of L2 allophonic patterns, namely cases where the L2 has a new set of allophones not present in L1 versus situations where the L2 has a different organization of allophones already present in L1. These two scenarios involve learning a new set of sounds, on the one hand, and recategorizing already existing sounds, on the other. In addition, L1 allophones might present gradience, especially when they result from weakening processes. This project further examines how L2 learners deal with allophonic alternations that are gradient in L1 speech in the two scenarios mentioned above, i.e., new sounds vs. recategorization. To this end, this study compares the L2 acquisition of the Spanish voiced and voiceless stop allophones by L1 American English learners. Spanish /b, d, g/ present an allophonic alternation between voiced stops and approximants, the latter occurring after non-continuant sounds, while English /b, d, g/ alternate between voiced and unaspirated voiceless stops. Thus, L2 learners need to acquire a new set of allophones (approximants) but also recategorize their voiceless stop allophones. Moreover, Spanish /p, t, k/ are unaspirated in all environments, while English /p, t, k/ are produced as aspirated or unaspirated allophones. This is another element in the recategorization of the stop allophones that L2 learners of Spanish need to acquire.

Previous studies on L2 Spanish voiced stops focus mainly on the acquisition of the approximant allophones (e.g. [3], [4], [5]), while some work examines stop allophones production for Spanish voiceless and voiced stops (see [6]). However, a more comprehensive analysis that includes all allophones of voiced and voiceless stops would allow us to compare the acquisition of the two scenarios mentioned earlier. Furthermore, while L2 Spanish approximants and aspirated voiceless stops have been examined gradiently in separate stdudies, by analyzing the acoustic cues of CV-intensity ratio (e.g. [7]) and VOT duration (e.g. [8]), this project brings a new analytical approach by comparing the gradience of the two allophones to each other. Thus, our research questions are, (i) is there a difference in the rate of acquisition of approximant vs. voiced stop vs. voiceless stop allophones for L1 American English learners of Spanish? (ii) is there a difference in the degree of gradience of their approximant and voiceless stop allophones?

II. Methodology

The data comes from a bigger project that combines pedagogy and research. More precisely, it comes from a teaching module developed for college-level Spanish Pronunciation courses at a major Midwestern university in the US. In this module, students record themselves reading a list of words in isolation via a web-based interface and get instant feedback on their pronunciation via that interface. Students complete the module at the beginning (time point T1) and end of the semester (time point T2), which allows us to compare their production at T1 vs. T2 and analyze any changes as manifestations of the students' acquisition process. We examined data from 27 learners who received the same Spanish pronunciation curriculum and teaching methodology. Tokens of /b, d, g/ and /p, t, k/ in two different contexts, i.e., word medial vs. initial, were analyzed following a two-step procedure. First, we categorized each token according to its production, i.e. voiced/voiceless stop, approximant, etc., based on acoustic information. Second, we measure the VOT of voiceless stop allophones of /p, t, k/ and the intensity of approximant allophones of /b, d, g/ to capture the gradience of the allophonic alternations. We tested the effect of time point (T1 vs. T2), stress, word position, and stop place of articulation on the three dependent variables, type of allophone, intensity ratio and VOT, using linear and multinomial regression. Furthermore, to directly compare the two acoustic measures, we scaled the VOT and intensity via z-scores to make them comparable and explore them with linear regression and correlations.

III. Results and Discussion

For /b, d, g/, our results show a significant change in the type of allophone. More precisely, we observe a significant increase in approximants, a decrease in voiced stops and a small change in voiceless stops in T2 compared to T1. As Fig. 1 shows, there is an effect of word position: most voiceless stops occur initially, as expected since this context is utterance-initial position in our data and voiceless allophones of /b, d, g/ in English are common in this context. This initial effect decreases in T2 in favor of voiced stops and even some approximants (Fig. 1). However, the change in type of allophone is greater in word-medial contexts where the rate of approximants changes from 50% to 73% (Fig. 1). As expected, virtually all tokens of /p, t, k/ are realized as such. In terms of the continuous measures, we find no significant change in VOT, which falls within the English aspirated range, although time point interacts with stress and place of articulation, and VOT decreases in stressed syllables and for /p, t/ in T1 vs. T2. In contrast, there is a significant increase in the intensity ratio in T2 compared to T1, indicating that approximants are produced as more weakened at the end of the semester. The intensity ratio is higher in unstressed positions and in word-medial positions, indicating that these contexts present the highest degree of weakening and mirroring the positional effects found for L1 Spanish speakers [9]. Finally, z-scores further show that the intensity ratio for approximants presents a greater change in T2 vs. T1 compared to the VOT of voiceless stops, and that a decrease in VOT, which would indicate less aspiration, is not correlated with increase in intensity, i.e., more weakening, suggesting that learners might improve their voiced stops without an accompanying improvements of their voiceless stops.

Based on these results, we argue that the greater change in approximants compared to voiceless stops for /b, d, g/ suggests that learners are more successful at acquiring a new set of sounds, i.e., approximants word-medially, than at recategorizing their allophones, i.e., producing voiced and not voiceless stops initially. In addition, we find that participants show a limited change in the VOT duration of voiceless stop allophones for /p, t, k/ but a robust change in the intensity for approximants which become weaker by the end of the semester. This finding from the acoustic analysis adds a dimension to the main conclusion in that learners show a positive change not only in terms of how frequently they produce approximants but also in modifying their gradient production as weakened sounds.

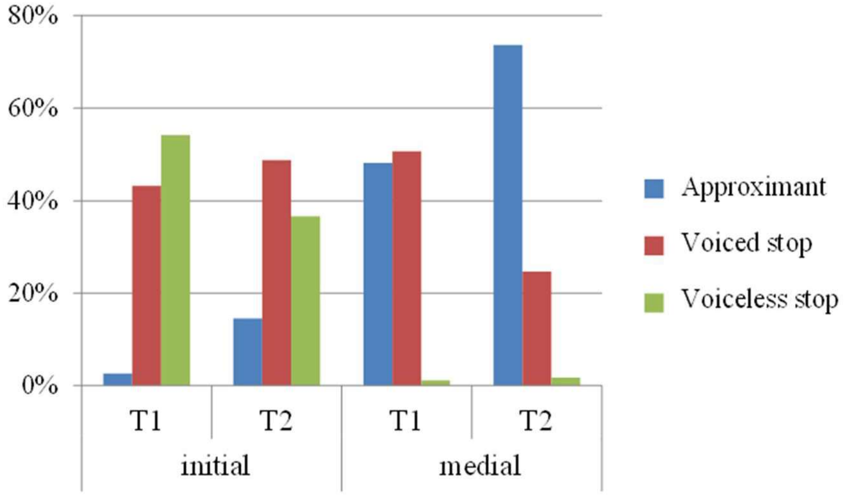


Fig. 1 Percentage of types of realization for /b, d, g/ by timepoint (T1 & T2) & position

References

[1] 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*. Timonium, MD: York Press, 233–277.

[2] Best, C. T. & Tyler, M. D. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In O. S. Bohn & M. J. Munro (Eds.), *Language experience in second-language speech learning: In honor of James Emil Flege*. John Benjamins, 13-34.

[3] Face, T. L. & Menke, M.R. 2009. Acquisition of the Spanish Voiced Spirants by Second Language Learners. In J. Collentine et al. (ed.), *Selected Proceedings of the 11th Hispanic Linguistics Symposium*. Cascadilla Proceedings Project, 39-52.

[4] Lord, G. 2010. The combined effects of instruction and immersion on second language pronunciation. *Foreign Language Annals* 43, 488-503.

[5] Alvord, S. M. & Christiansen, D. 2012. Factors influencing the acquisition of Spanish voiced stop spirantization during an extended stay abroad. *Studies in Spanish and Lusophone Linguistics* 6:2, 239–276.

[6] Face, T. & Menke, M.R. 2020. L2 Acquisition of Spanish VOT by English-Speaking Immigrants in Spain. *Studies in Hispanic and Lusophone linguistics* 13:2, 361-389.

[7] Rogers, B. & S. Alvord (2014). The Gradience of Spirantization: Factors Affecting L2 Production of Intervocalic Spanish [b,d,g]. *Spanish in Context*, 11:3.

[8] Casillas, J. 2020. The Longitudinal Development of Fine-Phonetic Detail: Stop Production in a Domestic Immersion Program. *Language Learning* 70:3, 768–806.

[9] Carrasco, P., J. I. Hualde & M. Simonet. 2012. Dialectal differences in Spanish voiced obstruent allophony: Costa Rican versus Iberian Spanish. *Phonetica* 69,149-179.

Assessing the speaker discriminatory power asymmetry of different acoustic-phonetic parameters

Julio Cesar Cavalcanti[1], Anders Eriksson[1], Plinio A. Barbosa[2]

[1]Stockholm University – SU, [2]University of Campinas – UNICAMP

Keywords — Speaker comparison, Speech tempo, Vowel formants, fundamental frequency

I. Introduction

The analysis of speech timing and melodic parameters is a common procedure within the forensic speaker comparison practice. However, the same level of speaker-discriminatory power should not be assumed when parameters deriving from different acoustic-phonetic dimensions are compared, as previous experimental studies suggest [1, 2, 3]. In view of that, the present study set out to assess what we call the speaker discriminatory power asymmetry regarding parameters from different phonetic dimensions in spontaneous speech, i.e., spectral, melodic, and temporal. We intended to do so by applying a more systematic approach while tackling the issue of data sampling on the discriminatory performance of the parameters.

II. Materials and method

A. Participants

The participants were 20 male subjects, Brazilian Portuguese speakers from the same dialectal area. The participants' age ranged between 19 and 35 years, a mean of 26.4 years. The subjects (10 identical twin pairs) were recruited from a twin research project. However, the focus here is drawn on the comparison among all speakers (i.e., 190 inter-speaker comparisons) rather than on individual twin pairs.

B. Speech material and Recordings

The speech material consisted of spontaneous telephone conversations between siblings. During the recording sessions, the subjects were placed in different rooms, not directly seeing, hearing, or interacting with each other. The speakers were encouraged to start a conversation using a mobile phone while being simultaneously recorded. In all cases, the conversation topics were decided beforehand by the speakers. The transcribed material presented an average duration of about 2:30 minutes per speaker. All recordings were carried out with a sample rate of 44.1 kHz and 16-bit amplitude resolution, using an external audio card (Focusrite Scarlett 2i2) and headset condenser microphones (DPA 4066-B). No audio degradation was applied to the original audio files.

C. Data segmentation and transcription

All speech material was segmented and transcribed in the Praat software [4], following acoustic and auditory criteria. The data annotation which is relevant for the present analysis comprised four distinct textgrid tiers, as follows:

1. Speech chunks: speech intervals on average 3 s long, in most cases corresponding to inter-pause intervals (i.e., stretches of speech between long silent pauses);
2. Vowel-to-vowel intervals: syllable-sized units defined as all the segments uttered between two consecutive vowel onsets;
3. Oral monophthongs: oral monophthongs contained within the speech chunks;
4. Silent pauses: silent pauses with a minimum duration threshold of 100 ms.

D. Acoustic-phonetic parameters

Overall, six acoustic-phonetic parameters were chosen for the comparison based on their relatively better discriminatory performance in relation to other parameters of the same category assessed in previous studies, cf. [5, 6, 7]. The Praat script "ProsodyDescriptorExtractor" [8] was used for the extractions:

1. *f0 median*: *f*0 median in semitones ref 1 Hz/ and in Hertz extracted from speech chunks;
2. *f0 baseline*: base value of *f*0 in semitones ref 1 Hz/ and in Hertz extracted from speech chunks;
3. *Speech rate*: defined as the number of V-V units in each speech chunk divided by its total duration (V-V units/second), including pauses;
4. *Articulation rate*: defined as the number of V-V units in each speech chunk divided by its total duration (V-V units/second), excluding silent pauses;
5. *F3*: the third formant frequency measured at the midpoints of oral monophthongs in Hertz;
6. *F4*: the fourth formant frequency measured at the midpoints of oral monophthongs in Hertz.

E. Statistical analyses

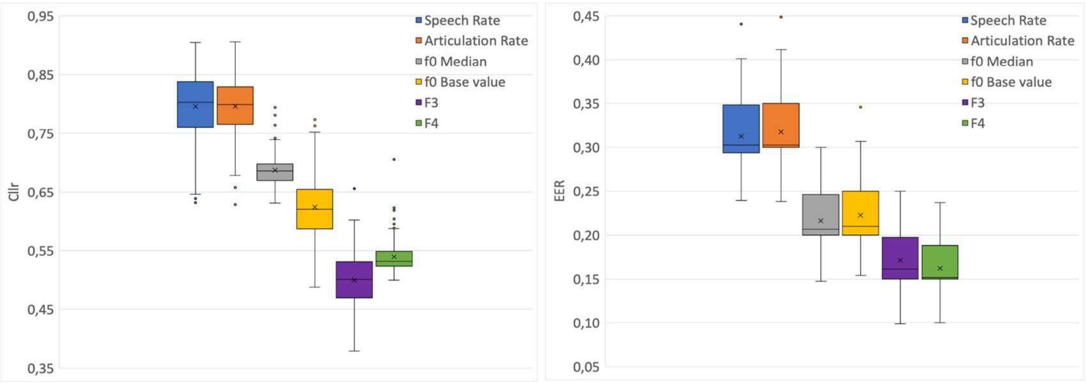
Regarding the assessment of discrimination, two estimates were examined via *cross-validation* as a function of the comparisons among all speakers in the study using the script 'fvclrr' [9]: Log-likelihood-ratio-cost (Cllr) and Equal Error Rate (EER) values. The first is an empirical estimate of the precision of likelihood ratios. It is a measure of validity, initially developed for use in automatic speaker recognition. The second estimate captures the point where the false reject rate (type I error) and false accept rate (type II error) are equal, being used to describe the overall accuracy of a system. Lower Cllr and EER values are compatible with better discriminatory performance, whereas higher Cllr and EER values suggest the opposite trend.

Given the nature of spontaneous speech, a discrepancy in the number of samples produced per subject was observed. In light of that, a random downsampling procedure was conducted to ensure that all participants were represented by the same number of data points in all tests. Such a procedure was repeated 200 times to avoid a selection bias. Both Cllr and EER values were reported after performing tests with the randomly selected data points and applying a calibration procedure using a logistic regression technique.

III. Results

The outcomes are summarized in Figure 1. A general pattern is suggested when inspecting the boxplots regarding Cllr and EER values as a function of acoustic-phonetic parameters. Parameters pertaining to the speech tempo category depicted the worse performance in terms of speaker discriminatory power when assessed in isolation. Such a trend is indicated by the relatively higher median and mean Cllr and EER values. Moreover, from the set of parameters assessed, high formant frequencies, i.e., F3 and F4, were the best performing estimates in terms of discriminability depicting the lowest EER and Cllr values. Based on the 200x resampling of the data, both the *f*0 median and base value displayed relatively similar EER values. However, in terms of Cllr, the *f*0 base value displayed the lowest median, mean and minimum values, suggesting a better overall performance.

Fig. 1. Calibrated Cllr and EER values as a function of the parameters assessed and data sampling.



IV. Final considerations

The results suggested a speaker discriminatory power asymmetry concerning different acoustic-phonetic parameters, in which speech tempo estimates presented a lower discriminatory power when compared to melodic and spectral parameters. Such a finding agrees with previous reports and bears important implications for forensic phonetics, indicating, for example, that the parameters usually most robust to the effect of degradation due to poor audio quality, such as temporal estimates, are not necessarily also the most discriminatory. Most importantly, the results suggest that data sampling appears to be of crucial relevance for the reliability of the results, given the observed variability of Cllr and EER values as a function of data selection.

References

[1] J. H. Künzel, "Some general phonetic and forensic aspects of speaking tempo," The International Journal of Speech, Language and the Law, v. 4, n. 1, pp. 48-83, 1997. Doi" https://doi.org/10.1558/ijsll.v4i1.48

[2] R. Lennon, L. Plug, E. Gold, "A Comparison of Multiple Speech Tempo Measures: Inter-Correlations and Discriminating Power," In: 19th International Congress of the Phonetic Sciences, pp. 785–789, 2019.

[3] V. Hughes; A. Brereton; E. Gold, "Reference sample size and the computation of numerical likelihood ratios using articulation rate," York Papers in Linguistics, University of York, v. 13, pp. 22–46, 2013.

[4] Boersma, Paul & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.10, retrieved 17 March 2022 from http://www.praat.org/

[5] J.C. Cavalcanti, A. Eriksson, A, P.A. Barbosa, Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison. Plos one, v. 16, n. 2, p. e0246645, 2021. Doi: https://doi.org/10.1371/journal.pone.0246645

[6] J.C. Cavalcanti, A. Eriksson, A, P.A. Barbosa, Multiparametric Analysis of Speaking Fundamental Frequency in Genetically Related Speakers Using Different Speech Materials: Some Forensic Implications. Journal of Voice, 2021. Doi: https://doi.org/10.1016/j.jvoice.2021.08.013

[7] J.C. Cavalcanti, A. Eriksson, A, P.A. Barbosa, Multi-parametric analysis of speech timing in inter-talker identical twin pairs and cross-pair comparisons: Some forensic implications. Plos one, v. 17, n. 1, p. e0262800, 2022. Doi: https://doi.org/10.1371/journal.pone.0262800

[8] P. A. Barbosa, Prosody descriptor extractor [Praat script], 2021. https://github.com/pabarbosa/prosody-scripts/tree/master/ProsodyDescriptorExtractor

[9] J. Lo, Fvclrr: Likelihood ratio calculation and testing in forensic voice comparison [R script], 2018. https://github.com/justinjhlo/fvclrr

And it's just like: The discourse-pragmatic and phonetic variation of just with applications to forensic voice comparison

Ben Gibb-Reid, Paul Foulkes, Vincent Hughes – University of York, UK

Keywords — forensic phonetics, discourse-pragmatic variation, socio-phonetics

I. INTRODUCTION

A particularly important issue in forensic voice comparison (FVC) is the lack of direct correspondence in the content of different recordings. That is, recordings are unlikely to share many of the same words. Therefore, a frequently used word (or other feature) in naturally occurring speech is of value to the FVC practitioner because it permits direct comparison. To examine the forensic value of any linguistic features, it is necessary to understand how variable it is between and within speakers, and the factors that affect it in different discourse positions or prosodic contexts. In the present study, the short discourse-pragmatic marker (DPM), *just* is analysed in this way for suitability as a diagnostic feature in FVC. In previous research, other DPMs such as filled pauses (*uh, um*) have been analysed as FVC features with promising results [1, 2]. For the present study, *just*, STRUT and filled pauses were analysed for 100 male Southern Standard British English speakers (DyViS corpus, [3]).

The polyfunctional word *just* was selected because of its high frequency in spontaneous speech, and because previous research describes its variation in pragmatic function. *Just* is the 27th most frequent word in the British National Corpus (2014) at 0.75 per 100 words [5]. Research also shows that *just* is increasing in frequency over time, as demonstrated for younger speakers in Toronto [4]. Other research on the occurrence of *just* [6, 7] has highlighted distinct functions where *just* is used as an adverb (a marker of precision: *That's **just** down the road* – or of time: *I've **just** finished my exams*), as a marker of restriction (*It was **just** me*), or as a discourse marker (to evaluate or minimize: *I **just** watched TV* – or to intensify: *I **just** can't remember*). It is also of interest to FVC whether speakers use *just* in different ways, and therefore these different discourse functions are also analyzed to aid speaker comparison.

II. RESULTS

1,276 tokens of *just* were extracted for analysis, tagged with their function, turn position and following/preceding environment. Each token was also transcribed to show segment elision and allow for formant readings to be taken from the vowel. As expected, *just* was highly frequent, occurring overall 0.88 times per 100 words. Midpoint formant measures for STRUT and the vowel of *um* were also extracted as points of comparison, allowing for likelihood ratio-based testing across 76 speakers. The vowel midpoints for all tokens are displayed in Fig. 1 along with the mean readings for STRUT and *um* vowels. Generally, the vowel in *just* is considerably more raised and/or fronted compared to STRUT or *um*. Fig. 1 also displays four speakers who had mean F1 and F2 values at the upper and lower extremes. Each speaker is represented by an ellipsis showing their standard deviation.

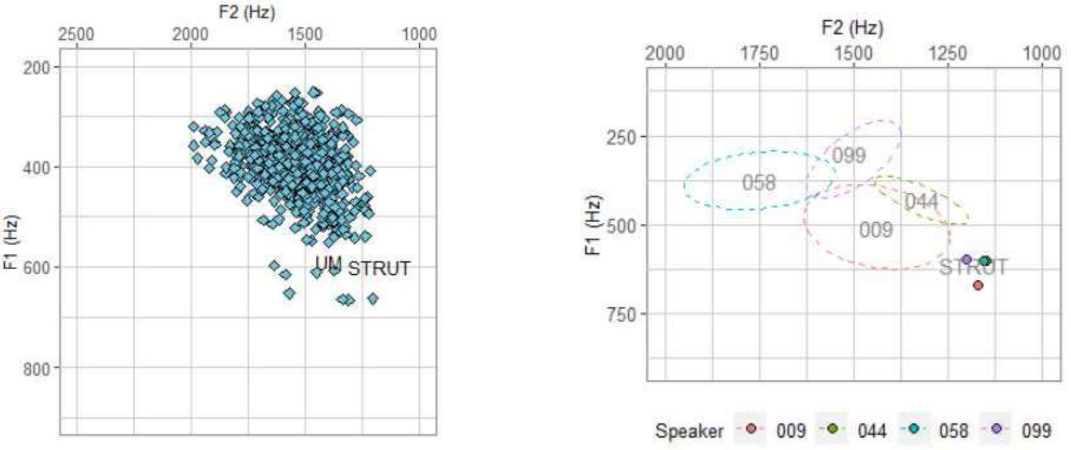


Fig. 1 F1-F2 plot of just vowel midpoints and STRUT and the vowel of *um* (left). F1-F2 of *just* with ellipses showing SD for four speakers (right).

In likelihood ratio-based testing, various models were run comparing acoustic measures of just. Just was also compared with STRUT and um in its discriminatory capacity. Fig. 2 shows the validity measures for these tests, where lower log LR cost (C_{llr}) and equal error rates (EER) correspond to a better-performing system. The left panel shows that F1-F3 of just outperforms the formants of STRUT. It has a lower C_{llr} than um but a very slightly higher EER. The right panel of Figure 2 displays the effect of adding discourse functions of just to speaker comparison models. Just F1-F3 without any function information performs best, whereas adding restrictive or discourse just information reduces model validity. Overall, just shows some promise for FVC application, performing better than um or STRUT. Adding information about just functions, however, does not aid the task of FVC. This is positive, as FVC analysts can treat all tokens of *just* the same, without needing to refer to specific functions – making *just* a broader idiosyncratic feature of the voice.

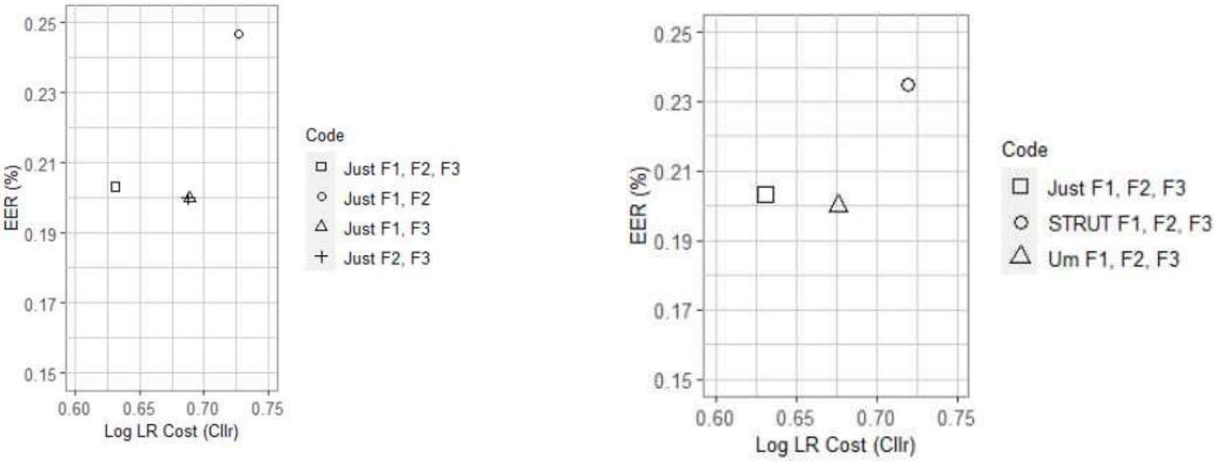


Fig. 2 Plot of log LR cost (Cllr) and equal error rate (EER%) for *just* vowel midpoints across formants (left) and for *just*, STRUT and *um* F1-F3 vowel midpoints (right). NOTE: lower scores on both dimensions are considered desirable for good FVC variables.

REFERENCES

[1] Tschäpe, N., Trouvain, J., Bauer, D., & Jessen, M. (2005). Idiosyncratic patterns of filled pauses. Paper presented at the 14th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Marrakesh, Morocco.

[2] Hughes, V., Wood, S., & Foulkes, P. (2016). “Strength of forensic voice comparison evidence from the acoustics of filled pauses”. International Journal of Speech, Language & the Law. 23(1), 99-132.

[3] Nolan, F., McDougall, K., De Jong, G. & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. International Journal of Speech Language and the Law 16(1): 31-57.

[4] Tagliamonte, S. (2016). Teen talk: The language of adolescents: Cambridge University Press.

[5] Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3): 319-344. DOI: 10.1075/ijcl.22.3.02lov

[6] Beeching, K. (2016). Pragmatic Markers in British English: Meaning in Social Interaction.

[7] Woolford, K. (2021). Just in Tyneside English. World Englishes. doi:10.1111/weng.12542

[8] CCC (forthcoming) It’s not just a sound change: linking phonetic and pragmatic change in a discourse-pragmatic marker.

[9] Drager, K. K. (2011). "Sociophonetic variation and the lemma." Journal of Phonetics 39(4): 694-707.

[10] Schleef, E. & Turton, D. (2016). "Sociophonetic variation of like in British dialects: effects of function, context and predictability." English Language and Linguistics 22(1): 35-75

Prosody training aids second language processing

Sabine Gosselke Berthelsen & Mikael Roll, Lund University

Keywords — prosody, processing, L2 learners, training

I. Introduction

In many languages, there is a systematic relationship between phonetic and prosodic features, on the one hand, and morphosyntax, on the other. If these systematicities are known and the relevant phonetic or prosodic features can easily be identified, a listener can use them actively during speech processing to narrow down the choice of possible continuations at word or sentence level. While native speakers do this automatically and effortlessly, second language learners generally struggle with the identification of phonetic and prosodic features in the second language (L2), especially if they are dissimilar from native language (L1) phonology. Consequentially, they fail to notice the features and make use of the systematicities associated with them. We will discuss the beneficial role of training for L2 learners' use of non-native prosodic features in L2 speech processing.

II. Prosody Processing

A. Prosody processing in a native langauge

Prosody often stands in tight interplay with morphosyntax both at the sentence and word level. It can, for instance, function as a cue to sentence structure [1] or as a cue to word structure [2]. With respect to the latter, prosodic alternations are often part of the inflectional system where they, either on their own [3] or in interplay with affixes [2,4,5], distinguish grammatical features. Different languages use different prosodic parameters in this respect: for example, Danish uses voice quality [3], Swedish uses pitch [2] and Spanish uses stress [5]. When prosody on the word stem is combined with grammatical suffixes, the prosodic cue allows listeners to accurately but typically preconsciously make predictions about how the word continues. When a Central Swedish listener, for instance, encounters low pitch on the word stem of *bil* 'car' in a sentence context like (1), they know immediately that the upcoming suffix is likely the definite singular suffix *-en* as only the singular but not the plural suffix *-ar* or a compound like *bilföraren* is preceded by a low pitch on the word stem. This type of pre-activation makes speech processing rapid and effective.

(1) *Jag såg bilen / bilar / bilföraren köra på cykelbanan.* 'I saw the car / cars / the car driver drive in the bicycle lane.'

The automatic prosody-based prediction has been demonstrated with eye-movement and behavioural data where native listeners quickly fixate and select the correct suffix upon hearing the prosodic cue [5]. It has also been shown with neurophysiological data where non-cued suffixes (e.g., plural *-ar* after low stem tone) elicit error-related responses in the brain while the cue itself produces a stronger pre-activation negativity the more restraining it is with respect to possible continuations [2,4].

B. Prosody processing in a second language

Second language learners are typically initially insensitive to prosody in their L2, especially if the prosodic cues differ from those in the L1 either in nature or in usage [6]. At early and intermediate acquisition stages, learners do not appear to make use of prosodic cues during speech processing. Thus, neither eye-tracking data nor neurophysiological data find evidence of prosody-based prediction at these levels [5,7,8]. L2 learners do not fixate cued words above chance, they do not produce differential pre-activation negativities for differently strong cues, and they do not process incorrectly cued suffixes as speech errors. At upper intermediate and advanced acquisition stages, learners show first signs of prediction-related activity, such as delayed response times to prosody-suffix mismatches and increased accuracy at selecting the correct suffix based on the prosodic cue [5,9]. However, prosody processing does not seem to become equally automatic and efficient in an L2 as it is for native speakers even for highly trained bilinguals [10].

C. The effect of training on prosody processing in a second language

In a series of behavioural and neurophysiological studies, we have since investigated whether focused training of L2 prosody could facilitate native-like speech processing. Notable progress in prosody processing was observed when learners used a prosody-based training app, specifically developed to focus on a wide range of systematic relations between prosody and grammatical suffixes in Swedish [11]. Nineteen participants showed significant improvement in selecting correct suffixes based on a prosodic cue. With just few hours of training, dispersed over ten training sessions, the effect of the acoustic combinatorial training had spread even to untrained words and, as a side effect, even improved production. Importantly, after training, learners' neurophysiological responses revealed native-like processing with differentially strong responses as a factor of prosodic cue strength [7].

We further found that learners who trained on pseudo words with familiar prosodic cues (tone) had an advantage over learners who were previously unfamiliar with the kind of prosodic features found in the novel L2 [12]. Thus, on two consecutive days, 23 participants with a tonal L1 and 23 participants with a non-tonal L1 were trained on 24 novel words in which tonal cues expressed a grammatical inflection (e.g., number). Words were presented auditorily followed by pictures that depicted the intended meaning including grammatical categories. Both participant groups became fairly accurate at matching the novel prosody to its designated meaning within the first training session (i.e., within two hours) [13]. L1-based familiarity with the prosodic cues led to prosody-facilitated grammar processing already at pre-attentive processing stages within just twenty minutes of training. The more familiar the prosody was, the earlier the processing was. Training also allowed for prosody-dependent grammar processing for unfamiliar prosodic cues. However, the new words were initially processed with later, more attentively controlled mechanisms [6,12]. Yet, familiarity advantage decreased with increased training, and earlier, less attention-dependent processing became possible for learners previously unfamiliar with the prosody already at the second training session, i.e., after two hours. These findings suggest that focused training can rapidly develop native-like processing even for unfamiliar word-level prosody.

III. Conclusions

In conclusion, focused training on the relation between prosody and grammar can improve grammatical processing in L2 learners. It can also accelerate the onset of the use of predictive processing strategies and help learners overcome disadvantages due to unfamiliarity with phonetic features or functions. These effects are visible both behaviourally and neuropsychologically.

References

[1] P. Söderström, M. Horne, P. Mannfolk, D. van Westen & M. Roll, "Rapid syntactic pre-activation in Broca's area: Concurrent electrophysiological and haemodynamic recordings," *Brain Research*, vol. 1697, p. 76-82, 2018.

[2] M. Roll, P. Söderström, P. Mannfolk, Y. Shtyrov, M. Johansson, D. van Westen & M. Horne, "Word tones cueing morphosyntactic structure: Neuroanatomical substrates and activation time course assessed by EEG and fMRI.," *Brain and Language*, vol. 150, pp. 14-21, 2015.

[3] A. Oomen, "Gender and plurality in Rendille," in Afroasiatic Linguistics, vol. 8, R. Hetzron and R. G. Schuh, Eds., Undenoa, 1981, pp. 35-75.

[4] A. Hjortdal & M. Roll, "Phonetic and phonological cues to prediction: Neurophysiology of Danish stød," submitted for publication.

[5] N. Sagarra & J. V. Casillas, "Suprasegmental information cues morphological anticipation during L1/L2 lexical access," *Journal of Second Language Studies*, vol. 1, no. 1, pp. 31-59, 2018.

[6] S. Gosselke Berthelsen, M. Horne, Y. Shtyrov & M. Roll, "Native language experience shapes pre-attentive foreign tone processing and guides rapid memory trace build-up: An ERP study," *Psychophysiology*, 2022.

[7] A. Hed, A. Schremm, M. Horne & M. Roll, "Neural correlates of second language acquisition of tone-grammar associations," *The Mental Lexicon*, vol. 14, no. 1, pp. 98-123, 2019.

[8] S. Gosselke Berthelsen, M. Horne, K. J. Brännström, Y. Shtyrov & M. Roll, "Neural processing of morphosyntactic tonal cues in second-language learners," *Journal of Neurolinguistics*, vol. 45, pp. 60-78, 2018.

[9] A. Schremm, P. Söderström, M. Horne & M. Roll, "Implicit acquisition of tone-suffix connections in L2 learners of Swedish," *The Mental Lexicon*, vol. 11, no. 1, pp. 55-75, 2016.

[10] C. Lozano-Argüelles, N. Sagarra & J. V. Casillas, "Slowly but surely: Interpreting facilitates L2 morphological anticipation based on suprasegmental and segmental information," *Bilingualism: Language and Cognition*, pp. 1-11, 2019.

[11] A. Schremm, A. Hed, M. Horne & M. Roll, "Training predictive L2 processing with a digital game: Prototype promotes acquisition of anticipatory use of tone-suffix associations," *Computers & Education*, vol. 114, pp. 206-221, 2017.

[12] S. Gosselke Berthelsen, Y. Shtyrov, M. Horne & M. Roll, "Different neural mechanisms for rapid acquisition of words with grammatical tone in leaners from tonal and non-tonal backgrounds: ERP evidence," *Brain Research*, vol. 1729, pp. 1-15, 2020.

[13] S. Gosselke Berthelsen, M. Horne, Y. Shtyrov & M. Roll, "Phonological transfer effects in novice learners: A learner's brain detects grammar errors only if the language sounds familiar," Bilingualism: Language and Cognition, vol. 24, no. 4, pp. 656-669, 2021.

Learning Sounds through Unconscious Associations

Grenon[1], C. Sheppard[2] and J. Archibald[3]

[1]The University of Tokyo, [2]Waseda University, [3]The University of Victoria

Keywords — Phonetic training-English consonants-Japanese speakers

I. Introduction

Rapidly learning new words when learning a new language (L2) is undisputedly desirable, but it may be harder to commit new words to memory when they contain unfamiliar sounds. Since improving learners' perceptual abilities can bootstrap lexical acquisition both during infancy [1] and adulthood [2, 3] it may be most beneficial to improve the perceptual abilities of language learners from the very onset of L2 acquisition. To date, however, the tasks used for phonetic training may require learners to be already familiar with the L2 phoneme-grapheme correspondence, or may be too strenuous for some populations (e.g. young children) making this kind of training difficult to implement from the very onset of learning. The current study tested the efficiency of the picture association training (PAT), which can be adapted for learners who are not yet literate in the L2 (e.g. Swedish learners of Arabic or Thai), as well as for young children in need of quickly mastering a new language (e.g. children from refugee or expatriate families) making this training implementable at the learning onset of any L2.

Early phonetic training experiments [4] reported significant improvement in the perception of non-native sound contrasts when using an identification task with varied stimuli. A typical identification task consists of presenting auditorily one word at a time to the learners (e.g. 'sheep') and ask them to choose which word was heard (e.g. *ship* or *sheep).* The learners only 'hear' the words, but the choices are written on the computer screen, requiring knowledge of the L2 grapheme-phoneme correspondence, or to be able to extract this information through training. While this task may be adapted for young children by changing the written forms to pictures of the words, the learners still need to have some level of phonological awareness to complete the task. The use of a discrimination task may similarly avoid the orthography problem by presenting auditorily two words consecutively (e.g. 'sheep' [silent interval] 'ship') to the learners and have them decide whether the words were the "same" or "different" (the choices can be written in any language or use symbols). Still, both tasks require considerable levels of sustained focus in order to hear acoustic differences the learners may be insensitive to, and thus, these tasks may be too demanding (or discouraging) for some learners.

Based on findings that visual word recognition triggers the activation of the auditory cortex [5], we hypothesized that it may be possible for learners to acquire new sound contrasts by creating unconscious associations between sounds and pictures. We developed the picture association training (PAT) where the learners are presented with a picture of a *sheep*, for instance, while "at the same time" hearing the word 'sheep' [ʃip] in the headphones. After a short interval, the learners see another picture, of a *ship* for instance, while hearing the word 'ship' [ʃɪp] in the headphones. The learners' task is to decide if the two "pictures" presented were the same or different, while specifically asked to ignore what they hear in the headphones (i.e., they do not need to hear any acoustic difference to answer correctly.) This paradigm was first tested with a vowel contrast ('sheep' vs. 'ship') with Japanese learners of English, who reported that the task was easy to do while their accuracy rate was near ceiling level (about 96%) from training onset. A post-training task evaluating learners' ability to identify the target vowels (i.e., without pictures) after 1 hour of training revealed that the trainees improved their perception of the vowels to the same extent as those trained with a discrimination task requiring learners to focus on the sounds (rather than pictures) [6]. For the current study we were interested in evaluating whether PAT (with focus on pictures) may yield comparable results to the audio-only AX discrimination task (with focus on sounds) when training with a consonantal contrast rather than a vowel contrast. The sound contrast in 'rose' and 'roads' ([roz]-[ro**d**z]) was used as the target contrast, with Japanese learners of English as the experimental group.

II. Method

A. Participants

Sixteen native Japanese speakers (university students), aged between 18 and 22 years old (M = 19), were recruited in Japan. None of them had ever spent more than 2 weeks in an English-speaking country (M = 0.6 week).

B. Stimuli

The stimuli were created from natural 'rose' and 'roads' samples recorded at 44,100Hz using Praat [7] by a female North American English speaker. The closure duration of the [d] in 'roads' was modified from 0 to 60 ms in steps of 10 ms using a script [8]. The vowel duration of each of the resulting tokens was modified from 210 to 300 ms in steps of 30 ms using the same script, which resulted in a total of 28 stimuli. All 28 stimuli were used for the pre-test and post-test, while a subset of 16 stimuli was used for training. The stimuli chosen for training were those at the extreme ends of the stop closure duration continuum, a cue that is used more categorically than vowel duration by English speakers to distinguish 'roads' from 'rose' [9]. The 16 audio stimuli were paired with pictures representing each word (the same two pictures were used for the entire training.) For instance, the 8 audio stimuli most consistently identified as *rose* by native English speakers were always paired with the picture of a *rose*.

C. Procedure

The pre-test and post-test were identical and used a two-alternative forced-choice identification task: the learner would hear the word 'rose', for instance, and had to decide if the word was *rose* or *roads* by pressing a response key. No picture nor feedback were presented during a test.

After the pre-test and before the post-test, the Japanese participants went through 1 hour of training (2 sessions of about 30 minutes). For PAT, the learner would first see a picture of a *rose* for a duration of 250ms, for instance, and at the same time hear the word 'rose' (chosen among the 8 'rose' stimuli). After an ISI of about 1500ms, the learner would see the picture of *roads* for a duration of 250ms while hearing the word 'roads'. The learners' task was to decide whether the two *pictures* were the same or different by pressing a response key. Each trial was followed by a message (feedback) indicating whether the choice was correct. PAT followed the same procedure and used the same set of audio stimuli as the audio-only AX discrimination task reported previously [9], which results were used to compare with the current results. During training, 16 stimulus pairs contrasted in terms of stop closure (these were 'different' pairs), and 16 pairs did not contrast in terms of stop closure (these were 'same' pairs). None of the words was paired with itself. There was a total of 256 training trials per session. The only - but crucial - difference between PAT and the audio-only discrimination task is that in PAT pictures of the words were presented at the same time as the audio stimuli, and participants were required to decide whether the two consecutive *pictures* presented were the same or different, while instructed to ignore the words they heard in the headphones. In contrast, in the audio-only AX discrimination task, participants were required to decide whether the two consecutive *aural stimuli* presented in the headphones were the same or different.

III. Results and Discussion

The research questions addressed by this study were whether any change in the use of vowel duration and the stop closure duration (as in 'rose' vs. 'roads') would be observed after PAT, and whether PAT (with focus on pictures) may trigger changes in perception comparable to the changes reported with an audio-only discrimination task (with focus on sounds). At pre-test, both training groups were comparable in their use of vowel and stop closure duration, and were both significantly different from English speakers' use of these cues [9]. At post-test, neither group changed their use of vowel duration, but there was a change in their use of closure duration towards English behavior. This change was near-significant after only one hour of training ($p = 0.06$) using ANOVA with sphericity corrections. Although this change has not reached a significant level (yet), there was no significant difference between the two training groups ($p = 0.90$).

It has been shown that adults can learn to contrast new sounds by listening "passively" to a contrastive distribution of the target sounds without performing any task [10]. To confirm that the improvement on PAT is related to the creation of unconscious associations between sounds and pictures—rather than being the result of passive listening to a contrastive distribution of the stimuli—we are preparing to test another task where the audio stimuli are associated with the "wrong" pictures for half of the trials. If only the distribution of the stimuli is sufficient to trigger the change, and this occur without any sound-picture associations, then we expect the "random association paradigm" to provide equivalent results to those reported here.

Although greater improvement may be obtained with an identification task [11], there may be cases where this task cannot be used, especially at the onset of L2 learning. The use of PAT may provide a means to bootstrap lexical acquisition by helping language learners to contrast novel sounds from the learning onset of any L2.

References

[1] Werker, J. and Yeung, H. H. (2005). Infant speech perception bootstraps word learning. *Trends in Cognitive Sciences 9*(11), 519-527.

[2] Perfors, A. F. and Dunbar, D. W. (2010). Phonetic training makes word learning easier. In S. Ohlsson and R. Catrambone (eds.) *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1613-1618). Portland, Oregon, August 11-14.

[3] Silbert, N. H., Smith, B. K., Jackson, S. R., Campbell, S. G., Hughes, M. M. and Tare, M. (2015). Non-native phonemic discrimination, phonological short term memory and word learning. *Journal of Phonetics 50,* 99-119.

[4] Logan, J. S., Lively, S. E. and Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *JASA, 89*(2), 874-886.

[5] Haist, F., Song, A. W., Wild, K., Faber, T. L., Popp, C. A. and Morris, R. D. (2001). Linking sight and sound: fMRI evidence of primary auditory cortex activation during visual word recognition. *Brain and Language 76*, 340-350.

[6] Anonymous (2019, will be added after review).

[7] Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.29, retrieved from http://www.praat.org/.

[8] Winn, M. (2014). Make duration continuum [Praat script]. Version August 2014, retrieved April 14, 2017 from http://www.mattwinn.com/praat.html.

[9] Anonymous (2019, will be added after review).

[10] Maye, J. and Gerken, L. (2000). Learning phonemes without minimal pairs. In S. Catherine Howell et al. (eds.) *Proceedings of BUCLD 24* (pp. 522-533). Somerville, MA: Cascadilla Press.

[11] Carlet, A. and Cebrian, J. (2022). The roles of task, segment type, and attention in L2 perceptual training. *Applied Psycholinguistics 43*(2), 1-29.

Revitalization and appreciation of local languages and phonetic training of English in Namibia

Katja Haapanen, Antti Saloranta, Kimmo U. Peltola, Henna Tamminen and Maija S. Peltola

Phonetics and Learning, Age & Bilingualism laboratory, University of Turku

Keywords — language revitalization, language preservation, language education, phonetic training

I. Introduction

Dozens of languages are currently spoken in Namibia. Some of them are originally local Bantu and Khoisan languages, and others are Germanic, having been spoken during the colonization of Namibia. The official language is English, and in addition to it, some of the more common languages are used in the school system at the lower levels before the mode of instruction changes to English only. Many of the original languages have very few speakers and are at an increasing risk of withering away or disappearing completely. Furthermore, the dominant role of English limits the educational possibilities of native speakers of certain local languages, who may struggle to reach the required level of English for studying other subjects. This abstract introduces a project that aims to tackle these issues.

"**Dance as a window to endangered languages and the phonetic world (Tanssi uhanalaisten kielten ja foneettisen maailman tulkkina, T&T&F)**" is a three-year project with two main goals: first, to preserve and revitalize some of the more threatened Namibian languages and cultures, and second, to make education more accessible to speakers of Bantu and Khoisan languages. The first goal will be achieved by recording and saving spoken accounts of historical events, and by increasing linguistic awareness of the languages through tandem language learning and multimodal dance art. The second goal will be tackled with tailored, language-specific training and teaching materials for spoken English.

II. Methods

The target languages in the project are divided into three levels. The languages on the first level are those that are at the highest risk of disappearing, and for them, the main goal is preservation through interview recordings. The speakers are asked to tell stories both about subjects that are important to them, and specific subjects that may be of interest to historians and folklorists. The interviews will be performed with no pre-prepared templates, but rather a list of historically interesting topics that the interviewees may wish to discuss, in addition to relating stories relevant to them personally. The historical topics will mainly concern historical and current presence of Finland in Namibia, such as the influence of the work of Finnish missionaries on local culture in the north of Namibia. The recordings will be archived for research purposes and their content will be analyzed in co-operation with Namibian and Finnish cultural historians.

Languages on the second level are those that are at a lower risk of disappearing, but still endangered. Stories and interviews about everyday topics will be recorded from the speakers with a focus on revitalization. The interviews will be used to create materials for the tandem language learning method. This method brings speakers of rarer and more common languages together, with the intention of both participants learning and gaining knowledge of each other's languages. For the interviews about everyday topics templates will be used to cover the same common topics about each language. In addition, vocabulary lists will be created so that specific words can be recorded for cross-linguistic phonetic analysis. The interview recordings will be given a surface-level phonetic analysis, focusing on vowel and consonant inventories and prosody. These analyses will be used to find the most suitable tandem study partners.

The third level consists of school languages that are commonly spoken in Namibia. These include the Bantu languages Oshikwanyama, Otjiherero and the Khoisan language Khoekhoegowab (Damara). The main goal is to conduct the same interviews covering everyday topics as the second level languages. This will allow for the comparison of phonetic inventories and vocabulary between the local languages and English. Thorough phonetic and acoustic analysis will be performed, focusing on identifying the most problematic sounds and phonemes for the perception and production of English. The results of the analyses will be used to create tailored training materials and exercises for the learning of spoken English.

The English training materials will particularly focus on listen-and-repeat training using individual words, as previous research has shown this to be an effective method of learning difficult second language (L2) sound contrasts [1-12]. Furthermore, we will continue this research by conducting listen-and-repeat training studies with speakers of various Namibian languages in order to broaden our current understanding of L2 phonetic learning.

Recorded material from the languages on the first two levels will also be used in an interpretative dance project, aiming to increase awareness of endangered languages both in Finland and in Namibia. Other methodologies will also be considered as the project progresses and the first phonetic analyses have been performed. At ISAPh2022 we will present the project in more detail in a poster, along with preliminary analyses of any interview recordings collected at that point.

REFERENCES

[1] Immonen, K., Peltola, K. U., Tamminen, H., Alku, P., & Peltola, M. S. (2022). Orthography does not hinder non-native production learning in children. Second Language Research. https://doi.org/10.1177/02676583221076645

[2] Immonen, K., Alku, P., & Peltola, M. S. (2022). Phonetic listen-and-repeat training alters 6–7-year-old children's non-native vowel contrast production after one training session. Journal of Second Language Pronunciation. http://doi.org/10.1075/jslp.21005.imm

[3] Peltola, K.U., Alku, P., Peltola, M.S. (2017). Non-native speech sound production changes even with passive listening training. Linguistica Lettica, 25, 158–172.

[4] Peltola, K.U., Rautaoja, T., Alku, P., Peltola, M.S. (2017). Adult learners and a one-day production training – Small changes but the native language sound system prevails. J Lang Teach Res, 8, 1–7.

[5] Peltola, K.U., Tamminen, H., Alku, P., Peltola, M.S. (2015). Non-native production training with an acoustic model and orthographic or transcription cues. Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK, Paper number 0236.

[6] Peltola, M.S., Lintunen, P., Tamminen, H. (2014). Advanced English learners benefit from explicit pronunciation teaching: an experiment with vowel duration and quality. In P. Lintunen, M.S. Peltola, M-L Varila (Eds.) AFinLA-e Soveltavan kielitieteen tutkimuksia 2014, 6, 86–98.

[7] Saloranta, A., Alku, P., & Peltola, M. S. (2020). Listen-and-repeat training improves perception of second language vowel duration: Evidence from mismatch negativity (MMN) and N1 responses and behavioral discrimination. International Journal of Psychophysiology, *147*, 72–82. https://doi.org/10.1016/j.ijpsycho.2019.11.005

[8] Saloranta, A., Alku, P., Peltola, M.S. (2017). Learning and generalization of vowel duration with production training: behavioral results. Linguistica Lettica, 25, 67–87.

[9] Saloranta, A., Tamminen, H., Alku, P., Peltola, M.S. (2015). Learning of a non-native vowel through instructed production training. Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, UK, Paper number 235.

[10] Savo, S., Peltola, M.S. 2019. Arabic-speakers learning Finnish vowels: Short-term phonetic training supports second language vowel production. J Lang Teach Res, 10, 45–50.

[11] Taimi, L., Jähi, K., Alku, P., Peltola, M.S. (2014). Children learning a non-native vowel – The effect of a two-day production training. J Lang Teach Res, 5, 1229–1235.

Tamminen, H., Peltola M.S., Kujala, T. & Näätänen, R. (2015). Phonetic training and non-native speech perception - new memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioural measures. *International Journal of Psychophysiology*, 97, 23-29

4th International Symposium on Applied Phonetics (ISAPh2022), September 14-16 2022, Lund University, Sweden

A forensic-phonetic analysis of automatic speech transcription errors

Lauren Harrington, Department of Language and Linguistic Science, University of York, York, UK

Keywords —forensic, transcription, speech technology, regional accent

I. Introduction

The automatic transcription of speech has become an increasingly common part of daily life for many people, in the form of home voice assistants, voice passwords or closed captions on videos (e.g. lectures, television programmes, online videos). Orthographic transcription of recordings, such as police-suspect interviews, is a common form of forensic speech evidence but is an extremely time-consuming process for forensic practitioners. There are many commercial online services which offer fast automatic transcription of video or audio recordings. In the present study, the performance of such systems in more challenging scenarios, such as noisy audio and regionally-accented speech, will be assessed in order to explore whether there is any scope for incorporating automatic methods into the transcription of forensic audio recordings. Quantified overall error rates that are usually used in automatic transcription assessment can be distracting in forensic contexts and obscure details about performance. For example, two systems could achieve the same score where one has fully deleted an utterance but the other has substituted key words, changing the meaning. As a result, this study will offer a fine-grained analysis of the errors made and the phonetic motivations behind such errors. Errors will also be compared across audio qualities and more generally across regional accents.

II. Methodology

8 utterances per speaker were extracted from 2 speakers of Standard Southern British English (retrieved from the Dynamic Variability in Speech database [1]) and 2 speakers of West Yorkshire English (retrieved from the West Yorkshire Regional English Database [2]). Utterances range from 14 to 20 words and 3 to 6 seconds in length, featuring a single speaker in studio quality conditions. In addition, the intensity of the 32 recordings was scaled to an average of 70 dB in Praat [3] and then mixed with speech-shaped noise in Audacity. Online speech-to-text service Rev.ai was chosen for this study due to its claims of resilience against noisy audio and its Global English language model which is trained on "a multitude of… accents/dialects from all over the world" [4]. A total of 64 recordings were uploaded to Rev.ai and the resulting transcripts were compiled in a spreadsheet, manually aligned with the corresponding reference transcript, and each word pairing was marked as a match or an error (substitution, insertion or deletion).

TABLE I. EXAMPLES OF THE THREE TYPES OF TRANSCRIPTION ERRORS.

Error Type	**Transcript**	
	Reference	***Automatic***
Substitution	There's a deer park	There's a diff Huck
Insertion	Um he's a tour guide	Um and he's a tour guide
Deletion	A boat house	A house

III. Preliminary Results

Initial analysis suggests two general patterns, the first of which is a disparity in performance across the two accents. In both listening conditions, total error rates were much higher for West Yorkshire English (the non-standard regional accent) than for Standard Southern British English (SSBE), with 37 more errors made for Yorkshire speakers in studio quality conditions and 34 more errors in poorer listening conditions. In forensic casework, substitutions and insertions can be judged as more detrimental than deletions since these types of errors add information that wasn't contained within the signal, which can cause more damage than reducing information [5]. Comparing substitution errors across accents shows much higher numbers of substitutions in West Yorkshire English than in SSBE, as well as higher percentages of substitutions involving content words. The second general pattern shows worse performance for poorer quality recordings, in which higher numbers of all error types are observed within both accents.

TABLE II. PERCENTAGES (AND RAW NUMBERS) OF EACH TYPE OF ERROR MADE, WITHIN EACH ACCENT AND EACH AUDIO QUALITY. SSN REPRESENTS THE DEGRADED AUDIO RECORDINGS WHICH WERE MIXED WITH SPEECH-SHAPED NOISE.

	Standard Southern British English		**West Yorkshire English**	
	Studio	***SSN***	***Studio***	***SSN***
Substitution	48% (26)	38% (31)	58% (53)	55% (63)
Insertion	0% (0)	3% (2)	9% (8)	9% (11)
Deletion	52% (28)	59% (48)	33% (30)	36% (41)
Total number of errors	54	81	91	115

IV. IMPLICATIONS

The implications of these findings will be discussed in relation to forensics, speech technology and sociophonetics. The current study is one part of a wider project that is situated at the intersection of these three fields, and which aims to improve methods in forensic casework and the understanding of the effect of regional accent on transcription performance.

REFERENCES

[1] F. Nolan, K. McDougall, G. de Jong and T. Hudson, "The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research.," *International Journal of Speech Language and the Law*, vol. 16(1), pp. 31-57, 2009.

[2] E. Gold, S. Ross and K. Earnshaw, "The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework," *Interspeech 2018: Speech Research for Emerging Markets in Multilingual Societies,* pp. 2748-2752, Hyderabad, India, 2018.

[3] P. Boersma and D. Weenink, "Doing phonetics by computer," [Computer program]. Version 6.2.09, retrieved 12th November 2021 from http://www.praat.org/.

[4] F. Manjiyani, "What is Rev.ai's accuracy?" Rev.ai Help Center. https://help.rev.ai/en/articles/3813288-what-is-rev-ai-s-accuracy, 2022.

N. Tchäpe & I. Wagner, "Analysis of disputed utterances: A proficiency test," [Conference presentation]. *21st International Association for Forensic Phonetics and Acoustics (IAFPA) Conference*, Santander, Spain, 2012.

The lexical and intonational realisation of backchannels is less constrained in spontaneous than task-based conversation

Alicia Janz, Simon Wehrle, Simona Sbranna & Martine Grice

IfL Phonetik, University of Cologne, Germany

Keywords — backchannels, conversational dynamics, intonation, lexical choice, dialogue

The production of feedback signals such as backchannels (BC) can be seen as reflecting mutual understanding. BCs play an important role in constructing and maintaining shared knowledge between interlocutors in a conversation [1, 2, 3]. In experimental settings, tasks like the Map Task [4] are commonly used to elicit conversations that contain a large number of feedback signals. In a Map Task, two participants collaborate to transfer a given route from one participant's map to the other. In intonation research, Map Task are usually recorded without visual contact, so that participants have to rely solely on the spoken communication channel.

Past research has shown that the use of backchannels can differ considerably depending on the conversational setting [5], as well as the level of proficiency of a speaker in a given language [6]. Task-based conversations, for example, in which participants have to reach a clear understanding of the *common ground* (the knowledge they share) [7], require more explicit positive evidence of mutual understanding than casual conversations without a set goal. In the latter, absolute certainty about the current status of common ground is not as important from a strictly functional perspective [5]. BCs might be seen here to primarily serve a wider range of social functions [8]. This might be reflected in the types of lexical tokens interlocutors use to express feedback, as well as in the intonation contours produced.

To investigate the interplay of lexical choice of BC and intonation contour within different conversational settings, we conducted a pilot study. Two dyads of monolingual German speakers were recorded in two different settings, first while having a spontaneous conversation and then while taking part in a Map Task. It is important to note that the visual channel was available only in the spontaneous conversations.

For a first exploration of our data set (Map Task: 198 BC tokens; spontaneous: 37 BC tokens), we categorized intonation contours into rising, falling, and level contours by measuring the difference in semitones between the beginning and the end of each token. We found that intonation contours differed according to conversational setting. Table 1 shows that, in task-oriented conversation, speakers used predominantly rising tokens (53.1%) while in spontaneous conversation most tokens were produced with falling or level intonation (fall: 45.9%, level: 40.5%). Previous research on West Germanic languages has proposed that rising intonation is typical for BCs [9], whereas falling or level intonation is associated more often with filled pauses, which stand in direct contrast to BCs in terms of dialogue management (serving to extend the ongoing turn of a speaker, rather than that of the interlocutor) [10]. Moreover, in our study speakers used more 'standard' lexical types of BC like "ja" (*yes*), "mmhm", and "okay" in Map Task conversations. In spontaneous conversations, on the other hand, almost a quarter of all utterances were 'non-standard' BC types like "stimmt" (*right*), "voll" (*totally*) or "cool", which in related work on task-oriented dialogues accounted for less than 10% of all tokens [11, 12] (see Figure 1).

These findings suggest that the production of backchannels differs between task-oriented and spontaneous conversations. While in Map Tasks there is an inherent, functional motivation for interlocutors to constantly update and confirm the current status of common ground using positive feedback signals, spontaneous conversation, without an explicit goal, does not necessarily require the same degree of precision. This may be behind the use of a greater variety of lexical types and prosodic realisations, suggesting that, overall, BCs are used in a more varied and flexible manner in the absence of a constraining and goal-oriented conversational context.

TABLE III. PROPORTIONS OF INTONATION CONTOUR BY CONDITION

Condition	**Proportions of intonation contour**		
	Fall	***Level***	***Rise***
Map Task	30.6 %	16.3 %	53.1 %
Spontaneous	45.9 %	40.5 %	13.5 %

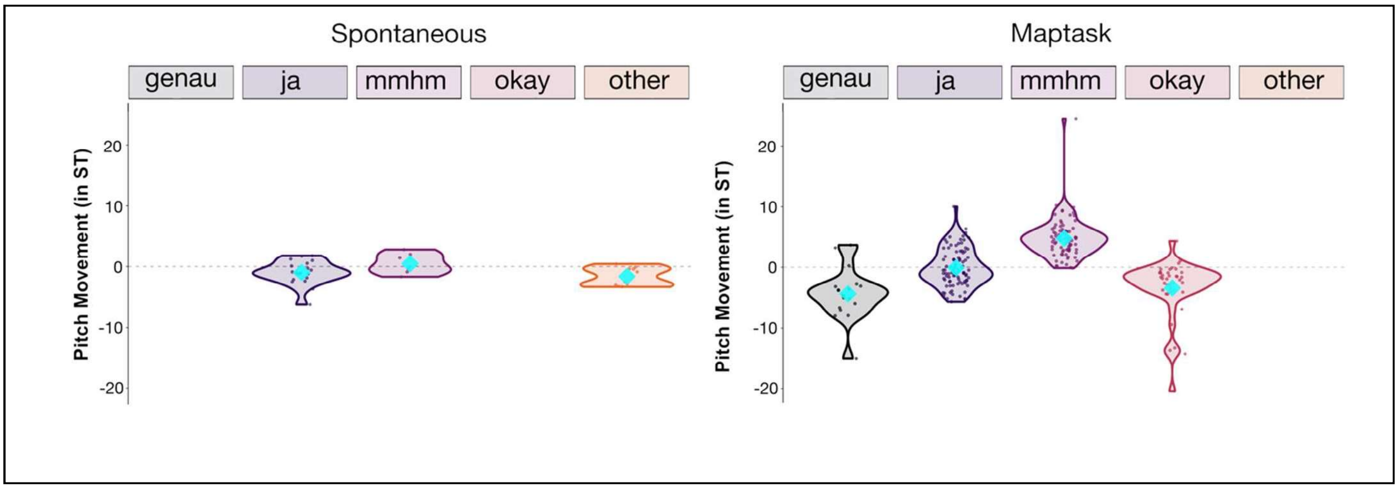


Fig. 2. Pitch movement in semitones for individual BC tokens in spontaneous and Map Task conversation. Cyan diamonds represent mean values; positive values represent rising contours; negative values represent falling contours.

REFERENCES

[1] Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive science*, *27*(2), 195-225.

[2] Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. Analyzing discourse: Text and talk, 71, 93.

[3] Caspers, J. (2000). Melodic characteristics of backchannels in Dutch Map Task dialogues. *Proc. of ICSLP 2000, Bejiing*, China.

[4] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... & Sotillo, C. (1991). The HCRC map task corpus. *Language and speech*, *34*(4), 351-366.

[5] Dideriksen, Christina, et al. Quantifying the interplay of conversational devices in building mutual understanding. *PsyArXiv. October*, 2020, 12. Jg.

[6] Cutrone, P. (2014). A cross-cultural examination of the backchannel behavior of Japanese and Americans: Considerations for Japanese EFL learners.

[7] Clark, H. H. (2009). Context and Common Ground. *May L. Jacob (Ed.), Concise Encyclopedia of Pragmatics*. 116–119.

[8] Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions.

[9] Benus, S., Gravano, A., & Hirschberg, J. B. (2007). The prosody of backchannels in American English.

[10] Belz, M., & Reichel, U. D. (2015). Pitch characteristics of filled pauses in spontaneous speech.

[11] Wehrle, S. (2022). A Multi-Dimensional Analysis of Conversation and Intonation in Autism Spectrum Disorder, *Ph.D. dissertation, University of Cologne.*

[12] Sbranna S., Möking E., Wehrle S., Grice M., (to appear). "Backchannelling across Languages: Rate, Lexical Choice and Intonation in L1 Italian, L1 German and L2 German". 11th International Conference on Speech Prosody, 2022, conference proceedings.

The role of pause location in perceived fluency and proficiency in L2 Finnish

Heini Kallio, Mikko Kuronen, and Liisa Koivusalo

University of Jyväskylä, Finland

Keywords — L2 fluency, pause location, automatic assessment

I. Introduction

Second or foreign language (L2) fluency has been widely studied from the perspective of temporal features related to speed and pausing. Studies have found measures such as speech and articulation rate as well as pause duration, pause frequency, and pause-time ratio to affect the perception of L2 fluency [1,2,3]. Pause locations have been less studied, but they have also been found to have a significant role in perceived fluency [4,5]. Fluent speech tends to have pauses at grammatical junctures, whereas disfluent speakers often pause within clauses or utterances [6,7]. Research on the effect of pausing on the perceived fluency and proficiency in L2 Finnish has yet remained marginal. In a recent study, we measured temporal fluency from spontaneous L2 Finnish speech and found that the rate of silent pauses (> 250 ms) and average duration of a composite break (> 250 ms) significantly affect the perceived proficiency and fluency [8]. However, this was not the case for all L2 speakers: some speakers with similarly high pause rate vary notably in their proficiency and fluency ratings. In the current study, we scrutinize this observation by analyzing pause locations in relation to perceived fluency and proficiency. This study is part of the (anonymized) project that investigates and develops automatic tools for spoken L2 assessment and practicing purposes.

II. Materials and methods

The speech data consists of a subset of the data used and described by Authors in [8]. Ten spontaneous speech samples with similar rate of silent pauses >250 ms were selected for pilot analysis: five speakers with beginner level (A-level) Finnish proficiency and five speakers with intermediate level (B-level) Finnish proficiency. The fluency ratings differed correspondingly between the speaker groups.

Previous studies have investigated the effect of pause location to speech fluency mainly with regards to their occurrence between or within clauses [4] or constituents [6]. Here we investigate pauses between and within clauses (defined as a constituent that links a predicate to a subject or object) and phrases (defined as a group of words that act together as a grammatical unit, but do not necessarily include a predicate). We analyze pauses between and within noun phrases, verb phrases, and adverbial phrases. Both clauses and phrases were considered to include more than one word. In addition, pauses within words, or between an incomplete and a corrected word, were measured. All pause types and their definitions are presented in Table 1.

TABLE I. Pause types

Pause type	Abbreviation	Variables computed
Pause (> 250 ms) between clauses	BCP	mean duration (meanBCP), rate (BCPrate), ratio (BCPratio)
Pause (> 250 ms) within clause	WCP	mean duration (meanWCP), rate (WCPrate), ratio (WCPratio)
Pause (> 250 ms) between phrases	BPP	mean duration (meanBPP), rate (BPPrate), ratio (BPPratio)
Pause (> 250 ms) within phrase	WPP	mean duration (meanWPP), rate (WPPrate), ratio (WPPratio)
Pause (> 250 ms) within word	WWP	mean duration (meanWWP), rate (WWPrate), ratio (WWPratio)

Pauses between and within clauses and phrases were manually annotated in spontaneous speech samples using Praat [9]. Pauses were considered to include either silence or hesitations such as “umm” or “hmm”. Pause rates were operationalized as number of pauses per minute, and pause ratios were operationalized as the relative proportion of pause type in response (total duration of pauses/total duration of response). Pause measures were computed from extracted annotation intervals using R [10].

III. Results

The pause variables were compared between the two speaker groups using Wilcoxon rank sum test with continuity correction. Clause-based and within-word pause measures failed to provide significant differences between groups. Instead, three phrase-based pause measures differed significantly between the speaker groups: meanBPP ($p < 0.05$), BPPrate ($p < 0.05$), and WPPratio (p <

0.05). The distributions of phrase-based pause measures are shown in Fig. 1. Differences in distributions are also visible in meanWPP, WPPrate, and BPPratio, although the differences remained non-significant. Expectedly, the mean durations of WPP and BPP were longer in A-level speakers than B-level speakers. Similarly, the WPP ratio was significantly higher in A-level group than in B-level group. WPP rate, however, did not provide significant differences between the groups. BPP rate, in turn, was significantly higher with B-level speakers than A-level speakers, indicating that intermediate speakers pause more frequently between phrases than beginners.

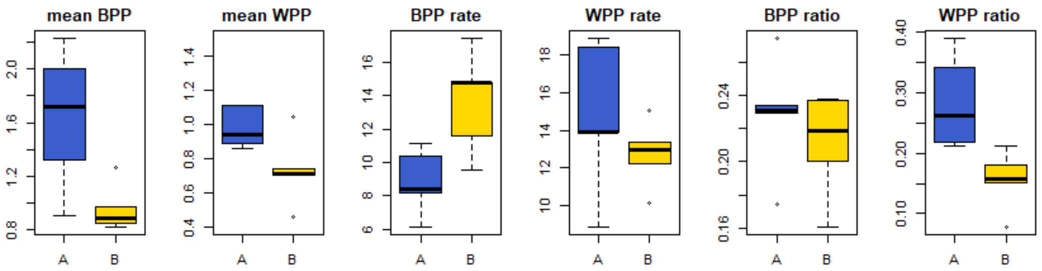


Fig. 1. The distributions of phrase-based pause measures within A-level (blue) and B-level speakers (yellow).

IV. Discussion and conclusions

Results from the pilot study indicate, in line with previous studies, that fluent and disfluent speakers do in fact differ with respect to their pause distributions [4,6]. Differing from previous studies, however, clause-based pause measures did not provide significant differences between speaker groups. Instead, the results show that beginner L2 Finnish speakers keep longer pauses both between and within phrases than intermediate L2 speakers. Interestingly, the rate of pauses between phrases was significantly higher with B-level speakers than with A-level speakers, while the rate of pauses within phrases was somewhat higher with A-level than B-level speakers, although the difference remained non-significant. The reason can be in the operationalization of pause rate: here we measured pause rate as average number of pauses per minute, but since the pauses of A-level speakers are significantly longer in duration than the ones of B-level speakers, a more relevant measure would be the number of pauses in relation to number of phrases in response [4]. Pauses within words provided no significant differences between the speaker groups, but the occurrence of such pauses was sparse.

To conclude, the results here indicate that intermediate L2 Finnish speakers seem to have a pausing pattern that supports the syntactic structure in speech better than beginner L2 Finnish speakers. Phrase-based pause measures seem to work better than clause-based pause measures, but the deficiencies in syntactic structure in L2 speech sometimes make the detection of phrases challenging. In the future, we will extend the investigation of pause conditions to a larger data collected from L2 Finnish speakers and test the effect of pause location to fluency and proficiency ratings with regression models. The results can have important implications for the development of automatic feedback for practicing L2 speaking skills as well as for the improvement of automatic L2 speaking assessment.

References

[1] H. R. Bosker, A.-F. Pinget, H. Quené, T. Sanders, and N. H. De Jong, “What makes speech sound fluent? The contributions of pauses, speed and repairs,” Language Testing, vol. 30, no. 2, pp. 159–175, 2013.

[2] H. Kallio, J. Simko, A. Huhta, R. Karhila, M. Vainio, E. Lindroos, R. Hildén, and M. Kurimo, “Towards the phonetic basis of spoken second language assessment: temporal features as indicators of perceived proficiency level,” AFinLA-e: Soveltavan kielitieteen tutkimuksia, no. 10, pp. 193–213, 2017.

[3] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech,” The Journal of the Acoustical Society of America, vol 111, no. 6, pp. 2862–2873, 2002.

[4] J. Kahng, “The effect of pause location on perceived fluency,” Applied Psycholinguistics, vol. 39, no. 3, pp. 569-591, 2017.

[5] S. Suzuki, and J. Kormos, "Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech," Studies in Second Language Acquisition 42.1, pp. 143-167, 2020.

[6] A. Riazantseva, "Second language proficiency and pausing a study of Russian speakers of English," Studies in Second Language Acquisition 23.4, pp. 497-526, 2001.

[7] J. Kahng, “Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall,” Language Learning, vol. 64, 809–854, 2014.

[8] Authors, 2022.

[9] Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer. Amsterdam: University of Amsterdam. http://www.fon.hum.uva.nl/praat/

[10] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rapid Adaptation to NNS Japanese Pronunciation

Naoko Kinoshita[1], Chris Sheppard[2]

Waseda University

[1]kinoshita@waseda.jp, [2]chris@waseda.jp.

Keywords —rapid adaptation, NS-NNS communication, L2 Pronunciation, Japanese Language

I. Introduction

A non-native accent can be detrimental to the quality of communication. Previous research has shown that foreign accents influence the comprehensibility and intelligibility of speech [1], the attitudes of native speakers to speakers of accented speech [2], and the quality of life of immigrants in their target language communities [3]. While much of the focus of L2 learning and teaching has been to remedy this situation by focusing on improving the accent of non-native speakers, it is also possible to approach the problem by training native speakers to listen to non-native accents [4]. Both experience in listening to non-native accents [5] and instruction [6] improves native speaker ability to comprehend non-native speech. Based on these results, Clarke and Garrett conducted research which demonstrated that reaction times identifying accented words improved after exposure to just 12 low-context sentences. The results lead them to conclude that native speakers are able to rapidly adjust to non-native speech with minimal exposure [7].

To date, much of the research in this area has been conducted with native speakers of English. Considering Japan will increase the number of foreign workers [8], it is necessary to determine if Japanese native speakers can adapt rapidly to non-native Japanese pronunciation. This paper reports research which attempted to confirm if the rapid adaptation to non-native English accents by native speakers could be replicated for Japanese native speakers listening to the Japanese speech of non-native Vietnamese speakers.

II. Method

40 Native Japanese university students were recruited and offered 3000 yen in exchange for their participation. They reported minimal contact with Vietnamese speakers of Japanese. A Japanese version of the Revised Speech Perception In Noise (SPIN-R) test [9], [10] was created. 32 three-mora (2-3 syllable) pairs of words with similar frequencies (10-50 wpm) were selected from the Balanced Corpus of Contemporary Written Japanese [11]. The pairs of words differed by just one phonetic element (ex. *chizu* (map)/ *chiizu* (cheese). Eight blocks of four words each were created ensuring each block had a similar phonetic make-up. Low context sentences of similar length were created as carrier sentences for the words (ex. *futon/futan ni tsuite ooku no hito kara fuman ga deta.* (Many people were dissatisfied with the "beds/burden"). The target words were placed at the beginning of the sentence, to account for Japanese language structure. (They are at the end in the English version [7]). Two blocks were selected for the practice tasks, four for the experiment tasks, and two for the baseline tasks. The sentences in the practice and baseline tasks were recorded by a female native Japanese speaker using a Sony PCM recorder (44 KHz). The four blocks in the experimental tasks were recorded by a Vietnamese female intermediate Japanese speaker and by a different female native Japanese speaker.

The experiment was conducted using the Gorilla experiment builder [12]. The participants were randomly assigned to three groups: the base group, the control group, and the experimental group. All groups completed the practice tasks first. The base group completed the four blocks recorded by the native speaker, the control group completed the first three blocks recorded by the native speaker and the final block by the Vietnamese speaker, and the experimental group completed all four blocks recorded by the non-native speaker. All groups completed the baseline tasks last. For each of the words, the participants completed a judgement task. Immediately after listening to the recording of the sentence, the participant was asked to judge if the word on the screen matched the word in the sentence. Once the answer was provided, the participant was given feedback on the accuracy of their choice. The reaction time between listening to the sentence and completing the judgment task and the correctness of the judgement was recorded.

The reaction times for each of the words was adjusted by subtracting the average baseline response time. Any incorrect responses (35) and responses that took longer than 3000 milliseconds (3) were removed from the data. The means were analyzed with a mixed effects ANOVA using the ezANOVA command in the ez package in R [13]. If rapid adaptation was taking place, reaction times would increase in the control group's 4th block when compared to the experimental group.

III. Results and Discussion

Fig. 1 shows the results of the adjusted reaction times for each of the groups by block. The results of the mixed effects ANOVA using the Huynh-Feldt Correction demonstrated that there was an overall effect for Group, (F (2, 37) = 3.47, p =.042), Block (F (3, 111) = 21.1, ε_{FE} = 0.87, p < .001) and an interaction effect between Group and Block (F (6,111) = 5.94,ε_{FE} = 0.87, p < .001). Follow-up analyses demonstrated there was a significant interaction between the Experiment Group and the Control group across Block 3 and Block 4 (F (1, 24) = 12.62, p =.002).

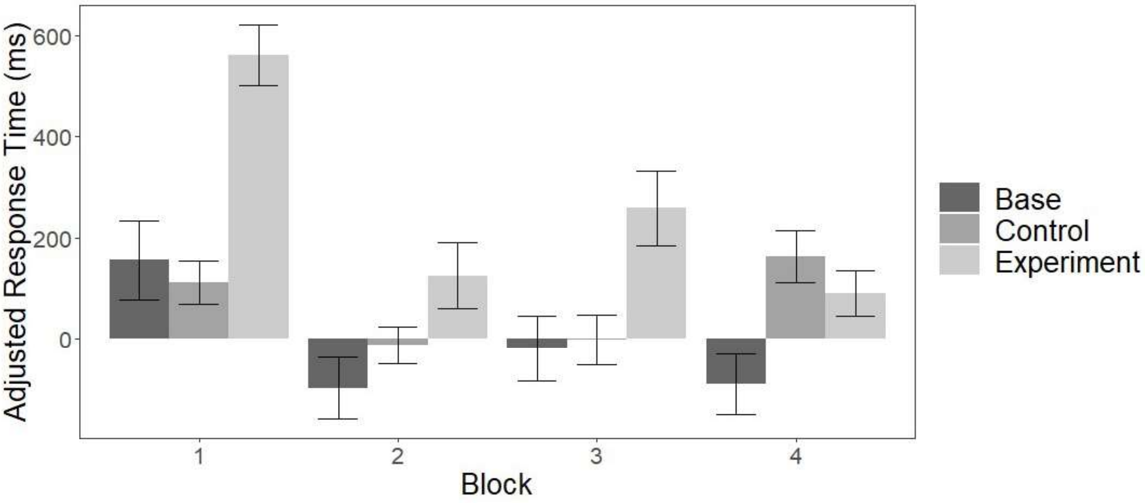


Fig. 3. The Adjusted Response Time (ms) for Group by Block (Error bars show standard error)

The results replicated those of Clarke and Garret [7], demonstrating that Japanese native speakers exhibit rapid adaptation to non-native Japanese pronunciation, in this case of a Vietnamese speaker. The significant Group (Control and Experiment) and Block (Block 3 and Block 4) interaction where the experiment group improves its response times, but the control group slows its response time clearly demonstrates that the experimental group has learned to identify non-native speech at a faster rate than the control group, likely, as a result of greater exposure to non-native speech. The results also showed that all groups improved on the task over time, as was expected. However there was some unexpected variability in the responses. This is most likely because of variable reactions to the words and some of the non-native speaker items were mistakenly identified more than others (*kookyuu* (luxurious) - 7 times), *kaban* (bag) - 5 times, *renzu* (lens) - 6 times). It may be useful to increase the block sizes in future research and select words which do not impact on comprehension to reduce error from this source. Future research would also need to confirm the robustness of this effect: how long does it last, and does this adaptation transfer to other non-native speakers.

REFERENCES

[1] M. J. Munro and T. M. Derwing, "Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech." Lang. Speech vol. 38, 1995, pp. 289–306.

[2] M. J. Munro, "A primer on accent discrimination in the Canadian context." TESL Canada Journal vol. 20. 2003, pp. 38-51.

[3] H. K. Carlson, and M. A. McHenry. "Effect of accent and dialect on employability." Journal of employment counseling vol. 43, 2006, pp. 70-83.

[4] T. M. Derwing and M. J. Munro, "Training native speakers to listen to L2 speech. " In Social Dynamics in Second Language Accent, J. M. Levis and A. M. Moyer, Eds. Berlin: De Guryter, 2014, pp. 219-238.

[5] A. R. Bradlow and T. Bent, T. "Listener adaptation to foreign accented English, " in Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, 2003, edited by M. J. Sole, D. Recasens, and J. Romero, 2003, pp. 2881–2884.

[6] C. M. Clarke, "Perceptual adjustment to foreign-accented English." J. Acoust. Soc. Am. vol. 107, 2000, p. 2856.

[7] C. M. Clarke and M. F. Garrett. "Rapid adaptation to foreign-accented English." J. Acoust. Soc. Am. vol. 116, 2004, pp. 3647-3658.

[8] S. Takizawa, "Japan's Immigration Policy 2015-2020: Implications for Human Security of Immigrant Workers and Refugees." Journal of Human Security Studies vol. 10, 2021, pp. 51-78.

[9] D. N. Kalikow, K. N. Stevens and L. L. Elliott, L. L., "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. " J. Acoust. Soc. Am. vol. 61, 1977, pp. 1337–1351.

[10] R. C. Bilger, "Manual for the Clinical Use of the Revised SPIN Test." Univ. of Illinois, Champaign, IL., 1984.

[11] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, and Y. Den. "Balanced corpus of contemporary written Japanese." Language resources and evaluation vol. 48, 2014, pp. 345-371.

[12] A.L.Anwyl-Irvine, J. Massonié, A. Flitton, N. Z. Kirkham, J. K. Evershed, "Gorilla in our midst: an online behavioural experiment builder. " Behavior Research Methods vol. 52, 2020, pp. 388-407.

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

[14] C. Ryan, "Language use in the United States, 2011. " Washington, DC: U.S. Census Bureau; 2013.

[15] C. Ikeguchi, "Intercultural adjustment-reconsidering the issues: The case of foreigners in Japan." Intercultural Communication Studies vol. 16, 2007, p. 99.

(This research was supported by JSPS KAKENHI (Grant Number: 21H04348)

4th International Symposium on Applied Phonetics (ISAPh2022), September 14-16 2022, Lund University, Sweden

EFL learners' L2 pronunciation noticing skills in an instructed learning context

Pekka Lintunen[1] & Hanna Kivistö-de Souza[2]

[1]University of Turku, Finland [2]Federal University of Santa Catarina in Florianópolis, Brazil

Keywords — Metaphonetic knowledge, phonological awareness, L2 speech learning, noticing

I. Introduction

Compared to many other language subskills, second language (L2) spoken skills and pronunciation learning are notoriously time-consuming and laborious tasks that the L2 learner frequently has to face alone, as pronunciation is not always sufficiently addressed in L2 classroom instruction [1, 2]. Input quality and quantity have been shown to be essential for the L2 pronunciation learning outcomes (e.g., [3, 4]). However, they do not always seem to be enough for L2 pronunciation development [5]. As a consequence, it has been argued in previous research that learners benefit from activities that bring the target pronunciation structure into the learners' attention through consciousness-raising, which can be achieved with explicit pronunciation instruction or phonetic training [6, 7]. Consciousness-raising aims at the learner's noticing of the target feature. Noticing can be divided into noticing the form (e.g., becoming aware of quality differences between the English /i-ɪ/) and noticing the gap between one's production and the target production [8]. Noticing the gap can occur spontaneously or through corrective feedback. Even though previous studies have indicated that noticing the gap is beneficial [9], earlier research also demonstrates that L2 learners have difficulties in noticing when their pronunciation is being corrected (e.g. [10]). The objective of our study was to examine the noticing of gap of Finnish learners of English after attending one semester course on English phonetics and phonology.

II. Data and methods

The participants of the study were 34 L1 Finnish university-level learners of English who can be considered highly proficient learners and users of English (CEFR level B2-C2, as shown by the LexTALE vocabulary test). The participants attended a 12-week long practical course in English phonetics and phonology, which focused on segmental and suprasegmental pronunciation features in teaching groups using either British or American English sound system as the model. At the beginning of the semester the participants read a list of CVC words containing difficult segments for Finnish EFL learners [11]: voiced plosives in initial and final position, voiceless initial plosives, and /i-ɪ/. At the end of the course, the participants completed a 'thinking about your pronunciation' activity. In this activity, the participants listened to their recordings and marked in a form whether their pronunciation of the given sound was accurate or not. Sounds marked as 'inaccurate' were taken as instances of noticing the gap between one's production and the target form. In order to increase noticing, each word was played three times and the participants were asked to focus on one sound segment at a time. The participants could additionally elaborate on the deviations they had noticed. The sound segments were analysed by pronunciation experts to confirm the accuracy of the productions and the instances of pronunciation deviations. Noticing scores were calculated by comparing the pronunciation deviations indicated by the pronunciation experts with the participant's verbalizations of inaccuracy. The data were analyzed quantitatively and qualitatively.

III. Results and discussion

An analysis of the accuracy data indicated that, overall, the participants produced the target sounds with high accuracy (*M*=90.45, *SD*=6.70), even though individual variation was observed, with accuracy in some categories for some participants being only 37.5% and for others 100% across categories. A one-way repeated measures ANOVA was conducted to compare the accuracy in initial consonants, final consonants and vowels. There was a significant effect for target sound [*F*(2,27)=8.65, *p*<.001, partial eta squared =.39.] Post hoc analysis with a Bonferroni adjustment revealed that accuracy in final consonants was significantly higher than in initial consonants (*p*<.001) or vowels (*p*=.036), but that the accuracy between initial consonants and vowels was not statistically different (*p*>.05). As for the noticing, the results indicate that, although the learners were in the possession of the necessary metalanguage due to their studies and the explicit nature of the task, the learners only reported noticing on average 25.91% (range: 0-100) of the target sounds. Taking a closer look at noticing, two patterns emerged: over-sensitivity and under-sensitivity. The first pattern referred to participants who identified pronunciation deviations in segments that were produced accurately. The second pattern referred to participants who failed to identify a pronunciation deviation when it had occurred (failure to notice the gap, reflected in the noticing score). A closer look at the oversensitivity data indicated large individual variation in this phenomenon: range: 0-22.22. A medium positive correlation was observed between participant's L2 proficiency level and the noticing score so that participants with higher L2 proficiency noticed more segmental deviations (*r*=.370, *p*<.0.5, *n*=29). When interpreting the results, it has to be borne in mind, that the participants' overall production accuracy was very high. In our presentation, we will discuss the findings from a pedagogical perspective, looking at both lack of noticing the gap and oversensitivity, and consider the role of phonological self-awareness for L2 speech learning in instructed learning contexts.

References

[1] T. M. Derwing, “Utopian goals for pronunciation teaching,” Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference, pp. 24-37, 2010.

[2] A. Henderson, D. Frost, E. Tergujeff, A. Kautzsch, D. Murphy, A. Kirkova-Naskova, E. Waniek-Klimczak, D. Levey, U. Cunnigham, and L. Curnick, “The English pronunciation teaching in Europe survey: Selected results,” Research in Language, vol. 10(1), pp. 5–27, 2012.

[3] J. E. Flege and O.-S. Bohn, “The revised Speech Learning Model (SLM-r)”. In Second language speech learning: Theoretical and impirical progress, R. Wayland, Ed., Cambridge: Cambridge University Press, 2021, pp. 3–83.

[4] T. Piske, I. R. A. MacKay, and J. E. Flege, “Factors affecting degree of foreign accent in an L2: A review,” Journal of Phonetics, vol. 29(2), pp. 191–215, 2001.

[5] E. M. Ingvalson, J. L. McClelland, and L. L. Holt, “Predicting native English-like performance by native Japanese speakers,” Journal of Phonetics, vol. 39(4), pp. 571–584, 2011.

[6] R. Lyster and K. Saito, “Oral feedback in classroom SLA: A meta-analysis,” Studies in Second Language Acquisition, vol. 32(2), pp. 265–302, 2010.

[7] R. I. Thomson and T. M. Derwing, “The effectiveness of L2 pronunciation instruction: A narrative review,” Applied Linguistics, vol. 36(3), pp. 326–344, 2015.

[8] R. Schmidt, “Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning.” In Attention and awareness in foreign language learning, R. Schmidt, Ed., Hawai’i: University of Hawai’i Press, 1995, pp. 1–64.

[9] W. Baker and P. Trofimovich, “Perceptual paths to accurate production of L2 vowels: The role of individual differences,” International Review of Applied Linguistics in Language Teaching, vol. 44, pp. 231–250, 2006.

[10] K. Saito, “Communicative focus on second language phonetic form: Teaching Japanese learners to perceive and produce English /ɹ/ without explicit instruction,” Applied Psycholinguistics, vol. 36, pp. 377–409, 2015.

[11] P. Lintunen, Pronunciation and Phonemic Transcription: a study of advanced learners of English. Turku: University of Turku, 2004.

4th International Symposium on Applied Phonetics (ISAPh2022), September 14-16 2022, Lund University, Sweden

Articulatory and temporal properties of Estonian palatalization by Russian L1 speakers

Anton Malmi[1], Pärtel Lippus[1], Einar Meister[2]

[1]University of Tartu, Estonia

[2]Tallinn University of Technology, Estonia

Keywords — palatalization, electromagnetic articulography, segmental duration, Estonian, Russian accent, second language acquisition

I. INTRODUCTION

We aim to determine whether the articulatory and temporal properties of palatalization of Russian L1 learners of Estonian deviate from the native Estonian production. The motivation for the study came from the students who reported [1] that they are not comfortable speaking Estonian because of their accented speech. Phonetic research on Estonian foreign-accented speech has mostly dealt with the production and perception of Estonian ternary quantity contrasts and Estonian vowel categories but not with palatalization.

A consonant is palatalized when coarticulated with a neighboring front vowel or a glide /j/. In Russian, palatalization is very salient and most of the consonants have a palatalized and non-palatalized variant. In Estonian only four alveolar consonants /l, n, s, t/ can be phonologically and phonetically palatalized. Unlike Estonian, the cue for palatalization in Russian is marked in orthography. Word-initial consonants are not palatalized in Estonian but are in Russian. In both languages' palatalization is realized by simultaneously raising the tongue to the palate while the primary consonant constriction remains. The tongue is also more anterior and wider with palatalization [2]. When the trigger for palatalization follows the palatalized consonant, the quality of the preceding vowel is consistently affected [3]. Presumably, because of the palatalization gesture, the duration of the vowels is lengthened too with palatalization as the body of the tongue has to move up to the hard palate [4]. More research is needed to draw the attention of the language learners and teachers on the possibility that the incorrect usage of palatalization might intensify Russian-accented speech in Estonian.

Theories that explain the second language (L2) acquisition and the underlying reasons for accented speech converge on the idea that the native language (L1) sets the boundaries to the limits in perceiving and producing L2 because learners perceive sounds in L2 in accordance with their L1. Most relevant to the current study are PAM(-L2) - Perceptual Assimilation Model [7] and SLM(-r) - Speech Learning Model [8]. SLM(-r) predicts the ease of token-by-token acquisition, but PAM(-L2) deals with the assimilability of contrasts. SLM predicts that when a sound in L1 is similar to the sound in L2, the learners will have to create a new category or split their existing L1 category. If the sound in L2 is new, the learners will have to create a new category altogether. The assimilation pathways posited by PAM(-L2) are very similar to the *equivalence classification* proposed by SLM framework. PAM(-L2) posits many pathways but two are relevant to the current study: (1) two-category assimilation in which the learner maps two contrasting L2 sounds to two different L1 categories. This is the easiest type of assimilation for the learners, and they should be able to discriminate between the two L2 sounds easily. (2) Single-category assimilation in which two L2 sounds are assimilated to a single category in L1.

Thus, the research questions and hypothesis for the current study are as follows: What are the articulatory and temporal properties that describe Russian L1 Estonian L2 palatalization in contrastive word pairs and in *i*-stemmed nouns? Are these properties similar to or different from Estonian L1 speakers' productions? Recognizing palatalization from the text might be a problem for the learner because it is not marked in orthography in Estonian as it is in Russian. Based on the predictions of SLM(-r) and PAM(-L2), it can be hypothesized that if there are differences between native and nonnative productions, then those differences will reflect the phonetic details of the native language of the speaker. Palatalization is a "similar" category for Russian speakers for which "two-category" or "single-category" assimilation will likely occur. Thus, we hypothesize that the position of their tongue and any temporal cues will be different from native speakers' productions.

II.MATERIALS AND METHODS

The participants were asked to fill out a short questionnaire, which included questions about their demographic background, the place of birth of their parents, what kind of language was spoken in their kindergarten and school, and their assessment of their language proficiency on a 5-point scale. The Russian L1 group consisted of 24 speakers (17 female, 7 male, mean age 26, SD 8.7). All of them were born and had been living in Estonia. They reported that their native language is Russian, and they spoke Russian at home with their parents. The control group consisted of 21 native Estonian speakers (11 female, 10 female, mean age 28, SD 6.2). They all went to an Estonian kindergarten and school. The data analyzed in this study includes two sets of words: a) 11 minimal pairs of monosyllabic words, where the word-final consonant is either palatalized or not; b) 26 monosyllabic /i/-stemmed words in the nominative case with a word-final palatalized consonant.

The articulatory data were recorded with a Carstens AG501 electromagnetic articulograph. The sensor on the anteo-dorsum was used to estimate the height of the tongue. The sensor on the medio-dorsum was used to estimate the anteriority of the tongue. Tongue lateral sensors were used to calculate the width of the tongue by subtracting the data from one sensor from the other. All productions were manually checked for any incongruencies, and some of the data had to be discarded.

Statistical analysis was carried out with the R software [9]. Generalized Additive Mixed Model (GAMM) was used from *mgcv* package [10] to estimate the effect of palatalization, L1, consonant, and the preceding vowel on the height, width, and the anteriority of the tongue while producing vowels and consonants, and to compare the trajectories of Russian L1 speakers with Estonian L1 speakers. Segmental duration values were tested with Linear Mixed Model from *lme4* package [11]. In all of the models, a random intercept for the speaker was included.

III. Results and discussion

First, we looked at palatalization in phonologically contrastive pairs. Russian L1 speakers' tongue dorsum was higher and more anterior with palatalization. The width of the tongue of Russian speakers varied without a systematic pattern. When we compared their palatalized tokens to Estonian L1 speakers, we found that palatalization in the Russian L1 group was inconsistent. Only in some vocalic and consonant contexts were similar to native production. In many cases, the tongue in the Russian L1 group dorsum was lower and more posterior than in the control group. The width of the tongue was similar between the groups. This indicates that the width of the tongue of Estonian L1 speakers in our current study also varied without a systematic pattern. These differences between the groups can be explained by the fact that the proficiency of the participants in our study varied a lot even though they were all born and had been living in Estonia their whole lives. Their self-assessment of their Estonian proficiency was relatively high (average 4.1 on a 5-point scale) and even came close to the native speakers' self-assessment. It must be noted that we had to discard ~50% of the test words in word pairs because they were not produced as expected. The participants palatalized consonants that were not palatalized and vice-a-versa. Besides the phonologically distinctive word pairs, we also looked at how Russian L1 speakers articulate *i*-stemmed nouns where the final consonant should be palatalized, but palatalization does not distinguish a meaning. Compared to Estonian speakers, the tongue dorsum of Russian L1 speakers was lower when producing these consonants and the preceding vowels. However, the anteriority and the width of the tongue were similar. Russian L1 speakers probably do palatalize word-final consonants in *i*-stemmed nouns but with a lower tongue dorsum. The results from both contrastive word pairs and *i*-stemmed words suggest a pattern of language-specific articulatory settings [5], [6] that differentiate both languages. Because of that, the realization of their segments is different.

Although articulatory movements showed that Russian L1 speakers palatalized consonants somewhat similarly to native speakers, the duration of the vowels was not systematically longer. Consonants, on the other hand, showed some tendency to be longer with palatalization. When we compared their results to native speakers, we found that both in phonological pairs and *i*-stemmed nouns, the duration of palatalized consonants and the preceding vowels of Russian L1 speakers were significantly shorter. In Russian, the palatalized consonant is sometimes lengthened because of the aspiration that accompanies the consonant.

The results in the light of SLM(-r) and PAM(-L2) suggest that Russian L1 speakers' palatalized tokens are different from native speakers because they are not sufficiently sensitive to the fine acoustic and articulatory details. Although participants in the study have been living in Estonia for their whole lives, they still produce palatalization differently than native speakers. As all of the participants used Russian at home all of their lives, they are attuned to their native language in the production of palatalization, and they use different motor patterns to achieve the segmental goals. The results suggest that a two-category assimilation might have happened and that obtaining a similar contrast in another language can be a problem for the learner.

References

[1] A. Malmi and P. Lippus, “Russian L1 speakers’ palatalization in Estonian and the effect of phonetic speech training,” *Est. Pap. Appl. Linguist.*, vol. 17, pp. 211–230, 2021, doi: http://dx.doi.org/10.5128/ERYa17.12.

[2] E. Meister and S. Werner, “Comparing palatography patterns of Estonian consonants across time,” in *ICPhS Proceedings*, 2015.

[3] A. Malmi and P. Lippus, “Keele asend eesti palatalisatsioonis,” *Eesti ja soome-ugri keeleteaduse ajakiri. J. Est. Finno-Ugric Linguist.*, vol. 10, no. 1, pp. 105–128, Dec. 2019, doi: 10.12697/jeful.2019.10.1.06.

[4] M. Ordin, “Palatalization and Intrinsic Prosodic Vowel Features in Russian,” *Lang. Speech*, vol. 54, no. 4, pp. 547–568, Dec. 2011, doi: 10.1177/0023830911404962.

[5] B. Gick, I. Wilson, K. Koch, and C. Cook, “Language-specific articulatory settings: evidence from inter-utterance rest position,” *Phonetica*, vol. 61, no. 4, pp. 220–233, 2004, doi: 10.1159/000084159.

[6] I. Wilson and B. Gick, “Bilinguals use language-specific articulatory settings,” *J. Speech, Lang. Hear. Res.*, vol. 57, no. 2, pp. 361–373, 2013, doi: 10.1044/2013_JSLHR-S-12-0345.

[7] 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.

[8] 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.

[9] R Core Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria, 2021, Accessed: 29-Nov-2018. [Online]. Available: https://www.r-project.org/.

[10] S. N. Wood, *Generalized additive models. An introduction with R*, 2nd editio. New York: Chapman and Hall/CRC, 2017.

[11] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting Linear Mixed-Effects Models Using lme4.,” *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: doi:10.18637/jss.v067.i01.

The role of reflection and retention intervals on earwitness performance in voice parades

Kirsty McDougall[1], Nikolas Pautz[2], Francis Nolan[1], Katrin Müller-Johnson[3], Harriet M.J. Smith[2] and Alice Paver[1]
[1] University of Cambridge, {kem37|fjn1|aep58}@cam.ac.uk; [2] Nottingham Trent University, {nikolas.pautz|harriet.smith02}@ntu.ac.uk;
[3] University of Oxford, katrin.mueller-johnson@crim.ox.ac.uk

Keywords — earwitness, voice identification, memory, reflection, system variables

There are certain crimes that occur where the perpetrator is heard, but not seen. In such cases, the 'earwitness' may be asked by the police to undertake a voice parade, where the witness is asked to identify whether the voice of the perpetrator is present among a series of similar sounding voices. While the utility of voice parades in these situations is undeniable, those familiar with research focusing on voice parades will be aware that voice parade experiments generally shows that accuracy rates are low, particularly in lab-based parades which simulate an innocent suspect (i.e., where the target voice is not included, e.g., [1]). As the results and recommendations stemming from voice identification research undertaken in a lab can influence policy changes, it is vital that the procedures used in these lab-based experiments are ecologically valid, i.e., they should reflect the 'real world' as much as possible to reduce the risk of inappropriate policy changes being recommended.

In most lab-based voice parades, participants are first exposed to a voice that simulates a listener hearing a perpetrator speaking. Following this exposure, participants are either asked to come back at a later point in time [2] or complete a 'filler' task [1] which is designed to simulate the effects of a temporal gap between the processes of encoding and storage, and later retrieval of the memory of the voice. However, it is not too broad a leap to imagine that if the earwitness realizes that what they have overheard involves a crime, there is a good chance that they will think back upon what they heard after the fact. In other words, there would be a period of reflection post-encoding. There are grounds to believe that such a period of reflection may be effective when trying to recall a voice at a later stage for comparison. For instance, reflection may promote cue utilization [3]. The cue utilization perspective posits that judgements can be cued by an individual's theories about encoding conditions and processes [4]. By encouraging individuals to reflect on a voice they have heard, it may cue them to recall task-relevant details that could influence their response criterion (i.e., the predilection of the earwitness to respond that the target is present or not present). We propose that to maximise the ecological validity of lab-based voice parades, this post-encoding reflection needs to be examined using both a temporal retention period and the more commonly used filler task. In this study, we present the results of two experiments that manipulated the role of post-encoding reflection in earwitness performance.

Experiment 1 used a factorial two (target presence: present, absent) by two (reflection: reflection, no reflection) design. Three male speakers of Standard Southern British English were selected from the *DyViS* database [5] as target speakers. 9-speaker voice parades were constructed for each speaker using similar-sounding speakers from *DyViS* as foils, selected on the basis of multi-dimensional scaling of listener ratings of the similarity of the voices [6], so as to approach the lower bound of earwitness performance by using voices highly similar in accent and personal voice quality. The parades were constructed using mock police interview material (*DyViS* Task 1), following the UK Home Office guidelines [7] only with 15-second parade samples. This experiment was hosted online using Gorilla and participants were recruited using Prolific. Listeners (N=180) were randomly assigned to one of the three targets and exposed to a 60-second encoding sample taken from the near-end of a telephone call (*DyViS* Task 2). Half of the participants were assigned to a reflection condition where they were given the following instruction: "*Imagine that the voice you have just heard is that of a criminal. You may be asked by the police to make an identification some time in the future. Take a few moments now to reflect on the voice.*" The other half were assigned to the control condition, where they were given a simple task which required the listeners to press spacebar as soon as a fixation point appeared on their screens. In both the instruction and the no-instruction condition, an interval of approximately 20 seconds elapsed after which listeners completed a 5-minute word-search task with lobby noise. The listeners were then given instructions to undertake a voice parade. Half the listeners were randomly assigned to target-absent parades, and half were assigned to parades where the target voice was randomly positioned (1-9) within the parade. The results indicate that there were no meaningful differences in identification performance between listeners who had post-encoding reflection and those who did not; nor was there a meaningful interaction between target presence and reflection. We found that responses to target-present parades were more likely to be accurate than those to target-absent parades, corresponding with previous findings [1]. Importantly, listeners had above-chance accuracy in target-absent parades, but only in the reflection condition (see Fig. 1 which displays the most likely parameter values for the different conditions for both experiments).

Experiment 2 (N=181) followed the same design as used in the first experiment, except instead of a filler task, participants were given a 20–28-hour period between encoding and testing (on average, 21.5 hours). As mentioned, the filler task is meant to mimic the impact that a 'real' time gap between encoding, storage and later retrieval would have on memory strength. Thus, it is important that any effects found in a parade using a filler task can be replicated using a more ecologically valid retention interval. The results showed no meaningful differences in identification performance between listeners who were in the reflection or control condition, nor was there any interaction between target presence and reflection. There was no evidence that listeners in target-present parades were more likely to be accurate than listeners in target-absent parades; this is surprising as target presence is a relatively robust effect

[1; 2]. Importantly, unlike in Experiment 1, we did not observe above-chance levels for target-absent parades in the reflection condition.

The results of the two experiments provide tentative evidence that the inclusion of a reflection period facilitates a conservative criterion shift, reducing the chances of a listener making a positive identification when the target is in fact not present in the parade (i.e., a false alarm), but not at the expense of hits. This effect, however, was present only in parades which used a five-minute filler task (Experiment 1) and not when using the longer retention interval (Experiment 2). The pattern of responses suggests the results of voice parades using a filler task can contribute to our understanding of voice parades, but also suggests that any outcomes need to be tested using a more ecologically valid retention interval. Future research would benefit by focusing on manipulating both the length of the reflection period and the wording of the reflection instructions.

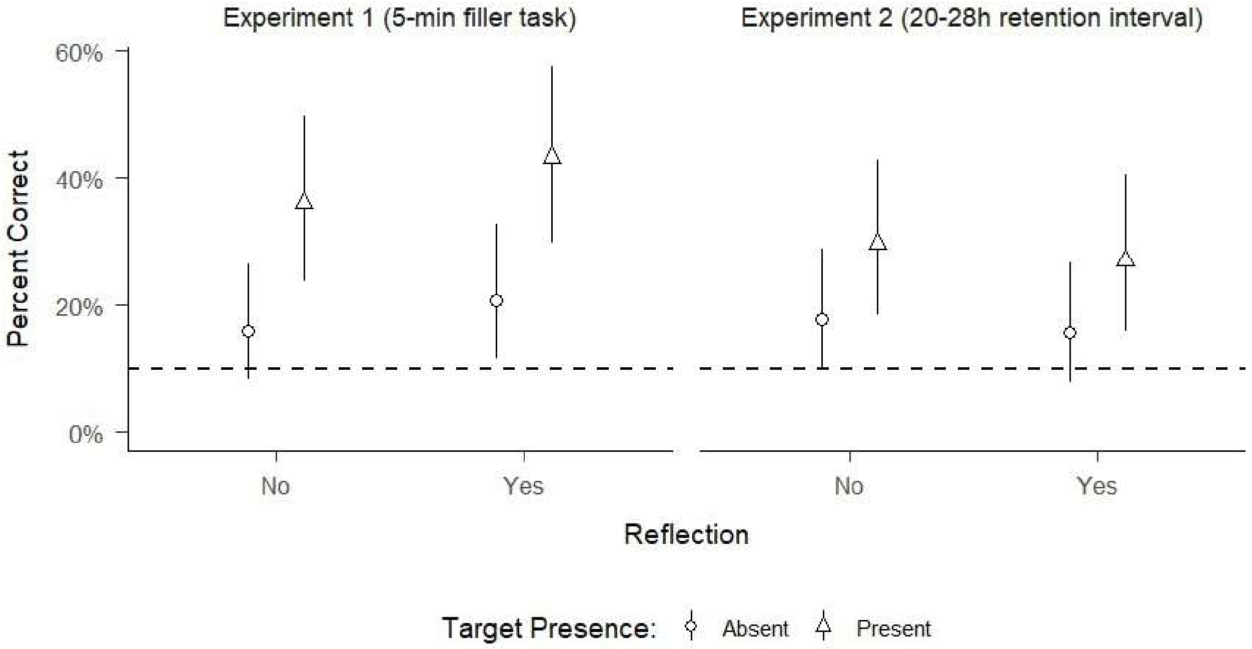


Fig 1. Point estimates and corresponding 95% Highest Density Interval (HDI) of accuracy extracted from the Bayesian logistic models.
Note. Dashed line represents chance level accuracy (10%)

REFERENCES

[1] Smith HM, Bird K, Roeser J, Robson J, Braber N, Wright D, Stacey PC. Voice parade procedures: optimising witness performance. Memory. 2020 Jan 2;28(1):2-17.

[2] McDougall K, Nolan F, Hudson T. Telephone transmission and earwitnesses: performance on voice parades controlled for voice similarity. Phonetica. 2015;72(4):257-72.

[3] Brewer N, Keast A, Rishworth A. The confidence-accuracy relationship in eyewitness identification: the effects of reflection and disconfirmation on correlation and calibration. Journal of Experimental Psychology: Applied. 2002 Mar;8(1):44.

[4] Koriat A. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. Journal of experimental psychology: General. 1997 Dec;126(4):349.

[5] Nolan F, McDougall K, De Jong G, Hudson T. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. International Journal of Speech Language and the Law. 2009 Sep 18;16(1):31-57.

[6] McDougall K. Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. International Journal of Speech, Language & the Law. 2013 Dec 1;20(2):163-172.

[7] Home Office. (2003) Advice on the use of voice identification parades. UK Home Office Circular 057/2003 from the Crime Reduction and Community Safety Group, Police Leadership and Powers Unit.

Ceiling effects and the limitations of intelligibility and accentedness ratings for advanced L2 English learners

Joaquín Romero, Universitat Rovira i Virgili, Tarragona, Spain

Keywords — ceiling effect, pronunciation assessment, ratings, intelligibility, accentedness

I. Introduction

Two main methods of pronunciation assessment have been proposed in the literature and used with varying success. The first relies on native rater judgments of notions like intelligibility and accentedness. A large body of work has proven the validity of this method in various contexts [1]. However, in certain situations the availability of raters is limited, which can introduce a high degree of variability in the ratings. The second method, acoustic/articulatory analysis, addresses the availability issue by removing raters from the process but at the expense of partly ignoring the communicative nature of language. A third approach, the use of technological tools for the automated assessment of pronunciation, while still not as widely used as the other two, has been gaining track in the last few years, as better, more reliable techniques are developed that could eventually replace or at least complement the subjectiveness of the human component [2].

On a different level, growing evidence suggests that native-like speech may not be the most realistic goal in many learning environments [3, 4]. For students training to become English teachers, however, achieving near-native pronunciation may still be a desirable goal, given the fact that, especially in non-immersive teaching situations where access to native speakers is very limited, teachers are often the main or only model that students have access to.

The current study investigates the assessment of L2 pronunciation in a group of L1 Spanish/Catalan advanced learners of English. Subjects were five 3rd year university students majoring in English and with a C1 or C2 overall language level. 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. Upon finishing their degree, many of these students were expected to go on to become English teachers at different levels of the educational system, with the largest numbers going to secondary schools as well as specialized language schools. Consequently, they showed great interest in achieving a level of English pronunciation that could guarantee that they would be adequate models in their teaching careers. For that reason, these students were considered the ideal population for the study.

II. Method

Participants were asked to record a set of English sentences, as well as a short paragraph and an improvised explanation of a complex situation from a newspaper headline. The sentences and the paragraph were created to make sure that they contained instances of advanced features of English pronunciation, such as a variety of allophonic variations of stops (aspirated, unaspirated, unreleased, etc.), as well as other phonological processes such as flapping of /t/ and /d/, glottalization of /t/, and palatalization of /d+j/ and /s+j/ sequences across word boundaries. The recordings were obtained in a sound-proof booth using a hand-held digital recorder.

For the intelligibility and accentedness ratings, productions were evaluated, on the one hand, by five native speaker judges and, on the other hand, by two expert phoneticians who were familiar with the phonetics and phonology course that the students had taken. In both cases the raters listened to the sentences by the five subjects as presented randomly and were asked to score them on two 5-item Likert scales, one aimed at assessing intelligibility and the other one accentedness.

In addition, an acoustic analysis was performed on two aspects, voice onset time as a measure of stop allophonic variation, and duration and relative intensity of the constriction associated with flapped /t/. The relative intensity was calculated as a difference between the intensity values of the flapped consonant and the following vowel.

III. Results

Preliminary findings show that, while the expert phonetician ratings were largely in agreement with the results of the acoustic analysis, showing a significant degree of speaker variability in their successful productions of stop allophonic variation and flapping, the native speaker ratings consistently evaluated the productions with very high scores, with little or no significant differences across subjects or raters.

These results can be indicative of the existence of a ceiling effect in the evaluation by the native speakers, who did not find the subtle differences in advanced features such as aspiration and flapping to be relevant in their evaluation of the subjects' productions. Though these features may in reality not be particularly important for effective communication in the L2, they are indicators of a high level of proficiency which could be of interest to future language teachers who strive to become accurate language models.

While still incomplete, the results seem to point to the need to reassess or fine-tune the use of intelligibility and/or accentedness ratings for the evaluation of L2 pronunciation, whether by making these constructs more precise or selecting raters to fit specific requirements, in order to accommodate the needs and goals of highly proficient learners.

REFERENCES

[1] M. J. Munro and T. M. Derwing, "The foundations of accent and intelligibility in pronunciation research," Language Teaching 44(03): 316-327, 2011

[2] A. van Moere and M. Suzuki, "Using Speech Processing Technology in Assessing Pronunciation," in O. Kang and A. Ginther (eds) Assessment in Second Language Pronunciation, Routledge, 2017.

[3] J. M. Murphy, "Intelligible, comprehensible, non-native models in ESL/EFL pronunciation teaching, ", System 42: 258-269, 2013.

[4] H. Bøhn and T. Hansen, "Assessing Pronunciation in an EFL Context: Teachers' Orientations towards Nativeness and Intelligibility," Language Assessment Quarterly, 14:1, 54-68, 2017.

4th International Symposium on Applied Phonetics (ISAPh2022), September 14-16 2022, Lund University, Sweden

Individual Voice Recognition Skills in Lay Speaker Identification Tasks

Sascha Schäfer, and Paul Foulkes
University of York, UK
sascha.schaefer|paul.foulkes@york.ac.uk

Keywords — forensic phonetics, lay speaker identification, voice parades, earwitnesses

I. Introduction

In many criminal cases investigators rely on the testimony of witnesses who have heard rather than seen a crime in order to establish the identity of the perpetrator. These 'earwitnesses' may be invited to take part in a voice parade (VP). In contrast to their well-established visual counterpart, however, auditory identification parades are rare because they are costly, being difficult to design, implement and interpret. It has therefore been a recurrent goal of both phonetic and psychological research to find ways of eliciting more reliable testimony from earwitnesses and to ultimately make VPs more viable.

A considerable amount of research has been undertaken to standardise and improve the VP procedure by e.g., optimising the quality [1], duration [1], [2], and presentation of the stimuli [3]. While these studies focus on variables that can be influenced by the investigator who sets up a VP ("system variables"), the present study aims to complement these findings with an analysis of inter-listener differences ("estimator variables"), which cannot be influenced by the investigator. The underlying assumption is that untrained listeners differ markedly in their abilities to recognise voices and might therefore not be equally suited for a standardised VP. Psychological tests, above all the *Bangor Voice Matching Test (BVMT)* [4] and the *Glasgow Voice Memory Test (GVMT)* [5], have already shown a wide range of listener performance ranging from developmental phonagnosia to 'super recognition'. Whether these results apply to earwitnesses is questionable, however, as the stimuli used in these studies were created from isolated vowels (GVMT) or syllables (BVMT), rather than naturalistic speech.

The present study addresses this problem. It is the first in a series of experiments intended to characterise the role of the listener in lay speaker identification tasks. Like the *BVMT*, it focuses on the immediate voice recognition skills of the listener, i.e. excluding memory processes as far as possible.

II. Methodology

A. *Stimulus Selection*

The stimuli for this experiment were taken from task 1 of the DyVis corpus [6], which consists of mock police interviews with 100 speakers of Standard Southern British English, aged 18 to 25. This dataset was chosen because the foil recordings used for the construction of a VP are usually taken from police interviews of unconnected cases [1]. Moreover, the relative homogeneity of speakers allowed control of language-related features of voice, such as accent, ensuring that the identification is predominantly informed by biological and habitual features of voice, such as f0, voice quality, and articulation rate. An expert constructing a VP would aim to create a similar type of variability. Some recordings were not considered for the stimuli, because they exhibited different recording conditions ($n = 3$), a different regional accent ($n = 1$), or a possible speech disorder ($n = 8$).

Forty-eight out of the remaining 88 recordings were used in this experiment (the other 40 recordings were used for further experiments within this study). Two 10s-long extracts were selected for each of the 48 speakers, one for the first exposure of the participant to the voice and one for subsequent identification. All 96 extracts were converted from stereo to mono and amplitude normalised to ensure that participants focused on the voice rather than the characteristics of the recording. Noise reduction was performed where necessary.

B. *The Task*

One hundred British participants (50 male, mean age = 36, SD = 13.8) took part in an AX discrimination task hosted on *Pavlovia*. For the stimuli, 96 pairs of various difficulty were created from the 96 extracts. Difficulty was determined by the average f0 difference between the extracts in a pair (range = 0 - 20Hz, mean = 5Hz). The parameter was chosen as previous studies have shown that f0 plays a pivotal role in judgments of voice similarity [3]. Three stimulus lists were created, that were deemed equally difficult. This measure was taken to avoid the possibility that significant results would be obtained by chance. Each list contained 16 same-speaker pairs and 16 different-speaker pairs. In each list the pairs were sampled from all 48 DyVis speakers.

Participants were assigned one of the three stimulus lists. They were asked to provide a same/different rating for each pair, the reaction time of which was also recorded via the *Pavlovia* software. Participants also reported their confidence on a 6-point Likert scale at the end of the experiment. It was hypothesised that participant performance would (1) be stable across the three experimental groups, but (2) vary significantly within each group.

III. Preliminary Findings

Signal Detection Theory (SDT) indices were calculated for all participants, taking into consideration *hits* (a same-speaker pair was correctly classified as such)*, false alarms* (a different-speaker pair was wrongly classified as same-speaker)*, misses* (a same-speaker pair was classified as different), and *correct rejections* (a different-speaker pair was correctly classified). Further relevant measures include the *percent correct* (PC), which is a combined measure of hit rates and correct rejection rates, and *d' (d prime)*, computed as the difference between standardised hit rates and false alarms. The d' value reflects the individual participant's sensitivity as a trade-off between hits and false alarms. In this connection, a d' value of 0 is equivalent to chance level while a value close to 3 indicates perfect discriminability. A negative d' score means that false alarms exceed the number of hits. The BVMT, which is also an AX discrimination task, assessed participant performance based on the PC alone, meaning that individual performance can only be assessed against the performance of the reference population. The d' score was therefore deliberately chosen in this experiment as it is reflective of participant behaviour rather than participant performance.

Results are summarised in Table I. The test produced an average PC score of 75, which is exactly halfway between performance at chance level and perfect performance, and a first indication that f0 was an adequate predictor of the test's difficulty. Both the *Jarque-Bera Test* and the *Shapiro-Wilk Test* suggested a normal distribution of PC scores. In support of hypothesis (1), participant performance was similar across the three experimental groups, with mean PCs of 75.4, 75.5 and 74.1, respectively. As predicted by hypothesis (2), the results also demonstrated a wide range of participant performance, including two "super recognisers" (>= 2SDs above the mean), and four participants at the opposite end of the spectrum (<= 2 SDs below the mean)[5]. A ceiling effect was avoided, in contrast to the GVMT, in which several participants produced a PC of 100. Participant behaviour produced a wide range of d' scores between 0 and 2.9. While the upper margin demonstrates high discriminability and is comparable to the GVMT's maximal d' value of 3.1, discriminability never dropped below chance level, as observed in the GVMT (min. d' = -0.7) [5]. Formal analysis of variation in the results correlating with demographic categories, reaction time and confidence is ongoing, but preliminary results suggest that earwitness behaviour varies greatly and that a criterial norm for this behaviour can be established via the d' index.

Table I. Summary Statistics of the present experiment

Voice Recognition Test, N = 100 (50 male)					
	Min.	*Max.*	*Mean*	*Median*	*SD*
Participant Age	18	68	36.0	38.9	13.8
PC	50.0	93.8	75.0	75.0	9.1
d'	0.0	2.9	1.4	1.4	0.6

IV. Outlook

The 100 participants who took part in this experiment were invited to two further experiments focusing on the memorisation abilities of lay listeners as well as task awareness, which is intended to shed light on the interactions between these different estimator variables. While certain characteristics of a crime scene, such as the psychological impact of the crime on the witness, are not addressed here, these controlled tests are a steppingstone to establishing credibility in the individual witness rather than a procedure.

References

[1] K. McDougall, "Ear-catching versus eye-catching? Some developments and current challenges in earwitness identification evidence," 2021. [Online]. Available: https://www.phonetics.mmll.cam.ac.uk/ivip/

[2] H. M. J. Smith, T. S. Baguley, J. Robson, A. K. Dunn, and P. C. Stacey, "Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance," *Applied Cognitive Psychology*, vol. 33, no. 2, pp. 272–287, Mar. 2019, doi: 10.1002/acp.3478.

[3] H. M. J. Smith *et al.*, "Voice parade procedures: optimising witness performance," *Memory*, vol. 28, no. 1, pp. 2–17, Jan. 2020, doi: 10.1080/09658211.2019.1673427.

[4] C. Mühl, O. Sheil, L. Jarutytė, and P. E. G. Bestelmeyer, "The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability," *Behavior Research Methods*, vol. 50, no. 6, pp. 2184–2192, Dec. 2018, doi: 10.3758/s13428-017-0985-4.

[5] V. Aglieri, R. Watson, C. Pernet, M. Latinus, L. Garrido, and P. Belin, "The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices," *Behavior Research Methods*, vol. 49, no. 1, pp. 97–110, Feb. 2017, doi: 10.3758/s13428-015-0689-6.

[6] F. Nolan, K. McDougall, G. de Jong, and T. Hudson, "The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research," *International Journal of Speech Language and the Law*, vol. 16, no. 1, pp. 31–57, Sep. 2009, doi: 10.1558/ijsll.v16i1.31.

The role of dialectology in L2 vowel acquisition; evidence from Mandarin Chinese

Robert Squizzero, University of Washington

Keywords — vowels, acoustics, vowel distance and dynamics, dialectology, Chinese

I. Introduction

This talk reports on phonetic differences in production of Mandarin vowels between first language (L1) Mandarin Chinese speakers and L1 American English, second language (L2) Mandarin speakers. The investigation focuses on the vowels /i u/, which appear to be similar between the two languages, but which, in fact, differ significantly.

Dialectal and sociolectal differences both in learners' L1s and in their target L2s are often overlooked in second language pronunciation research and teaching. While a phonemic approach to language learning and teaching has clear benefits in terms of simplicity, sub-phonemic differences between language varieties, such as dialectal differences, can prevent the faithful acquisition of L2 sounds, adversely affecting speaker intelligibility [1] and accentedness [2].

Part of the reason why sub-phonemic differences are overlooked is because of the uncritical use of traditional transcription symbols of L1 and L2 phonemes. One widely cited description of American English lists its vowel inventory as /i ɪ e ɛ æ ɑ ɔ o ʊ u ʌ ɝ/ [3], and one widely cited description of Mandarin lists its monophthongal vowel inventory as /i y u ə ɤ a/ [4]. Based on these descriptions, one could reasonably conclude that /i u/ are the phonemic monophthongs common to both languages. The phonetic realizations of phonemic vowels, however, often differ from the symbols chosen to represent them, due to regional and social differences affecting language varieties, as well as the passage of time. But conceiving of sounds in terms of the phonetic values assigned to traditional transcription symbols instead of the phonetic values present both in the L1 of the learner and the target variety of the L2 risks the obscuring of salient sub-phonemic differences.

Two such differences are relevant in comparisons of 1) /i/ preceding /ŋ/ and 2) /u/ in all phonological environments. In most dialects of Mandarin, including Beijing, /iŋ/ is realized as [i^{ə}ŋ], even though the presence of a schwa offglide is considered by some to be nonstandard or uneducated [5]. Regarding the second comparison, in many, if not most dialects of American English, /u/ is realized as [ʉ] [6]. In terms of acoustic backness, [ʉ] is in an intermediate position between Mandarin /u/ and /y/, which are typically realized in their cardinal positions [4]. Yet L2 Mandarin curricula for English learners generally include only /y/ and not /u/, potentially resulting in an unfaithfully fronted pronunciation of /u/ and less distance between these two high rounded vowels in the L2. Additionally, the common production of Mandarin /iŋ/ as [i^{ə}ŋ] is also often omitted from L2 Mandarin curricula, despite its possible perceptual relevance for L2 listeners.

II. Methods and Materials

Data come from a corpus of 44.1 kHz, 24-bit recordings made in a sound-attenuated studio at the Ohio State University [7]. The corpus consists of recordings of 10 L1 Mandarin speakers who were born and raised in Beijing and 21 American L1 English, L2 Mandarin speakers studying Chinese as a Foreign Language at the Ohio State University. Of the L2 speakers, 11 were of advanced proficiency and 10 were of intermediate proficiency. Speakers read three repetitions of 24 target sentences embedded in conversational scenarios. Recordings were automatically word and phone-aligned, and the first, second, and third formants of each vowel were measured at 20 equally spaced time intervals. Outliers were detected using Mahalanobis distance and were hand-corrected and re-measured prior to normalization.

III. Results and Discussion

Generalized additive mixed models were constructed for each vowel and formant to examine both their normalized formant values and their time-varying properties. Each model included F1 or F2 as the dependent variable, a fixed parametric term for L1/L2 status, a fixed smooth term for temporal measurement point, an interaction smooth term for measurement point and L1/L2 status, and a random smooth term for speaker.

Significant effects were observed both for F2 of /i/ preceding /ŋ/ and for /u/. The significant interaction smooth term for /i/, shown in Table I, indicates an overall difference in the shape of the trajectory of the vowel based on L1/L2 status. The difference in shape can be observed in the left panel of Fig. 1, where F2 of /i/ rises for both L1 and L2 speakers, but then decreases towards the end of the vowel's duration for L1 speakers (bottom trajectory) only. The significant fixed parametric term for /u/, shown in Table II, indicates an overall difference in the value of F1 or F2 of a vowel based on L1/L2 status. The difference in backness can be observed in the right panel of Fig. 2, where there is a visible gap between the L2 speakers (top trajectory) and the L1 speakers.

The results for /i/ and /u/ suggest that significant sub-phonemic differences exist between L1 and L2 Mandarin speakers despite the ostensible similarity of the two vowels between Mandarin and American English. Given that such differences can interfere with speakers' intelligibility and result in increased perceived accentedness, teachers of Mandarin pronunciation should consider adjusting their curricula to teach pronunciation of /u/ and of /iŋ/. More importantly, the results of this study demonstrate that second language learners and teachers would benefit from taking a critical view of the sounds that differ between the learners' L1s and L2s, a view that goes beyond simple comparison of phonemic inventories.

TABLE I. GENERALIZED ADDITIVE MIXED MODEL FOR F2 OF /i/ before /ŋ/ BY L1/L2 STATUS - MODEL COEFFICIENTS

Parametric coefficients:				
	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	1.79532	0.05347	39.653	< 2e-16 ***
L2	0.03647	0.05518	0.661	0.509
Approximate significance of smooth terms:				
	edf	*Ref.df*	*F*	*p-value*
s(Measurement point)	2.795	2.96	25.814	< 2e-16 ***
s(Measurement point):L2	1.00	1.00	5.484	0.0192*
s(Measurement point,speaker)	36.005	120.000	5.858	< 2e-16 ***

TABLE II. GENERALIZED ADDITIVE MIXED MODEL FOR F2 OF /u/ BY L1/L2 STATUS - MODEL COEFFICIENTS

Parametric coefficients:				
	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	0.75590	0.02117	35.698	< 2e-16 ***
L2	0.11361	0.02594	4.379	1.23e-05 ***
Approximate significance of smooth terms:				
	edf	*Ref.df*	*F*	*p-value*
s(Measurement point)	2.891	2.976	37.690	< 2e-16 ***
s(Measurement point):L2	1.822	2.184	2.074	0.11
s(Measurement point,speaker)	25.562	120.000	1.819	< 2e-16 ***

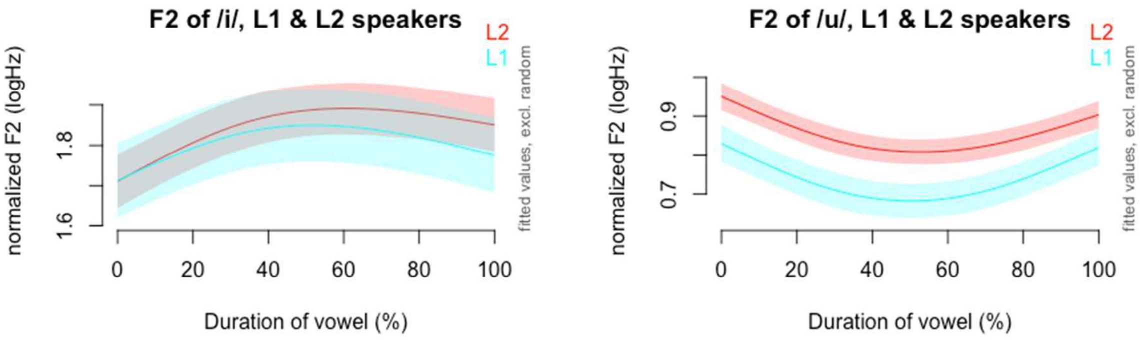


Fig. 3. Predicted F2 trajectories for /i/ preceding /ŋ/ (left panel), and for /u/ (right panel), with 95% confidence intervals (top trajectories: L2 speakers). Both panels are based on generalized additive models with L1/L2 status as both a fixed parametric term and an interaction smooth term with the temporal measurement point.

REFERENCES

[1] V. Porretta & B. V. Tucker. "Intelligibility of foreign-accented words: Acoustic distances and gradient foreign accentedness." ICPhS 18 Proceedings, 2015.

[2] E. A. McCullough. Acoustic correlates of perceived foreign accent in non-native English. Doctoral dissertation. The Ohio State University, 2013.

[3] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler. "Acoustic characteristics of American English vowels." JASA, vol. 95:7, pp. 3099-3111, 1995.

[4] W. Lee and E. Zee. "Standard Chinese (Beijing)." JIPA, vol. 33:1, pp. 109-112, 2003.

[5] Li, C. W.-C. "Conflicting notions of language purity: The interplay of archaising, ethnographic, reformist, elitist and xenophobic purism in the perception of Standard Chinese." *Language and Communication*, *24*(2), 97–133, 2004

[6] W. Labov, S. Ash, and C. Boberg. The Atlas of North American English. Berlin: De Gruyter. 2006.

[7] C. Yang. The Acquisition of Mandarin Prosody by American Learners of Chinese as a Foreign Language (CFL). Doctoral dissertation. The Ohio State University, 2011.

Role of Within Vowel Formants in Forensic Speaker Comparison: A Study Based on Vowels in Marwari as Spoken in Bikaner

Nikita Suthar[1], Peter French[2]

[1,2]Department of Language and Linguistic Science, University of York, United Kingdom

Keywords —formants, vowels, forensic speaker discrimination, within formant features, discriminant analysis

I. Introduction

It has been a goal of forensic research to assess the potential of the human voice for identifying a person. Formant analysis has been used as one of several methods for speaker discriminant studies[[1]- [3]. Most studies have focussed only on formant centre frequencies, trajectories and/or, to a more limited extent, bandwidths [4]- [6]. The current work takes this method further by adding ***within-formant*** measures to help increase the accuracy of formant analysis-based speaker discrimination. It reports on work conducted on formant bandwidth, centre of gravity, skewness, and kurtosis of the formant peaks as speaker discriminants. A total of forty-five female Marwari (Indo-Aryan language) monolingual speakers of three different caste dialects (Bishnoi, Jaat, and Brahmin – 15 speakers per variety) from the Bikaner district (Rajasthan, India) were recruited for the study. For the part of the research reported here, the speakers were treated as a single undifferentiated group. The recordings were collected from spontaneous and non-spontaneous speech and focused on eight different vowels. Three modes of data collection were employed. The first mode of data collection was a list of 80 words (10 tokens per vowel) that the participants were asked to read aloud. The second mode was a picture description task, i.e., participants were shown a picture of local deities and were asked to narrate a story associated with the deity. The third method was a conversation where participants were paired and asked to have a natural conversation on a topic of their choice or choose a topic from a provided list.

II. Data Analysis

An ANOVA conducted in R showed a significant inter-variety and vowel difference in the Marwari language for wordlist and story data. Once these differences had been established, the goal was to look at individual speaker discrimination. Spectral measures of eight acoustic features were extracted from the first four formants. These acoustic features were the centre of gravity, relative amplitude, spectral bandwidth, LPC bandwidth, spectral peaks, skewness, kurtosis, and standard deviation. Both manual and automatic formant extractions were conducted with the help of a Praat script. To make this script more accurate, eight different settings were selected. The settings provided manual control over deciding how to extract the within-formant features from the centre frequencies. The feature extraction relied upon the amplitude drop (-3dB vs -1dB) and different smoothing settings for the harmonics. This process was crucial to minimising the number of errors produced by the Praat script. The settings producing the lowest error rate were selected for further analysis; for -3dB, the error rate was 23%, and for -1dB, it was 6.9%, with the smoothing of 300Hz, 500Hz, 600Hz and 700Hz for F1, F2, F3 and F4 respectively.

As a next step, two models were created, with the first model treating the choice of setting as a factor and the second model treating vowel and variety as factors. In both models, the participants were a random variable. Results showed that for every spectral measure, all three factors play a significant role in their own right and that combinations of these factors also yield interesting results for speaker discrimination. A discriminant analysis was conducted on the features extracted from every variety to predict the classification rate of these measures in identifying individual participants.

Table 1. A CLASSIFICATION RATE OF INDIVIDUAL FEATURES FOR THE CENTRE FREQUENCIES F1-F4

Acoustic Measures	**Classification Rates**	**Times Above Chance**	**Classification Rates**	**Times Above Chance**	**Classification Rates**	**Times Above Chance**
	Wordlist		*Story*		*Conversation*	
F1+F2+F3+F4	15%	6.5 times	11%	4.5 times	70%	10.6 times
Amplitude F1-F4	13%	5.5 times	13%	5.5 times	70%	10.6 times
Spec Peak F1-F4	13%	5.5 times	12%	5 times	66%	10 times
Spec BW F1-F4	9%	3.5 times	8%	3 times	60%	9 times
LPC BW F1-F4	11%	4.5 times	10%	4 times	56%	8.3 times
COG F1-F4	15%	6.5 times	12%	5 times	56%	8.3 times
SD F1-F4	9%	3.5 times	9%	3.5 times	60%	9 times
Skewness F1-F4	7%	3 times	7%	2.5 times	30%	4 times
Kurtosis F1-F4	8%	3 times	9%	3.5 times	16%	2 times

III. Results

The initial results showed that all the spectral measures increased the classification rates minimum of 2.5 times. Some features performed better at classifying participants than others. Table 1 presents the initial results of the discriminant analysis and the classification rates of the individual features for the wordlist, story and conversational data. The results show that the centre of gravity, amplitude and spectral peaks were the best performing features of both datasets.

One of the most interesting findings from these results is that amplitude and standard deviation gave the exact same classification rates for both datasets. Once the individual classification rates were acquired further subsets of the vowels were created. Fig 1 shows the CR analysis of different vowels for the wordlist data. As represented in Fig 1, some measures performed better than the others in classifying speakers for individual vowels. The number of occurrences of the centre of gravity was almost as high as the centre formant frequencies.

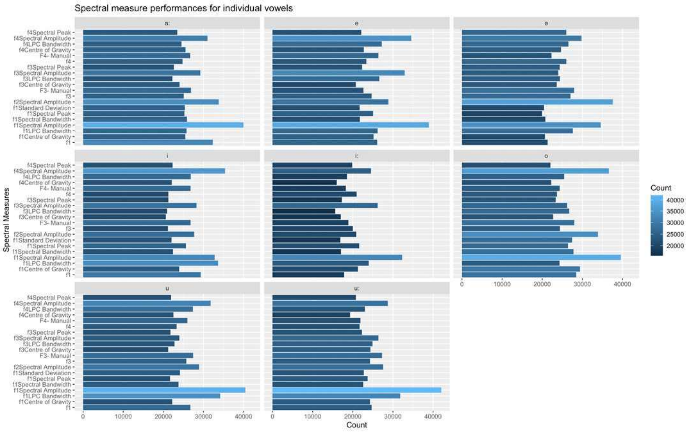


Fig. 1. *A classification rate of participants for different vowels for individual measures*

Individual CR analysis of the vowels showed a radical increase in the classification of the participants. Table 2 shows that the vowel /e/ was the best performing vowel, and /ə/ was the worst.

Table 2. A classification rate of individual Vowels for every feature together(times above chance)

Vowel	Wordlist	Story	Conversation
/a:/	31	24	11
/e/	22.5	39.5	13
/o/	28	33	15.6
/ɪ/	25	43	15.5
/i:/	30	24.5	15.6
/ə/	20.5	30	15.1
/ʊ/	23.5	33	15
/u:/	40	28.5	15

IV. Conclusion and Further Research

The analysis has shown that analysing within-formant measures along with already established features would increase the accuracy of the speaker discrimination studies. Further analysis on multiple combinations of the features also needs to be conducted to determine the possibility of identifying the best feature cluster that can be used in speaker discrimination studies. The same analysis will be further performed on the subsets of three different caste dialects to verify the accuracy of the results throughout these subcategories. As Marwari language was used as a testbed, the hypothesis will also be tested on other language data sets to verify the significance of language-independent significance of the features.

V. References

[1] T. Becker, M. Jessen, and C. Grigoras, “Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models,” 2008.

[2] H. Cao and V. Dellwo, “The role of the first five formants in three vowels of mandarin for forensic voice analysis”, doi: 10.5167/uzh-177494.

[3] P. Foulkes and P. French, “Forensic Speaker Comparison: A Linguistic-Acoustic Perspective,” in The Oxford Handbook of Language and Law, Oxford University Press, 2012. doi: 10.1093/oxfordhb/9780199572120.013.0041.

[4] R. D. Kent and H. K. Vorperian, “Static measurements of vowel formant frequencies and bandwidths: A review,” Journal of Communication Disorders, vol. 74. Elsevier Inc., pp. 74–97, Jul. 01, 2018. doi: 10.1016/j.jcomdis.2018.05.004.

[5] J. Gonzalez-Rodriguez, “Speaker recognition using temporal trajectories in linguistic units: the case of formant and formant-bandwidth contours.”

[6] M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, and D. Mürbe, “Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall,” Biomechanics and Modeling in Mechanobiology, vol. 14, no. 4, pp. 719–733, Aug. 2015, doi: 10.1007/s10237-014-0632-2.

Does accented speech affect attention and information retention?

Trisha Thomas[12], Gerard Llorach[7], Clara Martin[16], Sendy Caffarra[1345]

[1] Basque Center on Cognition, Brain and Language, San Sebastian, Spain, [2] University of the Basque Country (UPV/EHU) Bilbao, Spain, [3] University School of Medicine, Stanford, USA[4] Stanford University Graduate School of Education, USA, [5] University of Modena and Reggio Emilia, Italy
[6] Basque Foundation for Science (Ikerbasque), [7] University of Oldenburg, Germany

Keywords —accented speech perception, information retention, listening effort

I. Introduction

Millions of people interact on a daily basis with non-native speakers of their language. While we are increasingly likely to have conversations with foreign speakers, it has been shown that attending to foreign accented speech presents us with unique difficulties [1]. Accent can lead to differences in both attention and short-term memory representations [2][3]; however, little is known about the possible relationship between them. Do we struggle to attend to and memorize the details of a doctor's instructions when the doctor has an unfamiliar accent? Do we forget a lesson faster if taught by someone from abroad? This study aims to answer these kinds of questions by examining the effect of different accents (i.e., foreign, native, dialectal) on cognitive demands and subsequent information retention. Three different models have been proposed on the relationship between attention and memory within the context of accented speech:

- The Cognitive Load Theory hypothesizes that an attentionally demanding situation would lead to a drop in memorization performance [4].
- The Cognitive Effort Theory hypothesizes that an attentionally demanding situation would lead to better memorization performance [5][6].
- The Indexical Free Theory hypothesizes that accent is an attentionally demanding situation that should not affect memory performances [7].

To test these theories, we used a dual task paradigm where a primary (i.e., listening to and repeating spoken accented sentences) and a secondary task (i.e., digital circle tracking) were provided to the participants. This paradigm can assess attention through listening effort, which is a normalized measure of the cognitive effort required by a dual task as compared to a baseline (i.e., single task; see formula in (1) as described in [8]. To examine the effect of listening effort on information load and retention, participants were given a memory task, around 15 minutes after listening (session 1) and again 4 to 6 days later (session 2).

$$(Listening\ Effort = 100 \times [(Baseline - Dual\ Task) / Baseline]) \qquad (1)$$

For the dual task we predict that foreign accent should show a greater listening effort as compared to native accents, meaning that there would be an increased demand on attentional resources in the primary listening task, and a consequent decreased performance on the secondary task of circle tracking. For the memory task different predictions can be made based on the higher attentional demands of the foreign accent:

- The Cognitive Load Theory predicts poorer memory retention performance for foreign compared to native accents
- The Cognitive Effort Theory predicts better memory retention performance for foreign compared to native accents.
- The Indexical-Free Theory predicts no difference in the memory retention performance for foreign compared to native accents.

II. Methods

36 participants (22 Female, mean age=25.89, SD=4.48, range=19-37) listened to and repeated sentences in foreign, dialectal and native accents (British, Argentinean and Northern Spanish Spanish) while completing the secondary task of tracking a moving circle on the screen online. Participants performed a memory task immediately after the dual task (session 1), as well as in a delayed session (4 to 6 days later, session 2). In session 2, they also completed a subjective listening effort rating and a language background survey. Listening effort scores were obtained for the dual task. D' scores were calculated for the memory sessions.

III. Results

A. *Attention*

A two-way ANOVA with accent as a three-level factor (native, dialectal and foreign) and response type (listen and repeat) as a two-level factor was run on the tracking listening effort data. A significant main effect of response type was found (F(1,210)=4.12, p=0.04), where listening was associated with higher effort than repeating. No significant main effect of accent was found (F(2,210)=2.33, p=0.10) and no interaction between accent and response type (F(2,210)=0.41, p=0.66). Exploratory follow-up t-tests showed that foreign accent seemed to have the highest listening effort scores (foreign vs native: t(35)=1.85, p=.07; foreign vs dialectal: t(35)=2.26, p=.03; dialectal vs native: t(35)=1.40, p=0.17). On average, participants rated foreign accent as more difficult to pay attention to than native accents (Foreign:4.7 effort out of 10, sd: 2.2, Native:1.9, sd:1.9, Dialectal:3.4, sd:1.9). A repeated measure ANOVA was also run on these subjective listening effort measures. There was a main effect of accent (F(2,90)=13.76, p<0.0001), suggesting that foreign accent was perceived as more difficult than native (t(30)=5.61, p<0.0001) and dialectal accent (t(30)= 2.91,p=0.003).

B. Memory

A two-way ANOVA with accent as a three-level factor (native, dialectal, foreign) and session as a two-level factor (session 1, session 2) was conducted on the calculated d' scores of the memory trials. A significant effect of session was found, where the session 1 corresponded to higher d' scores overall than the session 2 (F(1,210)=76.85, p=6.31e-16). A significant effect of accent was also found (F(2,210)=3.93, p=.02), suggesting that foreign accent was better memorized than native accent (t(35)=3.28, p=0.002). There was no significant interaction between session and accent (F(2,210)=0.75, p=0.47). See Fig. 1.

C. Attention -Memory relationship

A regression model was fitted to test whether the relationship between listening effort and memory performance was affected by accent. A significant negative relationship between attention and memory was present in native accent (β =-4.97 , SE =2.14, p =0.02), but this effect disappeared for foreign (β = -1.42, SE = 1.34, p = 0.29) and dialectal accent (β = 0.07, SE = 2.03, p = 0.97).

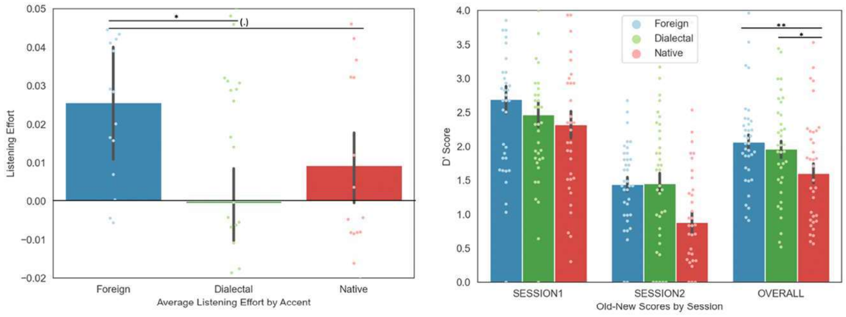


Fig.1. Average listening effort and d' scores for each of the accented conditions on session 1 and 2 of testing.

IV. Conclusions

A. Attention

Accent strongly influenced self-reported effort measures, showing that foreign accent had the highest perceived effort reports. However, listening effort measures were only mildly affected, with exploratory analyses showing the highest scores for foreign accent. This seems to suggest that foreign accent might require the highest cognitive effort among the accents examined here.

B. Memory

Foreign accent was also associated with better memory performances as compared to native accent.

C. Attention – Memory relationship

While in native accent the higher the listening effort the lower the memory performance is (in line with the Cognitive Load Theory), foreign accent seems to disrupt this attention-memory trade-off. Specifically, when dealing with foreign accent, high levels of listening effort do not imply lower memory performances. Instead, foreign accent has a beneficial effect on memory, which is at least partially in line with the Cognitive Effort Theory. Hence, the relationship between attention and memory does not seem to be fixed but rather may change as a function of accent.

References

[1] Lev-Ari, S. (2014). Comprehending non-native speakers: Theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*.

[2] Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology*, *30*(1), 117–139.

[3] Lev-Ari, S. (2017). Talking to fewer people leads to having more malleable linguistic representations. *PLOS ONE*, *12*(8), e0183593.

[4] Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. Educational Psychologist, 38(1), 63–71.

[5] Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95*(2), 419-425.

[6] Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. Journal of Experimental Psychology. Human Learning and Memory, 5(6), 607.

[7] Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. Frontiers in Human Neuroscience, 9.

[8] Kemper, S., Schmalzried, R., Herman, R., Leedahl, S., & Mohankumar, D. (2009). The effects of aging and dual task performance on language production. Aging, Neuropsychology, and Cognition, 16, 241-259.

Pretest-Posttest Production and Perception Results of Ultrasound Pronunciation Training

Noriko Yamane[1], Kunyang Sun[1], Jeremy Perkins[2], Ian Wilson[2]

[1] Hiroshima University, Japan, [2] University of Aizu, Japan

Keywords —ultrasound, pretest-posttest, English [l] and [ɹ], intelligibility, Japanese

I. Introduction

There is extensive research into the perception and production of English /l/ and /ɹ/ by Japanese native speakers. The phonemes /l/ and /ɹ/ are not used underlyingly in Japanese – instead /ɾ/ (or /ɽ/) is used, depending on the position in the syllable. However, these three sounds differ in their articulatory complexity: /ɹ/ has three gestures (tongue tip, tongue root, and lips), /l/ has two gestures (tongue tip and tongue dorsum), while /ɾ/ has only one gesture (tongue tip). Furthermore, articulation of these liquid consonants in syllable-coda position often includes a schwa-like element [1, 2], not found in L1 Japanese. Thus, Japanese native speakers often have difficulty acquiring these English sounds.

Previous studies have found that articulatory training by using ultrasound visual feedback is beneficial for learners acquiring /l/ and /ɹ/ [3, 4, 5]. The training in such research has usually occurred outside Japan, where learners are already somewhat immersed in English-speaking countries and can get continuous phonetic input. The cost of the ultrasound equipment is still too expensive for general use, but it's decreasing and may be affordable in the future. It remains to be answered whether Japanese university students in Japan can acquire the ability to distinguish English /l/ and /ɹ/ in their speech production after short training sessions, whether they can do so for both onset and coda liquids, and whether the contrast of tongue contours between /l/ and /ɹ/ corresponds to perceived intelligibility. This study investigates 1) to what extent Japanese university students are able to articulatorily distinguish English /l/ and /ɹ/ after a total of 3 hours of articulatory training with explicit instructions with the aid of tongue movies, 2) whether it is more challenging for them to acquire those sounds in onset or coda position, and 3) to what extent (and for what parts of the tongue) their articulatory contrast correlates with the intelligibility of these sounds for North American English listeners.

II. Method

A. *Participants and Procedure*

An instructor recruited participants from freshman English communication classes in a Japanese university. Six participants (one male; five female) were chosen on a first-come-first-serve basis. All participants were born and raised in Japan and had no experience living abroad for more than 2 weeks. Their English level was 'Independent User level' (equivalent to CEFR B1-B2, TOEIC average 657.5). Based on pre-questionnaires, none of them judged their own /l/ and /ɹ/ pronunciation to be perfect. None reported any auditory or visual disabilities.

The group training sessions and recordings were conducted at a language lab in fall-winter of 2019. Participants were instructed to come to 3 weekly group training sessions (1 hour each). Data was collected once pretest, once posttest, and once delayed posttest, and participants were instructed to complete pre- and post-questionnaires at home.

B. *Stimuli and Training*

The reading materials for practice and recording contained minimal pairs differing in /l/ versus /ɹ/. They were read in a carrier phrase. The pretest consisted of 46 sentences of 23 minimal pairs, while the posttest consisted of 69 sentences – the pretest list plus 23 novel sentences. Each participant read each sentence 14 times from a randomized list. The minimal pairs of words included those with /l/ or /ɹ/ in word-initial, word-medial, and word-final positions, as well as consonant clusters and challenging sequences. The first week of training focused on word-initial, the second week focused on word-medial, and the third week focused on word-final and challenging words. Participants received explicit instructions about the articulatory differences between /l/, /ɹ/, and the Japanese /ɾ/. Both groups' training included watching eNunciate videos (https://enunciate.arts.ubc.ca/), listening to the trainer's individual feedback, viewing ultrasound movies of target sounds produced by a Canadian English speaker, and observing peer training. The training and feedback focused on the movement of the tongue tip/blade and the tongue root and tongue bracing during the production of these sounds. While all participants saw ultrasound images of the native speaker's tongue, three of the participants (U-F1, U-F2, U-M1) also watched their own dynamic tongue shape by holding the probe on their own while being given feedback.

The tongue movies for all participants were recorded with ultrasound and saved on AAA [6]. Using the data, the tongue shape differences between /l/ and /ɹ/ in onset and coda position were analyzed using ultrasound imaging. Degree of differences in multiple regions (tongue tip, tongue body, and tongue root) were computed for pretest and posttest production results. In order to test whether the magnitude of the lingual difference would influence the intelligibility of those sounds, all sound files were presented in a random order in an intelligibility test taken by native English listeners on Amazon Mechanical Turk (AMT). Statistical analyses were performed in R to determine which production differences had the greatest effect on native listeners' accuracy in the intelligibility test.

III. Results

A t-test (computed using the built-in function in AAA) showed significant differences in specific lingual areas. Total improvement of each participant was shown in terms of the increased number of lingual areas of significance, the decrease in variability in tongue position, and the proximity of tongue shape to that of a model speaker.

The results indicate that four of the original participants successfully increased the articulatory differences between /l/ and /ɹ/ in both onset and coda, and the tongue tip/blade and tongue root were the main areas in which participants came to control more in posttest compared to pretest. This indicates that explicit articulatory learning can facilitate the accuracy of fine-grained control of sounds challenging to L2 learners, and ultrasound imaging is a potentially useful method for detecting the differences over tongue contours.

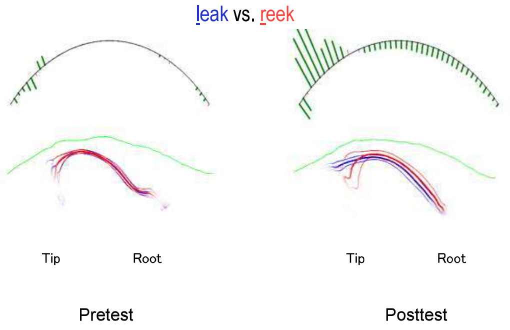


Fig. 1. Subject S-F3

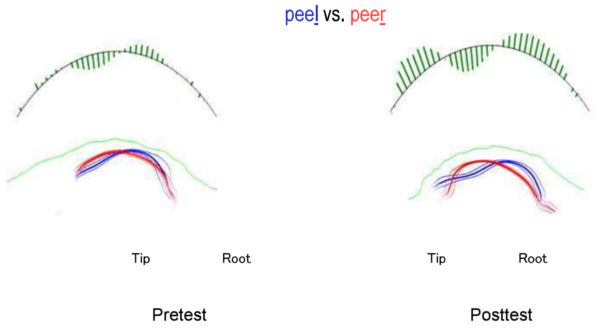


Fig. 2. Subject S-F2

The effect of phonological position varied among participants. For onset position, the largest pretest-posttest difference was observed in S-F3 (Fig. 1). For coda position, it was S-F2 (Fig. 2). Both participants did not see their own tongue while practicing, thus shadowing may have been sufficient to improve their pronunciation. In spite of the individual differences, overall, tongue tip in posttest showed more variability than tongue body, suggesting that the training was effective to facilitate the tongue tip to make a contrast between the sounds in question. Furthermore, for coda /l/, S-F2, U-F1, U-F2 and U-M1 showed raising/retraction of tongue dorsum after the training.

Based on the scoring system we developed, most participants (except S-F3) showed greater improvement in coda than in onset. Such an asymmetry was not expected, because coda position should be more complex in production. Perhaps the schwa-like quality or reduction of contact in the anterior portion of the palate in coda [7] also enhanced the perceptual difference with the Japanese flap, which facilitated the learning of the new category [8]. Due to space constraints, AMT results will be reported at the conference itself and in the proceedings paper.

References

[1] Gick, B., & Wilson, I. (2006). Excrescent schwa and vowel laxing: Cross-linguistic responses to conflicting articulatory targets. In L. Goldstein, D. H. Whalen, and C. T. Best, eds., *Laboratory Phonology* 8, pp. 635–659. Mouton de Gruyter, Berlin.

[2] Lawson, E., Leplatre, G., Stuart-Smith, J., & Scobbie, J. M. (2019). The effects of syllable and utterance position on tongue shape and gestural magnitude in /l/ and /r/. In *Proc. of the 19th International Congress of Phonetic Sciences (ICPhS)*. International Phonetic Association. 432–436.

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

[4] Tsui, H. M.-L. (2012). Ultrasound speech training for Japanese adults learning English as a second language (MSc thesis, University of British Columbia).

[5] Tateishi, M., & Winters, S. (2013). Does ultrasound training lead to improved perception of a non-native sound contrast? Evidence from Japanese learners of English. In *Proc. 2013 Annual Conference of the Canadian Linguistic Association* (pp. 1–15).

[6] Wrench, A. (2012). Articulate Assistant Advanced User Guide. Version 2.17.02. Edinburgh: Articulate Instruments Ltd.

[7] Kochetov, A. (2022, March). Production of English phonemic contrasts and allophony by Japanese learners: Electropalatographic evidence. Handout presented at Phonology Festa, Japan.

[8] Flege, J. E. (2016, June). The role of phonetic category formation in second language speech acquisition. In *8th International Conference on Second Language Speech (New Sounds 2016)*.

4th International Symposium on Applied Phonetics (ISAPh2022), September 14-16 2022, Lund University, Sweden

Patterns in the acquisition of /s/clusters in Spanish-English bilingual children with phonological disorders

Mehmet Yavaş Florida International University

Keywords: Spanish-English bilingual children, #sC clusters, Disordered phonology

English two-member initial #sC clusters[1] behave differently from other clusters. While non-/s/-clusters do not allow homorganicity (/pw, /dl/ are not allowed), some #sC clusters such as /st/, /sn/ do. While non /s/-clusters follow the Sonority Sequencing Principle (SSP) in that the sonority rises from the first consonant to the second (e.g. /pl/, /kr/), in some /sC/clusters such as /sp/. /st/ this is not the case. In addition to such structural differences from other clusters, data from acquisition studies also suggest that /sC/ clusters should be examined separately. Since Spanish does not have initial /s/-clusters, an investigation of Spanish-English bilingual children with phonological disorders provides us the opportunity to compare the patterns that are found in monolingual English-speaking children, as well as in typically developing Spanish-English bilingual children.

This study investigated the development of English initial two-member #sC clusters in 33 Spanish-English bilingual children (mean age 5;1) with phonological disorders. The objective was to determine whether sonority sequencing of the targets or the phonetic quality of C2 (the consonant following the initial /s/) could account for any sub-groupings, and if not, to find out what the governing patterns are in children's renditions of these targets. Data were collected via picture naming. The target words were #sC clusters with different combinations (/sp, st, sk, sm, sn, sl, sw/). Children's productions were analyzed with respect to their accuracy as well as their reduction patterns in incorrect renditions.

While a great deal of variability occurred both within and across children, certain general patterns emerged. Children showed higher accuracy on those targets in which C2 was [+continuant] (i.e. /sl/ or /sw/, whereby C1 and C2 with shared continuances) compared to the ones in which C2 was [-continuant] (i.e., /s+stop/, /s+nasal/ whereby C1 and C2 have opposite continuances). The patterns that are found in reductions were also, in general, supportive of the above stated binary grouping, whereby the preferred retained consonant was the C2 for '/s/+[-continuant]' targets, but was the C1 (i.e., /s/) for '/s/+ [+continuant]' targets. As such, results show many similarities to those found in another study with a English-speaking group of children with phonological disorders [1], and in typically developing Spanish-English bilingual children [2], as well as in monolingual English-speaking children [3]. The significance of the binary grouping by pulling stops and nasals (as C2) together (that is, [-continuant]) reiterates their common property of completely blocking the airflow through the center of the vocal tract, as opposed to '/sl, sw/ (i.e., '/s/+ [approximant]' clusters whereby C2 is [+continuant].

These targets are treated differently than the canonical clusters by various scholars in the literature. Accordingly, some researchers have proposed that */s/* is organized outside the onset constituent that contains the following consonant. Others have proposed that this sort of analysis holds only for a subset of sC clusters; those that rise in sonority, for example in *snake*, are represented in the same fashion as branching onsets. Yet others have argued that some sC clusters such as 's + stop' form complex segments. For a discussion of these issues, see [4].

REFERENCES

[1] M. Yavaş and S. McLeod, "Acquisition of /s/ clusters in English-speaking children with phonological disorders," Clinical Linguistics and Phonetics 24:3, pp.177-187. 2010.

[2] M. Yavaş and J. Barlow, "Acquisition of #sC clusters in Spanish-English bilingual children, "Journal of Multilingual Communication Disorders, 4, pp.182-193. 2006.

[3] M. Yavaş and C. Core, "Acquisition of #sC clusters in English speaking children," Journal of Multilingual Communication Disorders, 4, pp. 169-181. 2006

[4] H. Goad, " The representation of sC clusters," The Blackwell Companion to Phonology, M. van Oostendorp, C. Ewen, E. Hume & K. Rice (Eds.) pp.898-923. Oxford; Wiley-Blackwell. 2011.

4th International Symposium on Applied Phonetics (ISAPh2022), September 14-16 2022, Lund University, Sweden

Listening comprehension of World English pronunciation: How effective are awareness-raising activities?

Katsuya Yokomoto[1], Aki Tsunemoto[2], and Yui Suzukida[3]

[1] Center for Language Education and Research, Sophia University, [2] Department of Education, Concordia University,
[3] School of Medicine, Juntendo University

Keywords — World Englishes, awareness-raising activities, listening comprehension

I. Introduction

In the increasingly globalized world, in which English is used for communication by second language (L2) users more often than by first language (L1) users, the status and nature of English has diversified. As reflected in the terms World Englishes (WEs) or English as a lingua franca (ELF), efforts to appreciate the various features of English as used by L2 users for functional purposes have been made. According to this view, L2 users can learn and use regional varieties of English for successful communication rather than acquire so-called standard varieties of English. While WEs and ELF have been discussed widely in the field of sociolinguistics, they have rarely been discussed in the context of L2 teaching. To reflect the diversity of English, language education should incorporate understanding, awareness, literacy, and competence in the varieties of English [1]. Therefore, building on the evidence demonstrating the positive effects of incorporating WE instruction in raising learners' awareness of different varieties of English in language classes [2], the current study aimed to examine whether instruction actually facilitated learners' listening comprehension of different English varieties.

II. Literature review

According to speech recognition theories [3], the understanding of speech input relies on top-down and bottom-up processing. In particular, bottom-up processing (the identification of segmental and suprasegmental features) is essential for low-proficiency learners because they are unable to resort to top-down processing due to their lack of lexical knowledge, which suggests that the instructional priority in classroom settings should be familiarization with the *phonological* characteristics of WE pronunciation.

In actual communication, familiarity with the phonological features of the English variety used by the interlocutor is regarded as beneficial for listeners' effective speech comprehension. However, few classroom-level attempts have been made to increase the acceptance and use of WEs, particularly in the English as a foreign language (EFL) context. The available evidence has demonstrated that providing awareness-raising activities, such as increasing the exposure to different varieties of English via listening activities (e.g., letting learners listen to dialogues between L2 users) and learners' self-regulated learning inside and outside of classrooms succeeds in raising the learners' awareness and promoting favorable attitudes toward WEs and ELF [4]. Nonetheless, no study has yet explored how guided awareness-raising activities can improve learners' *listening comprehension* of WE varieties. Therefore, whether awareness-raising activities that focus on L2 users' pronunciation promote learners' listening comprehension remains an empirical question.

Accordingly, this study examined whether and the extent to which guided awareness-raising activities (particularly focusing on the phonological features of WEs) could promote accuracy in the bottom-up processing of WE varieties, thus helping L2 users to understand WE pronunciation. Specifically, the study focused on Japanese learners of English because they need to understand varieties of L2 English due to the increasing opportunities to engage in business dealings with people from non-Western countries, including China and Korea, who use ELF for communication [5].

III. Method

Forty-eight university students who were enrolled in a mandatory freshman English course at a university in Tokyo participated in this study, including a Chinese student, a Korean student, and 46 Japanese students; their proficiency level in English was B1 according to the Common European Framework of Reference.

A quasi-experimental design without a control group was employed. In the pre-test, the participants transcribed three speech samples of the Korean and three of the Chinese varieties that were played in random order. Five speech samples of the Korean and Chinese varieties were transcribed in the post-tests. All the speech samples were one-sentence extracts from the monologues available in the International Corpus Network of Asian Learners of English (ICNALE) [6].

The participants received a first treatment session focusing on the Korean varieties of English, followed by the Chinese varieties in the following week. Each week, the participants attended two 100-minute sessions in which they learned the linguistic characteristics of the varieties via guided awareness-raising activities (e.g., reading articles about the phonological characteristics of the target varieties). The objective of the sessions was to learn explicit knowledge about the target varieties; exposure to the authentic pronunciation of the varieties was limited to short video clips that introduced the phonological characteristics.

In the analysis, the score for the correctly transcribed words was calculated by dividing the number of correctly transcribed words by the total number of words following the conventional approach in L2 speech research [7]. A paired-samples *t*-test was administered to examine whether there was a statistically significant improvement in the scores for the correctly transcribed words; Cohen's *d* was calculated to examine the effect size.

IV. Results

Table 1 summarizes the results for the paired-samples *t*-tests. The results revealed a statistically significant improvement in the scores for the correctly transcribed words from the pre- to the post-tests for the Korean variety. After engaging in the guided awareness-raising activities, the students' transcriptions were, on average, 19.39% more accurate in the post-test compared to the pre-test results. Similarly, for the Chinese variety of English, the results revealed that there was a statistically significant improvement in the scores for the correctly transcribed words; the students' transcriptions in the post-test were 8.01% more accurate on average than they were in the pre-test.

Table 1. Results of the paired-samples t-tests ($n = 48$)

				95% CI					
Variety	*M*	*SD*	*SE*	Lower	Upper	*t*	*df*	*p*	*d*
Korean	19.39	12.09	1.75	22.90	15.88	11.11	47	< .001	1.55
Chinese	8.01	10.14	1.46	10.95	5.06	5.47	47	< .001	0.81

V. Discussion

The results revealed that the Japanese university students' comprehension of Korean and Chinese varieties of English improved significantly after participating in the guided awareness-raising activities, thus suggesting that the activities might have improved their ability to process speech in WE varieties accurately.

In this study, the students gained explicit knowledge about the typical segmental and suprasegmental features of the Korean and Chinese varieties of English, which may have enhanced their awareness of and perceptual ability to understand Chinese- and Korean-accented pronunciation in English. Such increased awareness and perceptual sensitivity may have enabled them to identify L1-accented sounds and to segment connected speech efficiently [8]. Thus, learning the typical segmental and suprasegmental features of Korean and Chinese accents in the pronunciation of English may have resulted in improving the comprehension of these English varieties. In addition, L2 pronunciations in unfamiliar accents are acknowledged to be difficult to understand [9]. The students' exposure to a particular accent via the awareness-raising activities enhanced their familiarity with it and led to a better understanding of the variety [10]. Thus, awareness-raising activities targeting the typical pronunciation characteristics of particular varieties of English might benefit the listening comprehension thereof.

In terms of the differences between the two varieties of English, the Japanese university students understood the Korean variety more easily than they did the Chinese equivalent, as shown by the large ($d = 1.55$) and medium-to-large ($d = 0.81$) effect sizes. This supports the findings of previous studies that showed that Japanese learners understood Korean speakers' English better than they did that of Chinese speakers [11]. In summary, our findings generally highlighted the potentially beneficial role of awareness-raising activities in improving EFL learners' listening comprehension of WE pronunciation.

References

[1] M. Berns, "World Englishes and communicative competence," in *The Handbook of World Englishes*, 2nd ed., C. L. Nelson, Z. G. Proshina, and D. R. Davis, Eds. West Sussex, UK: Wiley-Blackwell, 2020, pp. 674–685.

[2] A. D. Ören, A. Öztüfekçi, A. C. Kapçık, A. Kaplan, and Ç. Yılmaz Uzunkaya, "Building awareness of world Englishes among university preparatory students," *International Online Journal of Education and Teaching*, vol. 4, pp. 483–508, 2017.

[3] J. Field, "An insight into listeners' problems: Too much bottom-up or too much top-down?" *System*, vol. 32, pp. 363–377, 2004.

[4] L. Vandergrift, and C. C. Goh, "Teaching and learning second language listening: Metacognition in action," Oxon, UK: Routledge, 2012.

[5] A. Mizuta, "The unchanged images of English in changing Japan: From modernization to globalization," *J. of Intercultural Communication Studies*, vol. 18, pp. 38–53, 2009.

[6] S. I. Ishikawa, "Design of the ICNALE-Spoken: A new database of multi-modal contrastive interlanguage analysis," *Learner corpus studies in Asia and the world*, vol. 2, pp. 63–76, 2014.

[7] T. M. Derwing and M. J. Munro, "Accent, intelligibility, and comprehensibility: Evidence from four L1s,' *Studies in Second Language Acquisition*, vol. 19, pp. 1–16, 1997.

[8] E. M. Kissling, "Pronunciation instruction can improve L2 learners' bottom-up processing for listening," *The Modern Language J*, vol. 59, pp. 653–675, 2018.

[9] G. J. Ockey, and R. French, "From one to multiple accents on a test of L2 listening comprehension," *Applied Linguistics*, vol. 37, pp. 693–715, 2016.

[10] T. M. Derwing, M. J. Rossiter, and M. J. Munro, "Teaching native speakers to listen to foreign-accented speech," *J. of Multilingual and Multicultural Development,* vol. 23, pp. 245–259, 2002.

[11] M. Orikasa, "The intelligibility of varieties of English in Japan," *World Englishes*, vol. 35, pp. 355–371, 2016.

Explicit Rules or Implicit Imitation: A Comparative Study on the Approaches of Teaching English Prosody

Xiaodan. Zhang & Joaquin. Romero·
Rovira i Virgili University (Spain)

Keywords — sentence stress, implicit approach, explicit approach, prosody teaching

I. INTRODUCTION

Improper English prosody is one primary cause of foreign accent which can lead to unintelligibility of an EFL learner's speech [1]. Early research on pronunciation focused almost exclusively on segmental instruction. However, in more recent studies, a paradigm shift has been witnessed from a focus on segmentals to an increased emphasis on suprasegmental aspects and prosody [2]. Several studies have attempted to prove the effectiveness of introducing rhythm instruction within the English L2 class to improve EFL learners' prosodic skills [3] [4]. Nevertheless, there is still a scarcity of research exploring the efficacy of prosody instruction, especially to Chinese EFL learners whose L1 has a completely different rhythm from English. Therefore, the current research intends to investigate if Chinese EFL learners' pronunciation could be refined by a period of training as far as word stress and sentence stress are concerned and whether improvements vary from different training approaches, i.e., explicit rule explanation vs implicit imitation. The following two major research questions are formulated:

RQ1. Does prosody training help Chinese EFL learners improve their pronunciation?

RQ2. Which approach (explicit vs. implicit) is more effective in improving Chinese EFL learners' pronunciation?

II. METHOD

28 Chinese university students (20F, 8M) aged from 21 to 25 with an intermediate level of English participated in the research. The participants were divided equally into two experimental groups (explicit and implicit) and attended 8 online pronunciation training sessions of 30 minutes each. The whole training was done remotely online, as the researcher and instructor was residing in Spain while the participants were in China. The pronunciation module was designed to instruct placement of word stress and sentence stress. Following the same syllabus, however, the teaching approaches and materials varied between the two experimental groups. The explicit group was exposed to the pronunciation rules explicitly. Meanwhile, the implicit group was provided with native speaker recordings and were required to imitate after the audios. At the end of the implementation, 7 participants from the explicit group and 12 from the implicit group completed all 8 sessions. All participants did a pre-test before the treatment and then a post-test upon completion of the training. Both tests comprised a controlled reading task (words and sentences), followed by a spontaneous speaking task with pictures and questions as prompts. To accurately analyze the differences in the participants' pronunciation between the pre-test and the post-test, the content of the controlled reading remained the same, whereas the prompts for spontaneous speaking varied slightly so as to discard the influence of familiarity with the speaking task. After collecting the recordings, the controlled reading data were analyzed acoustically using the PRAAT software. In terms of the spontaneous speaking data, they were judged by 8 native raters on a 5-point Likert scale in terms of comprehensibility, accentedness and fluency. Due to time and space limitation, in this paper, results only present the acoustic analysis of a selection of the controlled reading data from the pre- and post-test as an example. As far as word stress is concerned, each syllable in the word was identified in the acoustic signal and a reading of the intensity peak was obtained. These values were then processed in order to obtain differences in amplitude between the stressed and the unstressed syllables.

III. RESULTS & DISCUSSION

The acoustic patterns of the word 'satisfactory' from a participant's pre-data and post-data are presented in Figure 1 and 2, respectively. The amplitude values are represented by the yellow line and an obvious variance is detected among syllables in the post-data (Figure 2) compared with the pattern in the pre-data (Figure 1).

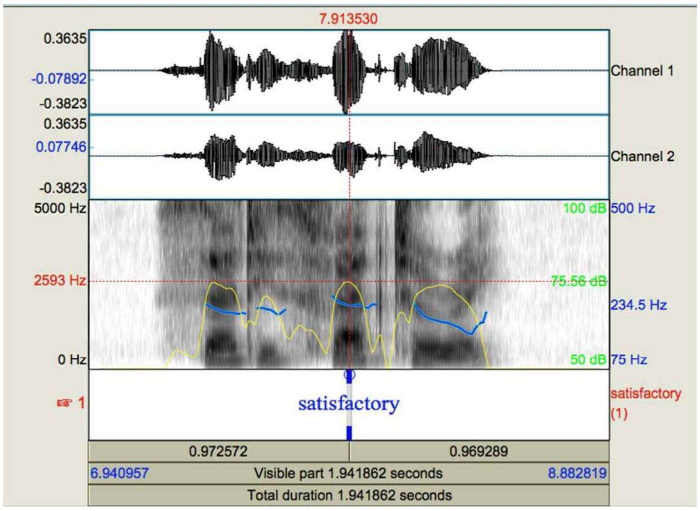


Fig. 1. Acoustic pattern of 'satisfactory' from pre-data

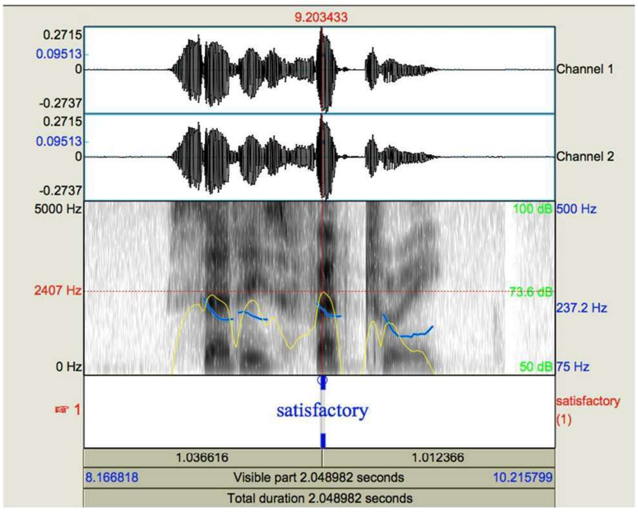


Fig. 2. Acoustic pattern of 'satisfactory' from post-data

Preliminary results show clear differences in amplitude between stressed and unstressed syllables in both groups' pre and post data. Yet, a more precise statistical analysis is needed to investigate whether there is a significant difference.

In terms of sentence stress, the accuracy of participants' placement of tonic stress/contrastive stress/emphatic stress was examined. The audio productions were analyzed in PRAAT to check whether the stress was placed on the correct words. The following Figures 3 and 4 present a participant's pre and post-production. It can be observed that the participant successfully managed to shift the stress to the correct place (*doctor*/*nurse*) in the post reading.

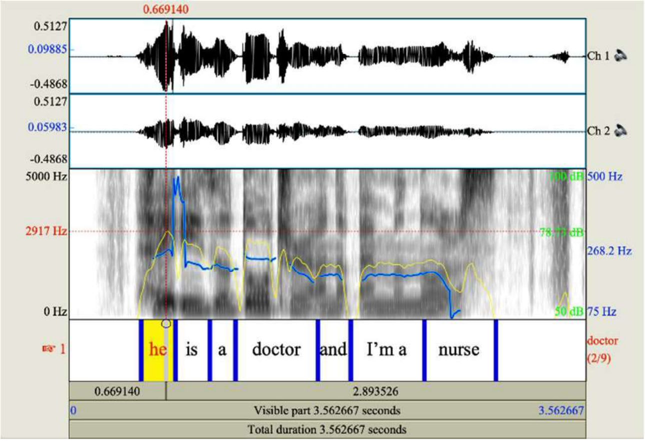


Fig. 3. Acoustic pattern of '*He is a doctor and I'm a nurse*' from pre-data

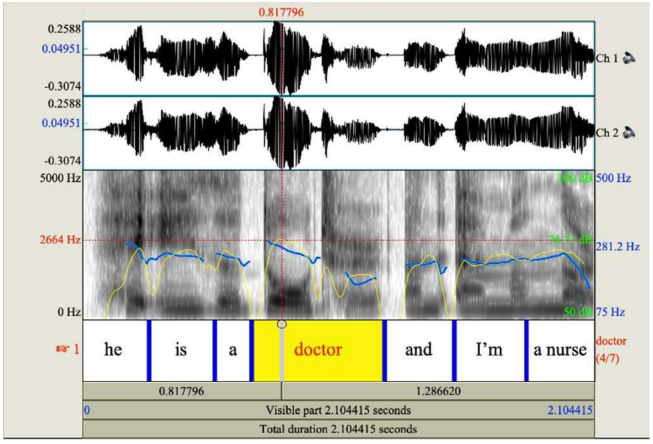


Fig. 4. Acoustic pattern of '*He is a doctor and I'm a nurse*' from post-data

Likewise, further analysis is still needed to investigate the statistical significance between the pre and post data from both groups, as well as to confirm which method of training is more effective. However, from the preliminary analysis, improvements have been detected from both groups after the treatment. This study seems to provide evidence for the contribution of suprasegmental instruction in the learning of pronunciation by EFL learners.

Reference

[1] Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL quarterly*, 38(2), 201-223.

[2] Busà, M. G. (2012). The role of prosody in pronunciation teaching: Agrowing appreciation. In M. G. Busà & A. Stella (Eds.),*Methodologicalperspectives on second language prosody. papers from ML2P 2012*(p. 101-106). Padua: Cleup

[3] Vázquez, L. Q., & Romero, J. (2018). The improvement of Spanish/Catalan EFL students' prosody by means of explicit rhythm instruction. In *ISAPh 2018 International Symposium on Applied Phonetics. Aizuwakamatsu, Japan.*

[4] Tsiartsioni, E. 2011. Can pronunciation be taught? Teaching English speech rhythm to Greek students. In Eliza Kitis, Nikolas Lavidas, Nina Tpointzi and Tasos Tsangalidis (eds.), *Selected papers from the 19th International Symposium on Theoretical and Applied Linguistics* (*ISTAL 19*), 447-458.

Transition or insertion? Acoustic analysis of the vocalic section of Mandarin *xia* in the production of Hungarian learners of Chinese

Andrea Deme[1] and Kornélia Juhász[1,2]
[1]Eötvös Loránd University, [2]Hungarian Research Centre for Linguistics

Keywords — alveolopalatal /ɕ/, coarticulatory transition, Mandarin Chinese, orthographic effect, L2 acquisition

I. INTRODUCTION & HYPOTHESES

In terms of L2 acquisition, many features of the Mandarin Chinese /ɕa/ [ɕja] (pinyin: *xia*) sequence are of interest from the perspective of Hungarian speaking learners of Chinese (HLCh). The alveolopalatal [ɕ] is a palatalized postalveolar sibilant, which means that apart from the constriction at the post-alveolar region, the dorsum also approximates the palate creating a long and narrow palatal channel. Due to this palatalization gesture, when the sibilant constriction is released into the following vowel, an audible [j] element is produced ([ɕja]) [1,2]. Both the alveolopalatal [ɕ] and palatalization as a secondary articulatory gesture are new to HLCh, on which basis we may expect second language accent effects in the production of the sequence at hand both in the fricative and the vocalic (sonorant) phases. With respect to the fricative phase, we investigated the place of articulation, and the degree of palatalization of the sibilant in HLCh [3] in 夏*xià* /ɕa/. As for the place of articulation, our results showed that HLCh produced the Chinese /ɕ/ identically to the Chinese /s/ (as in 萨*sà* /sâ/), similarly to Chinese natives. As for the degree of palatalization, however, we found differences as captured by F_2 frequency measured at the onset of the vocalic section: beginners showed patterns similar to what was observed in natives, but advanced learners produced /ɕ/ acoustically more palatalized (i.e., more back) than native speakers. Patterns in advanced learners suggests a difference in the coarticulatory effect exerted by the vocalic context, and that in advanced learners, the vocoid exerting this effect is more back than that we would find in natives (i.e., [i]-like and not [j]-like) [3,4].

In the present study we move our focus to the vocalic section following the alveolopalatal /ɕ/, to test the above assumption. In the Chinese /ɕa/ sequence, there is only two articulatory-acoustic targets (corresponding to /ɕ/ and /a/), and these are linked by a coarticulatory transition, the audible [j] element. By contrast, the pinyin transcription *(xia)* contains three graphemes, which seems to index three different articulatory-acoustical targets in the syllable. It is well known that the orthographical denotation of the L2 segments influences L2 perception and production [5]. On this basis, we hypothesize that HLCh's production is biased by the pinyin transcription, and that they produce an additional articulatory-acoustic target, i.e., an [i], inserted between the sibilant and the vowel. This means that Hungarians are expected to produce two acoustic targets instead of one, in the realization of the /ɕa/ sequence's vocalic section: we assume that duration of vocalic sections are longer in HLCh than in natives, and that HLCh produce longer vocalic phases in /ɕa/ that in other CV-structured Chinese syllables, e.g., /sa/ or /ʂa/. Furthermore, we expect to find differences in the F_2 curves between HLCh and natives, as well. Results of [6] showed that in /ɕa/, the F_2 formant frequency (which is positively correlated with the length of the cavity behind the constriction) decreases gradually and shows a concave curve with no positive excursions or stable phases at its onset, which mirrors a simple coarticulatory transition from the palatalized [ɕ] consonant to the following central-low [a] vowel and no further vocalic targets. Based on this, and the fact that HLCh are expected to insert a palatal [i] in the sequence at hand, we anticipate that we find a positive F_2-excursion at the beginning of the vocalic section in HLCh, and not the native-like gliding transition. In accord with the above, we assume also that HLCh produce the Chinese /ɕa/ identically to the Hungarian /sia:/ sequence.

II. METHOD

We compared the production of three groups: two groups of HLCh (beginners and advanced learners), and Chinese native speakers. We analysed three isolated words: 夏*xià* /ɕa/ [ɕjâ] 'summer'; 萨*sà* /sâ/ [sâ] 'a surname'; and 厦 *shà* /ʂâ/ [ʂâ] 'castle' that were read eight times in a randomized order by all speakers. We also recorded Hungarian pseudowords, /sa:/, /sja:/, /sia:/, and /sija:/, in isolation, produced by HLCh. We segmented, labelled, and analysed all sound samples in Praat [7]. We extracted F_2 automatically at every 5th millisecond throughout the whole quasiperiodic signal phase. Duration of the vocalic section of all syllables was also measured. In the statistical analysis, we compared durations measured in Chinese syllables produced by all speakers using a linear mixed model (with the fixed factors of speaker group and word) in R [8]. F_2 curves were submitted to generalized additive modelling (GAMM), where we analysed the effect of the normalized timepoint predictor on the dependent variable of F_2 and added the parametric term of word and random smooth by each trajectory. One model was built for comparing the production of Chinese /ɕa/ across groups, and another for comparing Chinese and Hungarian words in the production of HLCh.

III. Results & Discussion

Regarding duration, the mixed model showed significant interaction effect of speaker group and word (*F* (4, 485) = 9,01, *p* < .001), and while natives and advanced learners produced the Chinese /sa/, /ʂa/ and /ɕa/ with the same duration, beginners produced the vocalic section of /ɕa/ significantly longer than that of the other two syllables (*p* < .001) (Fig 1, left) which suggest the insertion of an extra articulatory-acoustic target in beginners. With respect to F_2 curvature, F2 curves of words in both GAM models showed non-linearity (*p* < .001). We found the expected concave coarticulatory offglide at the vocalic onset in natives (Fig 1, right). However, F_2 curves found in both groups of HLCh differed from this pattern with the production of beginners approximating native patterns better. While beginners featured an F_2 curve with a constant, or slightly decreasing interval at the onset, advanced learners featured a domed curve, that reached relatively higher frequencies (Fig 1, right). Comparing Hungarian and Chinese syllables in HLCh, we find steeper slopes and curves skewed more to the right in Hungarian words than in Chinese, that are the result of Hungarian long /aː/ being produced longer than Chinese /a/ Fig 2). In the case of advanced learners, the onset of the vocalic phase in /ɕa/ and in Hungarian /sjaː/ completely overlapped (Fig 2, right), while beginners produced a more transition-like onset in /ɕa/ with significantly lower F_2 than that in /sjaː/ (Fig 2, left). In sum, duration and F_2 curves showed opposing results: while in duration, production of advanced learners approximated native production more than that of beginners, in F_2 curves, beginners showed more native-like patterns, i.e., coarticulatory offglide transitions from the sibilant to the vocalic target. Comparison of Chinese and Hungarian words revealed that both groups produced the onset of the vocalic section in /ɕa/ similar to that of the Hungarian /sjaː/ sequence with advanced learners showing more similarities. This suggests that advanced learners realised the Chinese sequence with a more specified [j] target than beginners (while beginners produced a more native like transition between the sibilant and the final /a/), and that longer duration in beginners may be attributed to different causes. In our interpretation, longer duration in /ɕa/ than in /sâ/ and /ʂâ/ in beginners may be attributed to L1 transfer, that is, they produced /a/ in /ɕa/ similar to the Hungarian long /aː/. To conclude, we did not find the expected effect of orthography, but the assumption regarding an additional target insertion as a foreign language accent effect was partially confirmed. Our results contribute to the better understanding of L2 transfer and L2 speech sound acquisition.

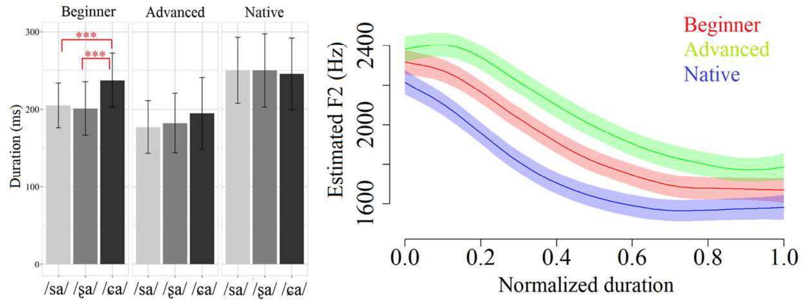


Fig. 1. Duration of the vocalic section of Chinese /sa/, /ʂa/, and /ɕia/ syllables (left), and the estimated F_2 curves of the /ɕa/ syllables produced by the three speaker groups within normalized duration (right)

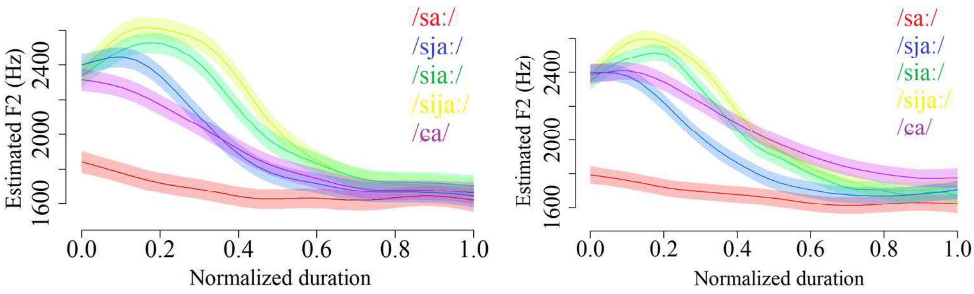


Fig. 2. Estimated F_2 curves of vocalic sections in the Chinese /ɕa/ and the Hungarian /saː/, /sjaː/, /siaː/, and /sijaː/ within the normalized duration, in the production of beginners (left) and advanced learners (right)

References

[1] D. Recasens, "On the articulatory classification of (alveolo)palatal consonants," *Journal of the International Phonetic Association*, vol. 43/1, pp. 1–22, 2013.

[2] P. Ladefoged, and I. Maddieson, *The sounds of the world's languages*. Oxford: Blackwell. 1996.

[3] K. Juhász, "The acoustic analysis of Mandarin Sibilants in the Production of Hungarian learners of Chinese." *Proceedings of International Symposium on Applied Phonetics 2021.* in press.

[4] K. Juhász, and A. Deme, „Mandarin kínai és magyar szibilánsok palatalizációja kínaiultanuló magyar anyanyelvűek ejtésében. [The palatalization of Mandarin and Hungarian sibilants in the production of Hungarian learners of Chinese." *Alkalmazott Nyelvtudomány*, vol. 2022/1. in press.

[5] B. Bassetti, "Orthographic input and second language phonology," In *Input Matters in SLA*, T. Piske, and M. Young-Scholten, Eds. Clevedon, UK: Multilingual Matters 2008. pp. 191–206.

[6] I. Hauser, "Coarticulation with alveopalatal sibilants in Mandarin and Polish: Phonetics or phonology?" *Proceedings of the Annual Meetings on Phonology*, vol. 8, pp. 1–8. 2020.

[7] P. Boersma, and D. Weenink, *Praat: doing phonetics by computer*. https://www.praat.org/. 2019. Version 6.1.15.

[8] R Core Team, *R Foundation for Statistical Computing*, https://www.R-project.org/. 2019. Version 3.6.1.

Short-term intervention effects on the development of pausing in read speech

Atsushi Fujimori (University of Shizuoka), Noriko Yamane (Hiroshima University),
Ikuyo Kaneko (Juntendo University), Brian Teaman (Osaka Jogakuin University)

Keywords —pausing, read speech, intervention effects, kinesthetic instruction

I. Introduction

This study investigated whether a short-term pedagogic intervention has an effect on the development of utterance fluency, particularly pausing, among Japanese EFL learners. Fluency, including not only speed but also pausing, plays an important role in L2 speech. Pause length, pause frequency and pause location are often utilized as measures of utterance fluency [1], as inappropriate prosody also affects comprehensibility.

A task of narrating picture stories was conducted with L2 learners of English, observing their longer periods of silence and more frequent pauses particularly in the middle of clauses than native speakers of English [2]. A similar finding was reported on read speech; Japanese learners of English in an EFL setting were asked to read aloud a monologue and the low level group (CEFR A2 level) produced inappropriately placed clause-internal pauses more frequently than the high level group (CEFR B2 level) and native speakers of English [3]. The current research question is whether low level learners of English can improve their pausing with a pedagogic intervention. To answer this question, we compared their oral performances in a pretest/posttest design.

II. Method

Participants were 18 Japanese undergraduates whose English proficiency was at CEFR A2 level. The L2 learners were divided into two groups: the Pause group and the Stress group. Both groups performed a kinesthetic practice while watching a kinesthetic video. We adapted a kinesthetic model for instruction which involves a specific physical movement in pronunciation training; Multimodal instructions are intuitively attractive and have been shown to be as effective as other visual and auditory instructions [4]. The Pause group watched a video where an instructor performed a downward hand chop with an outstretched arm at silent pauses including clause-internal and sentence final pauses. The Stress group performed the same downward chop aligned with stressed words, but not silent pauses, using the same text and sound materials. The training sessions lasted for four weeks, with a 20-minute training session each week. Before and after the training session, the participants were instructed to read it aloud and record it in Praat [5], after five minutes practicing at their own pace.

While 250 ms is often used as the minimum pause length [6], any silence longer than 150 ms was identified as a silent pause in the process of measurement, as the pause length produced by native speakers of English is quite short for read speech. The number of sentence-internal pauses was counted for each speaker. Also the mean length of sentence-internal pause (SIP) and the mean length of sentence-final pause (SFP) were calculated for each speaker [7].

III. Results

The mean values of relevant parameters in both pretest and posttest are summarized in Table 1. We ran ANOVAs for the two groups by the two tests and they showed that there was no significant difference between the two learner groups and within each group in number of SIPs and number of SFPs (number of SIPs: Group $F(1,32)$=.39, p=.845, Test $F(1,32)$=0.007, p=.933, Group x Test $F(1,32)$=0.096, p=.759). Meanwhile, the Pause group produced significantly longer SIPs and SFPs at the posttest than at the pretest (SIP: Group $F(1,32)$=3.383, p=.075, Test $F(1,32)$=2.908, p=.098, Group x Test $F(1,32)$=4.654, p=.039 (*), SFP: Group $F(1,32)$=1.425, p=.241, Test $F(1,32)$=1.378, p=.249, Group x Test $F(1,32)$=2.536, p=.121). Also their pause length was significantly longer than that of the Stress group at the posttest. Moreover, the total spoken time of the Pause group was significantly longer than that of the Stress group (Group $F(1,32)$=.787, p=.382, Test $F(1,32)$=.017, p=.897, Group x test $F(1,32)$=4.980, p=.033 (*)). The phonation time, the total spoken time minus total pause length, of the Stress group was significantly shortened at the posttest, while that of the Pause group remained the same after the intervention (Group $F(1,32)$=1.684, p=.204, Test $F(1,32)$=4.345, p=.045 (*), Group x Test $F(1,32)$=4.951, p=.033 (*)). These data indicate that after the intervention where the instructor demonstrated long SIPs and SFPs with hand gestures, the Pause group produced clear-cut boundary pauses while their phonation time hardly changed. In contrast, all the parameters including pause length as well as phonation time were lowered in the Stress group, indicating that stress-based intervention improved their speed in read speech.

TABLE 1. SUMMARY OF PAUSING DATA

Group	Test	SIP (No.)	SFP (No.)	SIP (sec.)	SFP (sec.)	Spoken time (sec.)	Phonation time (sec.)
Pause	Pre	12.556	9	0.361	0.785	47.265	34.910
	Post	13.333	9	0.547	1.021	51.876	35.062
Stress	Pre	13.556	9	0.377	0.819	49.889	38.699
	Post	13.111	9	0.355	0.783	45.790	34.064
Instruction video (Pause)	n/a	12	9	0.858	1.692	58.876	32.493

The number of SIPs in both groups did not change much in the posttest. As Table 2 shows, however, where the participants placed pauses clause-internally varies between the groups. The Pause group improved the accuracy of pause placement after the intervention; The total numbers of SIPs for the pretest and posttest were 72 and 75, respectively, whereas the number of inappropriate pauses decreased from nine to three. This tendency was apparent with pausing immediately before a complementizer *that*. In addition, the number of pauses decreased after sentence-initial adverbial phrases while it increased before a conjunction *and*, after the pause-based intervention. In contrast, the Stress group kept producing inaccurate clause-internal pauses frequently, particularly, after the complementizer *that*. These findings suggest that not only input but also associated kinesthetic instruction are effective for improving pausing in read speech. The question remains as to how native speakers of English would evaluate the performances of the two groups in comprehensibility. The results of their rating will be presented at the conference and not shown here due to space limitations.

TABLE 2. COUNT OF APPROPRIATE (AND INAPPROPRIATE) PAUSES

Group	Test	_ *that* (*that*_)	_*and* (*and*_)	DP_ (D_)	VP_ (V_)	PP_ (P_)	AdvP_ (Adv_)	Total number of appropriate pauses (inappropriate pauses)
Pause	Pre	4 (10)	21 (2)	10 (4)	14 (5)	4 (0)	7 (1)	60 (22)
	Post	9 (11)	12 (3)	7 (2)	12 (2)	7 (3)	8 (0)	55 (21)
Stress	Pre	12 (7)	21 (0)	3 (1)	9 (1)	5 (0)	13 (0)	63 (9)
	Post	17 (2)	29 (0)	2 (1)	12 (0)	7 (0)	5 (0)	72 (3)
Instruction video (Pause)	n/a	2	3	0	2	1	0	8

REFERENCES

[1] Mora, J. & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly* 46: 610—641.

[2] Tavakoli, P. (2010). Pausing patterns: differences between L2 learners and native speakers. *ELT Journal* 65: 71-79.

[3] Yamane, N., Fujimori, A., Teaman, B., & Yoshimura, N. (to appear). Fluency of read speech of L2 English. *Ars Linguistica* 28.

[4] Yamane, N., Teaman, B., Fujimori, A., Wilson, I., & Yoshimura, N. (2018). The kinesthetic effect on EFL learners' intonation. *Proceedings of ISAPh 2018 International Symposium on Applied Phonetics*.

[5] Boersma, Paul & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.09, retrieved 15 February 2022 from http://www.praat.org/

[6] De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*(2), 223-243.

[7] Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics, 39*(3), 569-591.

A comprehensive web application for research on vocalic differences in World Englishes (Native and Accented)

Simon Gonzalez, The Australian National University, U1037706@anu.edu.au

Keywords — socio-phonetics, vowel visualization, speech database, accented English, shiny app

I. Introduction

Current technological advances have allowed speech researchers to access online data in an unprecedented way. This has been a positive contribution from digital technologies. One area that has been greatly influenced is the second language field, where we can find numerous databases on accented speech, available from speakers around the world. However, there are two main challenges with this. The first one is that it can be computationally difficult to gather many audio files from given sources. The second one is that doing this requires strong computational skills to extract phonetic data and process it efficiently. In turn, this stage makes the data prepping more prompt for drastic diverging ways of wrangling the data, which raises questions on whether results across different studies can be comparable.

Based on these challenges, the purpose of this paper is to present the development of a speech technology, an open-source linguistic tool that addresses these issues within a single socio-phonetic app. The app gives users access to both language data and visualisation tools in a single place for accurate socio-phonetic analysis. It also offers cutting-edge analysis methodologies following established socio-phonetic approaches to vowel studies. Results from the app are therefore comparable to other socio-phonetic studies, even ready for academic publications. On this last point, an advantage of being open source is that methodologies for acoustic normalisation and vowel visualisation can be expanded in future versions of the app. Finally, it gives users access to all the raw files as well as all forced-aligned *TextGrids* created to analyse all the data outside the app.

II. Source Data

The app data was collected from the Speech Accent Archive [1]. This database gathers speakers from around the world recording themselves reading the following prompt:

> *Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.*

This paragraph contains all of the sounds of English. The purpose of the Archive is then to have a sample from a variety of language backgrounds, for both native and non-native speakers of English. A great advantage of this resource is that it allows users to make comparisons based on demographic and linguistic features of speakers. For the purposes of the current app, we only kept five characteristics: gender, country, region, city, and native language of each speaker.

III. Our Data Processing

All available samples were downloaded from the *Archive* using an R [2] algorithm developed by the main author. For each audio file, we created a corresponding Praat [3] *TextGrid.* Each pair of files per speaker (audio file and TextGrid) was then used as the input for force-alignment using the *Montreal Forced Aligner* [4]. The outputs were TextGrids forced-aligned at word and phonetic levels. These outputs were then prepared for acoustic analysis, and especially focusing on vowel analysis for socio-phonetic research. In this process, we extracted 11 equidistant points per vowel trajectory, which allows us to measure vowels both in static and dynamic ways. Following standard phonetic practices, we normalised vowels using five of the well-known normalisation methods: bark [5], Labov [6], Lobanov [7], Neary [8], and Watt and Fabricius [9]. From all the files, we chose a representation of one male and one female speaker per city. The final data has 880 individual speakers, from 127 languages and 135 countries. For a final count of all vowels in the data, as well as the mean durations and overall percentages, see Table 1 below.

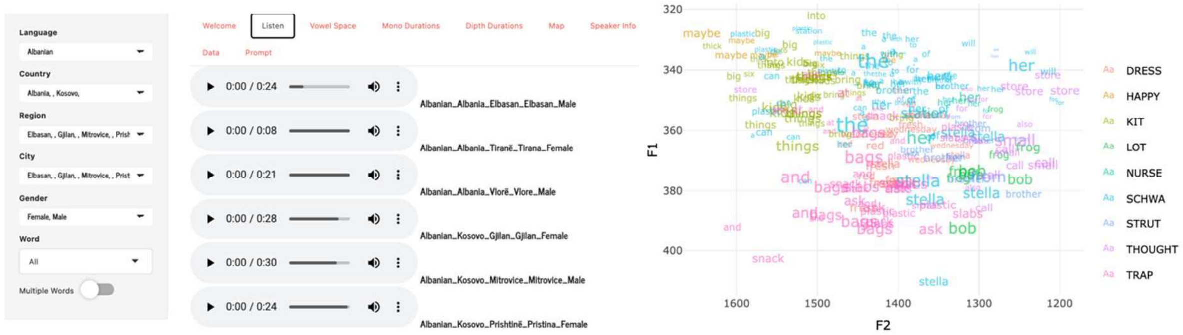
IV. App Development and Results

The App was developed using *Shiny* apps [10], an RStudio [2] web-based framework. The main motivation for using R is that it is widely used for speech analysis and, combined with Shiny apps, it gives users great control over applications that are both efficient and interactive. Additionally, Shiny applications are intrinsically reactive, which is invaluable when interacting with online apps. Another relevant advantage of R is that it allows us to create efficient speech visualisations that capture the many complexities of language data. For this, we decided to plot F2 values on the x-axis (reversed) and F1 values (reversed) on the y axis. For monophthongs, we plot one value (50%) and for diphthongs, two points, either 10%-90% trajectories or 20%-80% trajectories. All these options allow users to make meaningful and phonetically implemented vowel comparison methodologies. Another important aspect in terms of visualisations, is that it allows controlling all aspects of the graphics (colour, sizes, themes, texts) to be customised by users. In this way, it gives freedom to download vowel space plots ready for sharing, comparison, and publication.

The other relevant aspect of the app is accessing and selecting speakers to listen to in audio files. A major component of this app is to listen to all available data and observe vowel spaces of selected speakers. We have created a friendly user section to make selections based on native language, country, region, city, and gender. For each selection, the app creates an audio play widget for each speaker, as well as location pins in a world map, and the vocalic space for all selected speakers. Additionally, colours/groups in vowels can be changed to reflect different patterns, (vowels by language, country, region, city, or gender). All these features aim to give users control over data visualisation and selection, which, to our knowledge, has not been implemented previously in a free resource looking at accented speech on world Englishes.

TABLE I. ALL VOWELS IN THE DATA (MEAN DURATIONS, TOTAL COUNTS AND PERCENTAGES)

Vowel	Mean Duration (ms)	N	Percentage
SCHWA	93	12318	19%
FLEECE	114	8484	14%
KIT	84	7079	12%
TRAP	128	6648	11%
FACE	119	4247	7%
THOUGHT	92	4007	7%
DRESS	89	3415	6%
GOOSE	123	2943	5%
STRUT	75	2984	5%
GOAT	105	2348	4%
LOT	149	1551	3%
NURSE	122	1672	3%
HAPPY	104	1217	2%
CHOICE	205	605	1%


Language: Albanian
Country: Albania, , Kosovo,
Region: Elbasan, , Gjilan, , Mitrovice, , Prisł
City: Elbasan, , Gjilan, , Mitrovice, , Prist
Gender: Female, Male
Word: All
Multiple Words

0:00 / 0:24 Albanian_Albania_Elbasan_Elbasan_Male
0:00 / 0:08 Albanian_Albania_Tiranë_Tirana_Female
0:00 / 0:21 Albanian_Albania_Vlorë_Vlore_Male
0:00 / 0:28 Albanian_Kosovo_Gjilan_Gjilan_Female
0:00 / 0:30 Albanian_Kosovo_Mitrovice_Mitrovice_Male
0:00 / 0:24 Albanian_Kosovo_Prishtinë_Pristina_Female

Fig. 1. App controls and vowel visualisation spaces from the selected data

REFERENCES

[1] Weinberger, S. (2015). *Speech Accent Archive. George Mason University*. Retrieved from http://accent.gmu.edu

[2] RStudio Team (2021). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[3] Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

[4] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*.

[5] Hartmut, T. (1997). *Auditory scales of frequency representation*. [Online: http://www.ling.su.se/staff/hartmut/bark.htm]

[6] Labov, W., Ash, S. & Boberg, C. (2006). *The Atlas of North American English: Phonology, Phonetics, and Sound Change*. A Multimedia Reference Tool. Berlin: Mouton de Gruyter.

[7] Lobanov, B. M. (1971). Classification of Russian vowels spoken by different listeners. *Journal of the Acoustical Society of America* 49:606-08.)

[8] Nearey, T. M. (1977). *Phonetic Feature Systems for Vowels*. Dissertation, University of Alberta. Reprinted 1978 by the Indiana University Linguistics Club.

[9] Watt, D. & Fabricius, A. (2002). Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1 ~ F2 plane. In D. Nelson, Leeds, *Working Papers in Linguistics and Phonetics* 9:159-73.

[10] Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2019). *shiny: Web Application Framework for R* (Version 1.3.2) [R package]. Retrieved from https://CRAN.R-project.org/package=shiny

4th International Symposium on Applied Phonetics (ISAPh2022), September 14-16 2022, Lund University, Sweden

The interplay of tone and intonation: f0 contours produced by Hungarian learners of Mandarin

Kornélia Juhász and Huba Bartos
Hungarian Research Centre for Linguistics, Eötvös Loránd University

Keywords — Mandarin Chinese, f0 contour, intonation, L2 production

I. Introduction

In this acoustic analysis, we aim to shed light on how Hungarian learners of Mandarin Chinese contrast the intonation patterns (conceived as f0-contours) of unmarked yes-no questions and statements, by analysing 4-syllable utterances having the same tonal value, either rising (T2) or falling (T4), throughout. In tonal languages, such as Mandarin Chinese (MC), f0 serves for both the realization of lexical tones and intonation [1]. Firstly this means that f0 modulation is locally dependent primarily on tone values (or tonal contexts), further affected by segmental effects, as well. Moreover these local effects also interact with intonation patterns, yielding the actual f0 contour [2]. Finally, adjacent tones are also closely dependent on each other (tonal coarticulation).

Taking a broader view, in distinguishing MC statement and question intonation patterns, localized (e.g. terminal rise on the last (tonal) syllable) and global acoustic cues (raised f0 over the whole utterance) have been identified concerning f0 register and f0 range [3]. Regarding f0 register, according to Shen's MC intonation model [1], statements display a gradually descending pattern, while unmarked yes-no questions feature a significantly higher f0 throughout the whole utterance (compared to the declarative contour), complemented by a terminal rise. According to Shen's study, this elevated characteristic of the question intonation pattern affects not just the top line of the contour (the peaks of each syllable strung together), but also the base line (the valleys of each syllable), and there is no expansion of the f0 range. The aforementioned terminal rise can be attributed to the lack of lexical/syntactic cues (such as particle *ma*), thus prosodic cues are used exclusively to express the interrogation [3]. However the realization of the terminal rise is tone-dependent, meaning that in case of T2, the tonal f0 curve is realized with a widened pitch range, which rises much higher than in statements. In the case of T4 the f0 range remains intact both in question and statement intonation, but the f0-contour is realized at a higher level for questions [4].

In contrast to MC, Hungarian is a non-tonal language, thus f0-change is manifest only at clause level. In addition to this, while in Hungarian the declarative pattern is realized in a descending manner similar to MC, the prosodic structure of yes/no questions' f0 contour differs in MC and Hungarian: in the latter, the initial f0 value is low (compared to declaratives), and the f0 contour is characterized by a rising structure peaking on the penultimate syllable (unless fewer than two syllables follow the final phrasal stress), followed by a fall [5]. In L2 acquisition, prosodic patterns, e.g. intonation patterns, are transferred from L1 [6]. Due to L1 prosodic transfer, on the basis of the different structures of intonation patterns and the absence of tones in L1, we expect that synchronizing tone and intonation production poses problems to Hungarian learners. In particular, we hypothesize that Hungarian learners of MC favor the proper tone-production over intonation, in this manner produce unmarked questions similar to statements, without an elevated f0. In addition, we are seeking an answer for the question how language learners' f0 contours are shaped in T2 and T4 sequences.

II. Method

We analysed three adult speaker groups (5 women per group, altogether 15 speakers): 1. Hungarians with cca. one year language experience of MC (beginners); 2. Hungarians with 3-4 years of learning MC (advanced learners), and 3. a control group of Chinese natives. Regarding the material, we compared the production of declarative and syntactically unmarked yes/no interrogative sentences, all SVO with 4 syllables (2-1-1 syll.), matching pairwise in syllable structure and the number of voiced segments. Each syllable in a sentence has the same tonal value (either T2 or T4). Three interrogative and three declarative sentences for each tone were read for three times, presented as short question-answer dialogues, projected on a screen with Chinese characters and *pinyin*. The sentences were introduced to the speakers before recordings were made. Also, the instructions emphasized that the aim of the experiment was contrasting question and statement intonation. We recorded altogether 3600 syllables (2 tones × 3 sentences × 5 repetitions × 2 modalities × 15 speakers). To give an example for a question-answer pair for T4 sequences: A: 蒗雀卖电。*Làng Què mài diàn.* "Lang Que sells electricity."; Q: 蒗雀卖电? *Làng Què mài diàn?* "Does Lang Que sell electricity?" Within voiced segments, f0 was extracted by 5 millisecond intervals automatically in Praat [7]; f0 was converted to semitones (with a reference value of 50 Hz) and f0 curves were analysed by generalized additive mixed models (GAMM; [8]) in R [9]. Matching syllables of questions and answers were analysed together. For each syllable pair, a basic model was built where the f0 (in semitones) is the dependent variable and the normalized duration was the independent variable, and the models were further complemented by a factor variable of the modality produced by different speaker groups (equals to 6 levels).

III. Results & Discussion

The GAMMs showed significant differences between speaker groups as well as between the question and answer modalities in all 4 syllables of both tone sequences. In native Chinese speakers' production, there is a clear pattern in T4 sequences: the f0 curve of questions is realized in a higher frequency range compared to statements, and the difference between the two curves expands gradually towards the end of the utterance (Fig. 1, right). The same pattern was observed in case of T2 sequences as well, but the differences were notable in the last two, though less apparent in the first two, syllables of the utterances (Fig. 1. left). We should further notice that our results confirm [1]'s results: there is a significant range expansion on the last T2 syllable, while the T4 pattern remains roughly intact but elevated to a higher f0 range. Comparing beginners' and advanced learners' production, beginners are inconsistent in distinguishing question and statement patterns, while advanced learners tend to stick to the same f0 patterns in both. Regarding the first syllable of the sequences, in T2-production both learner groups approximated natives' f0 curves more or less. However, in T4-production, in defiance of the falling characteristic of this tone, both Hungarian groups produced a strikingly different, rising, curve, possibly due to preparing for the appropriate height for the fall of the subsequent syllable. As an example for tonal patterns, in case of the last two syllables of T4 sequences (especially the 3rd syllable, where the phase of the voiced initial is also included) a delayed target approximation is observed reaching the peak of the domed curve, while Hungarian production featured a moderately concave pattern. In sum, as expected, L2 learners did not approximate the elevated characteristic of natives' question-production. Besides, regardless of interrogative or declarative mood, L2 learners' tonal production yielded different patterns compared to natives, which is possibly due to fundamental problems in lexical tone production itself, regardless of intonation or tonal context. The results contribute to the understanding of the intonation acquisition of a tonal L2 language.

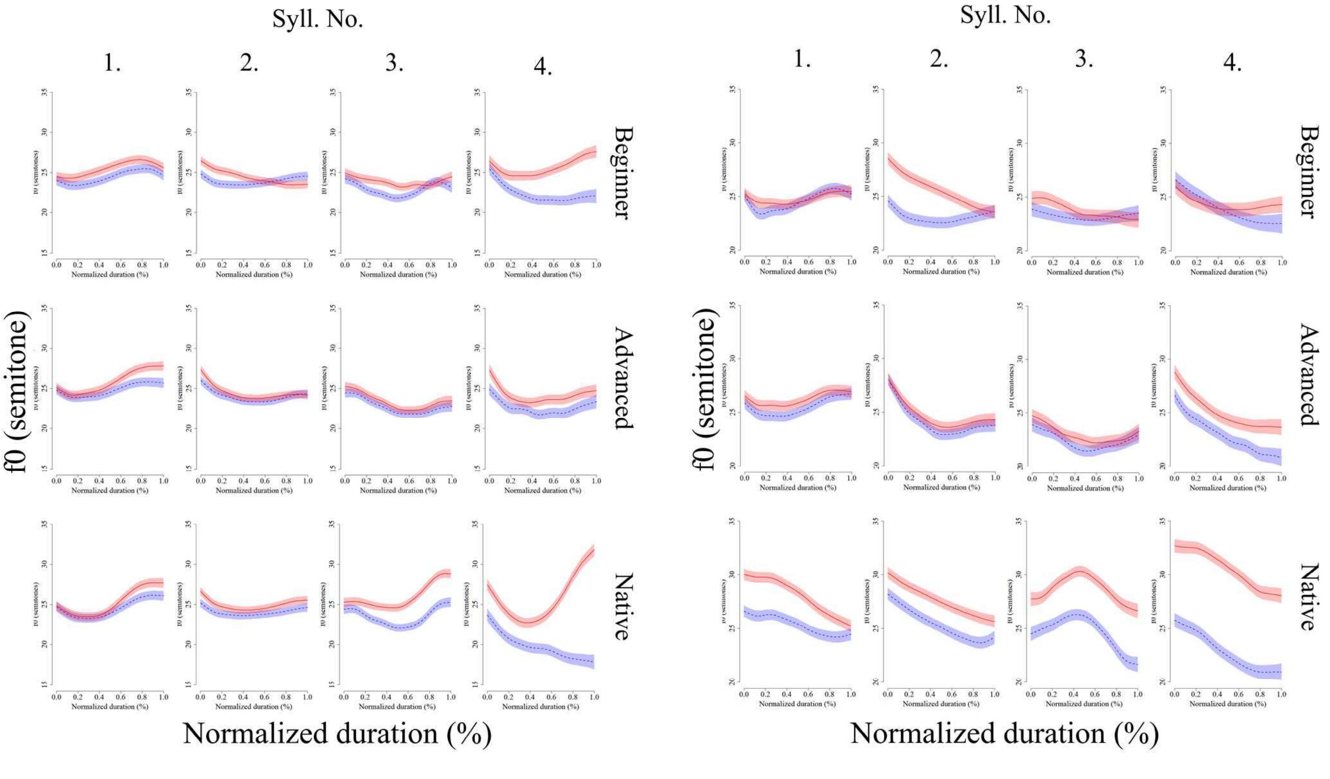


Fig. 1. T2 (left) and T4 (right) sequences' estimated f0 curves in questions (red solid line) and statements (blue dashed line) for each syllable in the production of beginners, advanced learners and natives

References

[1] X. N. Shen, *The Prosody of Mandarin Chinese*. California: University of California Press, 1990.
[2] C. L. Shih, "A Declination Model of Mandarin Chinese," *Intonation: Analysis, Modelling and Technology*, vol. 1, pp. 1–23, 2000.
[3] O. J. Lee, *The prosody of questions in Beijing Mandarin.* Doctoral Dissertation, Ohio State University, 2005.
[4] X. N. Shen, "Interplay of the four citation tones and intonation in Mandarin Chinese", *Journal of Chinese Linguistics,* vol. 17, pp. 61–74, 1989.
[5] G. Németh, and G. Olaszy, *A magyar beszéd.* Budapest: Akadémiai Kiadó, 2010.
[6] S. A. Jun, and M. Oh, "Acquisition of Second Language Intonation," *Acoustical Society of America Journal*. vol. 107, pp. 73–76, 2000.
[7] P. Boersma, and D.Weenink, *Praat: doing phonetics by computer*. v. 6.1.05, Available: https://www.praat.org/, 2019.
[8] S. N. Wood, *Generalized Additive Models: An Introduction with R.* New York: Chapman and Hall, 2017.
[9] R Core Team, *R: A language and environment for statistical computing.*, v. 3.6.1. Available: https://www.r-project.org/ 2019.

Acoustical cues as boundary markers in a left-headed language

Valéria Krepsz and Anna Huszár

Keywords — name grouping, boundary cues, prosody, acoustics, syntax-prosody interface

I. INTRODUCTION

Previous analyses based on some languages examined how differently coordinated names or numbers are realized prosodically (for example Ladd [1] and Wagner for English [2], Féry and Kentner for German and Hindi [3], [4], and Féry and Truckenbrodt for German [5] as well). Eg. Wagner [2] compared the acoustic implementation of embedded structures with different strengths: Among the first researchers, he introduced a methodology (that was later adopted by many other linguists), where the speech units either appeared as equals (Name1 and Name2 and Name3 and Name4) or implemented embedded categories of different depths, such as ((Name1 and Name2) or Name3 and Name4) or (((Name1 and Name2) and Name3) or Name4). Reading aloud these structures forces speakers to label the boundaries after the detached speech units by using different prosodic boundary cues.

According to the latest model of speech boundaries, the connection of the speech units or conversely the signaling of the boundary between two elements raised from the principles of 'Proximity' and 'Similarity'. Proximity is based on the adjacent grouped elements' syntactic constitution and reflects the syntactic boundaries regarding the pitch and duration in the prosodic structure of a given linguistic unit. In contrast, Anti-proximity means that elements in a separate group are realized with prosodic spacing, implementing boundary markers between the groups. In the case of the Similarity principle, syntactic embedding is determined by the similarity of the elements: the components of the elements belonging to a group are similar in the given prosodic structure, while units in other groups are different from them (Anti-Similarity).

The interaction between syntactic structure and the planning of prosodic boundary markers has been investigated mostly for English and German, but not for Hungarian. However, many prosodic features of Hungarian differ from the two languages mentioned. Hungarian prosody is typologically different from the prosody of Germanic languages. First, information structure in Hungarian is primarily expressed by word order, i.e. logical functions are linked to certain sentence positions. The position of focus is defined syntactically (it is immediately pre-verbal), while prosodic prominence marking plays only a secondary role and is partly optional [6]. Second, Hungarian is a left-headed head/edge-prominence language, while German is a right-headed head-prominence language. This means that in Hungarian, phrase-initial pitch accents mark the left edge of an accentual (minor) phrase simply by their position. Besides, unlike in English and German, main prominence is not expected to fall on content words towards the end (i.e. the right head) of an intonational phrase, unless the sentence lacks post-verbal units. Thus, it is expected that speech production in terms of prosodic planning is different in these typologically distinct languages.

Therefore, the aim of the present study is to examine how boundaries of different strengths based on the variously coordinated names are realized in Hungarian. The main question of the examination is how the prosodic boundary marker patterns differ from each other regarding their position and syntactic structure (left or right-headed structures).

II. METHODOLOGY

The structure of the research was based on Wagner's [2] experiment. The material included different name grouping structures consisting of 3 or 4 proper ordinary Hungarian first names. All of them were disyllabic and consisted of short vowels and sonorant consonants (eg. 'Vili', 'Mira', 'Lili', 'Vali' etc.). The names were conjuncted with 'és' (and) and 'vagy' (or) linking words. For example: (Vili és Mira) vagy (Lili és Vali). The trigger structures were embedded in a question-answer context. And the recordings were made using the SpeechRecorder software.

For the examination of the acoustic parameters of the words, the A / B / C [/ D] units were considered as a reference for both the 3-element and 4-element units, and the other word durations were compared to this. In the case of the 3-element-units, 2 other constructions were analyzed: (A and B) / C and A / (B and C), while in the case of 4-element-forms, 5 other structures were analyzed: A / B / (C and D); (A and B) / C / D; ((A and B) / C) and D; A and (B / (C and D)); (A and B) / (C and D). The order of the sentences was randomized.

25 monolingual speakers of Hungarian participated in the study. They were instructed to read aloud the structure to indicate which elements belong together. They were given as much time as they needed, and had the opportunity to correct themselves. If corrections were made, the last versions were taken into account. Altogether 9 differently structured units were recorded in 4 different forms (with different names) by 25 speakers in the task (900 elements).

Sound samples were recorded in lab circumstances with headphones. The annotation was made automatically by the MAUS online program, then corrected manually in Praat [7]. The following acoustic parameters of words and syllables were examined:

duration of the elements (words, syllables and pauses), features of the fundamental frequency (mean, range (semitones), min, max), appearance of creaky voice and intensity.

III. Results

Some of the key results are: In the reference sentence, the duration of the words were almost equal in all cases (no effect of position), while the modified structure induced increasing and decreasing tendencies, as well. The duration of each detached element was about 75% of the original ones, while the duration of the words directly before the embedded structure or after them were about 5-12.5% longer than the reference in the 4-element-units and a bit longer in the 3-element-long forms (Fig. 1.). The ratio of the f0 variance was significantly higher in the case of left-headed than in the right-headed structures, the average f0 of the first member of the related elements was lower, and that of the other was higher. The creaky voice showed notable individual differences and occurred more often at the end of right-headed structures. Intensity played a greater role in the realization of 4-element than 3-element units, and together with silent pauses, had a bigger effect on the realization of left-headed structures.

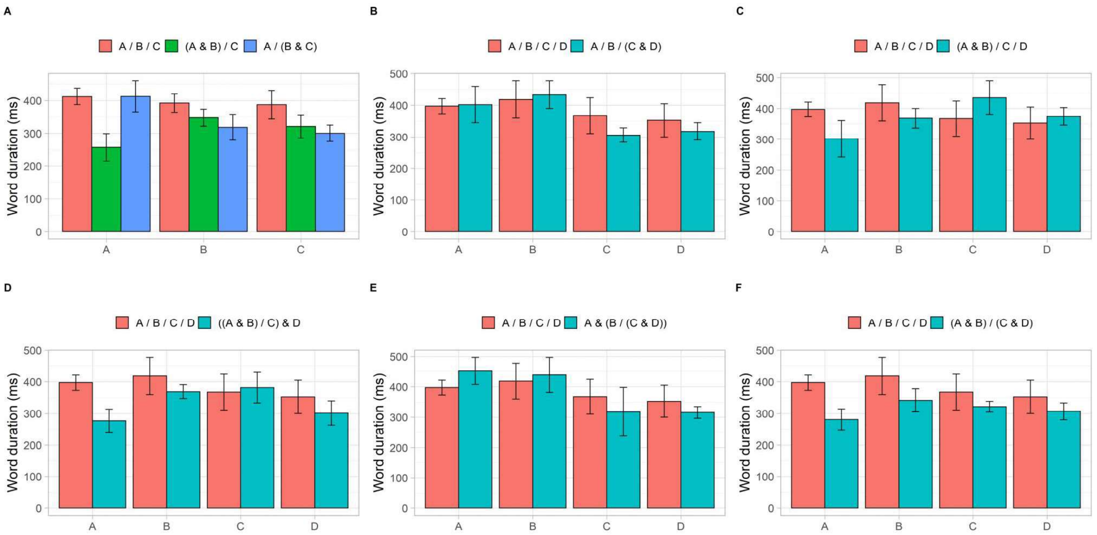


Fig. 1. Word duration regarding the reference structure (in reddish) and the modified structures (blueish) (firstly with 3-element-, then 4-element- units)

IV. Conclusions

In conclusion, the results for Hungarian showed that prosodic structure reflects syntactic grouping and embedding for timing patterns, pitch and voice quality. For single boundary structure (single bracket structure), the boundaries were well separated from the reference structure with stronger prosodic distinction, but for the production of secondary boundary markings, the (double bracket structure) prosodic indication of detachment was less perceptible. Proximity and similarity explain the implementation of prosodic structure that unfolds from the syntactic structure. Although some uniform tendencies emerged from the use of prosodic features, mainly in terms of word and pause duration and pitch, significant individual differences were observed in the strength of boundary markings and patterns of prosodic characteristics, as well. Based on the examination of the similarities and differences between the languages, the results of the research contribute to the teaching of Hungarian as a foreign language and to the more accurate operation of the text-to-speech process.

References

[1] D. Ladd, Compound prosodic domains. Occasional Papers of the Department of Linguistics. Edinburgh, Scotland: University of Edinburgh, 1992.
[2] M. Wagner, Prosody and recursion. Doctoral Dissertation. Cambridge, Mass.: Massachusetts Institute of Technology, 2005.
[3] C. Féry, and G. Kentner, “A new approach to prosodic grouping,” The Linguistic Review, vol. 30, pp. 277-311, 2013.
[4] C. Féry, and H. Truckenbrodt, “,” Studia linguistica vol. 59, pp, 2005.
[5] C. Féry, and H. Kentner, “The prosody of embedded coordinations in German and Hindi” Speech Prosody, 2010 USA, 2010.
[6] K. Mády, “Prosodic (non-)realisation of broad, narrow and contrastive focus in Hungarian: a production and a perception study,” Proc. Int.peech, Dresden, pp. 948–952, 2015.
[7] P. Boersma, and D. Weenink, “Praat: doing phonetics by computer,” Ver. 6.2.09, downloaded: 22 Feb. 2022 from http://www.praat.org/, 2022.

Acknowledgement

The research was supported by the Alexander von Humboldt Foundation.

Rhythm structure in Thai and Indian Englishes

Daniel Denian Lee, |Nanyang Technological University

Keywords — rhythm metrics, prosody, world Englishes, applied phonetics, language pedagogy

I. Background and objectives

The ontology of speech rhythm has been, and continues to remain, a controversial topic within laboratory phonology and theoretical linguistics more broadly [1, 2]. And nested within the debate on the metaphysical nature of linguistic rhythm is another point of contention: The usefulness of rhythm metrics [3], which extends to applications in linguistic research and language pedagogy. Against the backdrop of these developments in the fields of theoretical and applied linguistics, this paper aims to investigate the utility of rhythm metrics in world Englishes (WE) scholarship, and in doing so, develop an apologia for the use of rhythm metrics in clinical linguistics, language pedagogy, and other applied fields concerned with addressing rhythm-related issues.

To those ends, two state-of-the-art rhythm metrics—the pairwise variability index [PVI; 4] and variation coefficient for vocalic intervals [VarcoV; 5]—will be applied to acoustically examine the rhythm structure in Thai and Indian Englishes, which are relatively understudied varieties in the applied phonetics literature. By conducting this contrastive study on the two Asian Englishes, it will be demonstrated that utilising rhythm metrics, namely the PVI and VarcoV, is methodologically sound insofar as cross-varietal phonetic studies are concerned, thereby defending the continued employment of these tools to quantitatively classify rhythm structure in speech and ultimately use these metric-based findings to empirically inform language pedagogy in second- and foreign-language learning contexts, and other areas of applied phonetics more widely.

II. The past, present, and future of rhythm research

A. The three waves of rhythm research

Post and Payne [2] provide an overview of the history of rhythm research in the phonetics and phonology literature, broken down in terms of three distinct waves. Their account of the historical trajectory of rhythm scholarship is recapitulated in this subsection. The first wave of speech rhythm research can be most usefully defined as the era that applied the terminology 'stress-timed', 'syllable-timed', and 'mora-timed' [e.g. 6] based on the rhythm class hypothesis that describes the temporal organisation of speech rhythm percepts. Inspired by this hypothesis, the second wave subsequently saw a burgeoning of methods developed to quantitatively capture the production and perception of speech rhythm, chiefly through rhythm metrics. In the lead up to the metrics being actually proposed for application to phonetic research, [7] first provided evidence of the possibility of using durational information from acoustic data to quantitatively characterise speech rhythm across a continuum that has 'stress-timed' and 'syllable-timed' as its endpoints. In the following year, [8] made a similar attempt to capture rhythmic properties using consonantal and vocalic intervals. In the year after that, [4] established the PVI as a way of quantitatively capturing the durational characteristics of linguistic rhythm in Singapore English speech, and the PVI was then extended to analyse several other languages in [9]. In 2006, [5] formulated the variation coefficient method of describing speech rhythm, based on consonantal intervals (deltaC), and his metric has since been extended by speech rhythm researchers to vocalic intervals, thereby establishing the VarcoV.

Finally, the third wave may be characterised by scepticism [e.g. 3] towards the metric focus that defines the second wave. This sceptical attitude is mainly founded on putative methodological issues in using rhythm metrics for phonetic scholarship, especially cross-linguistic research [3]. For instance, because of the non-symmetry found in the phonological environment of passage stimuli across different languages, it is said that appropriate comparisons cannot be made across languages. Inasmuch as varietal phonetic research is concerned, however, this criticism should no longer be relevant as all the varieties in question should be using a single set of visual/written stimuli for speakers to produce speech. In the case of a WE project, the phonetician ideally has at least one set of stimuli that is orthographically identical for speakers of all the varieties investigated. Therefore, echoing the sentiments expounded in [2], the second and third waves need not be viewed as necessarily competing perspectives, but as complementary and belonging to an eclectic assemblage of tools to quantify speech rhythm. Hence, this paper will illustrate the utility of rhythm metrics as a way of providing an objective measure of speech rhythm in contrastive phonetic analyses.

B. Rhythmic patterning in Thai and Indian Englishes

Preliminary glimpses at what the rhythmic structure might look like in Thai and Indian Englishes (henceforth ThaiE and IndE respectively) are available, but they are few and far between. Moreover, earlier studies that have looked at the speech rhythm in both varieties have done so without comprehensive contrastive analyses. Notwithstanding these considerations, some observations on ThaiE and IndE are available at hand. In the Outer Circle literature, [10] reported evidence of mixed rhythmic characteristics in both American English and IndE speakers in her study, though the IndE speakers' speech exhibited an overall tendency to veer towards the syllable-timed end of the rhythmic patterning continuum. Fuchs [11] conducted a large-scale contrastive study that examined (educated) IndE speech rhythm opposite that of British English, drawing an Outer-Inner Circle comparison. His findings observed the syllable-timed nature of IndE speech rhythm, corroborating [10], and also highlighted that there are multiple exponents in the production and perception of linguistic rhythm, a view that can be traced back as early as in [4], which recognised the applicability of the PVI not just to inter-segmental duration but also amplitude, among other types of acoustic-phonetic data. However, direct comparisons between varieties of English outside of the so-called Anglosphere (Inner Circle) remain decidedly rare, and a contrastive study on the rhythmic properties of IndE against a variety like ThaiE, an Expanding Circle variety, has not been done yet. Within the Expanding Circle domain, there is a demonstrable dearth of fine-grained acoustic studies on ThaiE linguistic rhythm, and ThaiE speech in general. That said, [12] conducted a cross-varietal study that focused on varieties/languages in the Outer and Expanding Circles, comparing ThaiE to Thai, and Hong Kong and Singapore Englishes. Their usage of the PVI yielded robust findings across the Asian varieties, showing that ThaiE showed higher normalised PVI values compared to other varieties/languages in the study, thus evincing a more stress-timed tendency in ThaiE rhythm. Overall, while some studies have been done on the rhythmic patterning in IndE and ThaiE, a finer-grained inter-varietal examination is necessary to appreciate not only the diachronic developments of prosody in both varieties, but to re-examine the place of rhythm metrics in WE and applied phonetics research more widely. The following research questions (RQ) are thus posed:

1. What is the rhythmic patterning in ThaiE and IndE?
2. Do rhythm metrics provide an objective basis of rhythm comparison between varieties?
3. What implications do metric-based speech rhythm findings have for applied phonetics?

C. Methodology and implications of this study

Acoustic phonetic analysis will form the primary mode of inquiry in this study. Principles of vocalic measurements expounded in [10] will be strictly adhered to. Acoustic data retrieved from the NIESCEA [13] will be used to ascertain the rhythmic patterning in ThaiE and IndE. Ten speakers of ThaiE and IndE, with an equal distribution in gender in each variety, will be examined. ThaiE and IndE speakers' production of bespoke sentences devised in [10], designed specifically to elicit segmental durational variability where possible, will be subject to PVI and VarcoV analysis to ultimately yield quantitative characterisations of speech rhythm for both varieties. The PVI, normalised for speaking rate, i.e. nPVI, is expressed formulaically as:

$$nPVI = 100 \times \left[\sum_{k-1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| /(m-1) \right] \tag{1}$$

where m = number of vowels in an utterance and d = duration of the kth vowel. The VarcoV formula is given below:

$$varco\Delta V = \frac{\Delta V * 100}{meanV} \tag{2}$$

where V = duration of vocalic intervals and the multiplication of 100 generates the output as a percentage. For both metrics, a higher numerical output is taken to represent a tendency towards a stress-timed rhythmic patterning in a speaker or variety. The nPVI and VarcoV results of ThaiE and IndE will be compared to ascertain whether their rhythmic properties are quantitatively similar despite their divergent sociohistorical backgrounds. After a review of the empirical results to answer RQ1 and RQ2, practical implications for subfields within applied phonetics will be discussed, with special attention paid to language pedagogy and speech-language pathology (clinical phonetics).

References

[1] Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: A metaphor? Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1658), 20130396.

[2] Post, B., & Payne, E. (2018). Speech rhythm in development: What is the child acquiring? In N. Esteve-Gibert & J. Pilar Prieto (Eds.), *The development of prosody in first language acquisition* (Vol. 23, p. 125). Benjamins Publishing Company.

[3] Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351–373.

[4] Low, E. L., Grabe, E., & Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech*, *43*(4), 377–401.

[5] Dellwo, V., Karnowski, P., & Szigeti, I. (2006). *Rhythm and speech rate: A variation coefficient for deltaC* (pp. 231–241). Peter Lang.

[6] Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press.

[7] Low, E. L. (1998). *Prosodic prominence in Singapore English* [Unpublished doctoral dissertation]. University of Cambridge.

[8] Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292.

[9] Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, *7*(1982), 515–546.

[10] Krivokapić, J. (2013). Rhythm and convergence between speakers of American and Indian English. Laboratory Phonology, 4(1), 39–65.

[11] Fuchs, R. (2016). Speech rhythm in varieties of english. Springer.

[12] Sarmah, P., Gogoi, D. V., & Wiltshire, C. R. (2009). Thai English: Rhythm and vowels. *English World-Wide*, *30*(2), 196–217.

[13] Low, E. L. (2015). The NIE Spoken Corpus of English in Asia (NIESCEA).

Improving Spanish L1 learners' perception and production of Estonian vowels with the CAPT tool Estoñol: a pilot study

Katrin Leppik[1], Cristian Tejedor-García[2], Eva Liina Asu[1], Pärtel Lippus[1]
[1]University of Tartu, Institute of Estonian and General Linguistics, [2] Radboud University Nijmegen, Centre for Language and Speech Technology
katrin.leppik@ut.ee

Keywords —Estonian, Spanish, L2, perception and production, CAPT tool

I. Introduction

Computer-assisted pronunciation training (CAPT) tools have been shown to help second language (L2) learners to improve L2 pronunciation and perception [1], [2]. The current study investigates the effectiveness of a CAPT tool developed for the Spanish L1 learners of Estonian for practising Estonian vowels. Estonian and Spanish differ in their vowel inventory size: Spanish has 5 vowels /i, e, a, o, u/ while the Estonian vowel system consists of 9 phonemes /i, y, e, ø, æ, ɑ, ɤ, o, u/ [3], [4]. As it is known that in L2 acquisition, the learners' mother tongue (L1) plays a very important role [5], [6] it is vital to take this into account even when developing CAPT tools. Taking into account the differences of Estonian and Spanish a CAPT tool Estoñol [7] was developed to help native speakers of Spanish to train their perception and production of Estonian vowels. The tool's training program is partially based on the Native Cardinality Method [8] and involves seven vowel contrasts, /i-y/, /u-y/, /ɑ-o/, /ɑ-æ/, /e-æ/, /o-ø/, and /o-ɤ/, which have proven to be difficult for native speakers of Spanish who learn Estonian [9]. The training activities of the tool include theoretical videos and four training modes (exposure, discrimination, pronunciation and mixed) in every lesson [7]. The tool is integrated into a pre/post-test design experiment with native speakers of Spanish to assess the learners' perception and production improvement. It is expected that the tool will have a positive effect on the results, as has been shown in previous studies using similar methodology [1], [2], [10]. The goal of this study is to investigate to what extent the use of the tool Estoñol affects the results of the post-test.

II. Experiment

Six Spanish L1 speakers who had just begun learning Estonian participated in this pilot study. The participants attended Estonian classes twice a week and it was their first semester in Estonia. The participants were aged between 19 and 26, five were males and one female. All the participants were tested before and after they started using the CAPT tool Estoñol. The pre- and post-testing consisted of two perception tasks and a reading task. In the vowel identification tasks nine Estonian vowels were presented for three times in a random order, the participants could listen to each sound only once. In the second perception task the participants saw a minimal pair on the computer screen and heard a word, they were instructed to click on the word that they heard. In total 28 minimal pairs (56 words) were presented randomly. The reading task consisted of reading 56 words that were the same ones that formed minimal pairs of the perception task, e.g. *kola*, *kõla*, *melu*, *mälu*. During one week the participants had three training sessions of 60 minutes. The training sessions took place at the University of Tartu; the participants used a tablet and headphones. Between the training sessions there was one day when there was no training. In each session the participants worked with 2-3 vowel contrasts: first session /i-y/ and /u-y/, second session /ɑ-o/, /ɑ-æ/ and /e-æ/ and, third session /o-ø/ and /o-ɤ/. A research member was present and explained the procedure to participants at the beginning of the session.

III. Preliminary results

Fig. 1 presents the results of the vowel identification task grouped by participants and Fig. 2 presents the results by vowel categories. Focusing on the results of the pre-test it can be seen that the vowels that are identical in Estonian and Spanish (/i, e, o, u/) have the highest percentage of correct responses and the vowels that are "new" to the Spanish learners have less correct responses. The vowels /ø/ and /ɤ/ have the lowest number of correct responses and were identified correctly only in 28% and 33% of the cases respectively. These vowels were often confused with each other and with /y/ and /u/. The vowel /y/ was mostly confused with /i/, the vowel /æ/ with /ɑ/, and the vowel /ɑ/ with /o/. The Wilcoxon signed-rank test showed that there are statistically significant ($Z = -4.5$, $p < 0.001$) differences between the results of pre-test and post-test: four of the six participants achieved better results in the post-test than in the pre-test, while one participant had the same results in both tests and one participant had fewer correct responses in the post-test. Focusing on vowel categories the post-test results show that compared to the pre-test there are less mistakes in identifying /æ/, /ø/, /u/ and /y/, the results of /e/, /i/ and /ɤ/ remain the same and in the identification of /ɑ/ and /o/ there are more mistakes. The results of the minimal pairs identification task and the reading task are still being analysed and we hope to present them at the symposium.

IV. Discussion and summary

The preliminary results of the vowel identification task showed that the Spanish L1 learners' identification of Estonian vowels improved. The Revised Speech Learning Model [6] suggests that there is a strong bi-directional connection between production and perception. It is expected that after the training sessions the Spanish L1 learners have improved their production as well. A more intensive and longer training period might help to improve the Spanish L1 learners' perception and production of Estonian vowels even more.

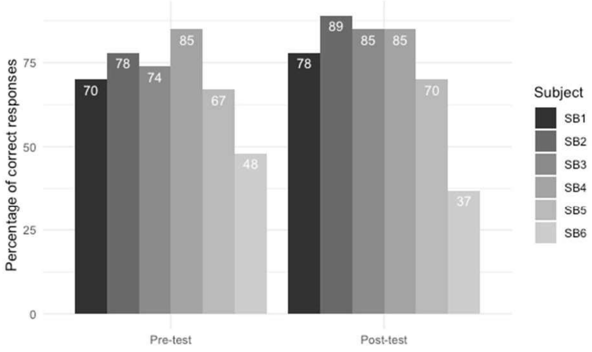


Fig. 2. The percentages of correct responses grouped by participants (marked by a gray scale) and testing time (left panel – pre-test, right panel – post-test).

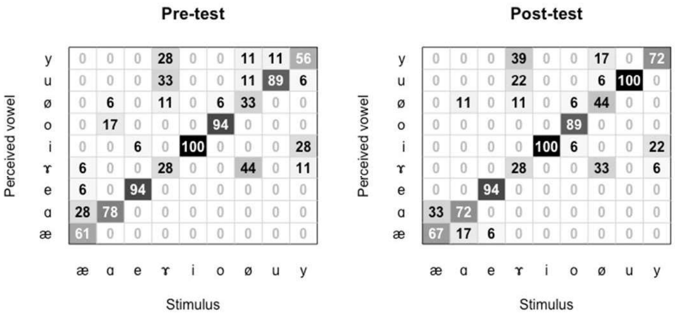


Fig. 3. The results of the vowel identification task grouped by stimulus vowel and testing time (left panel – pre-test, right panel – post-test). The figures indicate the number of responses (percentages) and the colour shows the mean reaction time of the responses (the darker the shade the faster the reaction time).

References

[1] C. Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, E. Cámara-Arenas, and V. Cardeñoso-Payo, 'Evaluating the efficiency of synthetic voice for providing corrective feedback in a pronunciation training tool based on minimal pairs', 2017, doi: 10.21437/SLaTE.2017-5.

[2] C. Tejedor-García, V. Cardeñoso-Payo, and D. Escudero-Mancebo, 'Automatic Speech Recognition (ASR) Systems Applied to Pronunciation Assessment of L2 Spanish for Japanese Speakers', *Applied Sciences*, vol. 11, no. 15, 2021, doi: 10.3390/app11156695.

[3] T. Navarro, *Studies in Spanish phonology*. Coral Gables: University of Miami Press, 1968.

[4] E. L. Asu and P. Teras, 'Illustrations of the IPA: Estonian', *Journal of the International Phonetic Association*, vol. 39, no. 3, pp. 367–372, 2009.

[5] C. T. Best and M. D. Tyler, 'Nonnative and second-language speech perception: Commonalities and complementarities', in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, M. J. Munro and O.-S. Bohn, Eds. Amsterdam: John Benjamins Publishing, 2007, pp. 13–34.

[6] 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, pp. 3–83. doi: 10.1017/9781108886901.002.

[7] K. Leppik and C. Tejedor-García, 'Estoñol, a computer-assisted pronunciation training tool for Spanish L1 speakers to improve the pronunciation and perception of Estonian vowels', *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, vol. 10, no. 1, Art. no. 1, Dec. 2019, doi: 10.12697/jeful.2019.10.1.05.

[8] E. Cámara-Arenas, 'The NCM and the reprogramming of latent phonological systems: A bilingual approach to the teaching of English sounds to Spanish Students', *Procedia - Social and Behavioral Sciences*, vol. 116, pp. 3044–3048, 2014, doi: http://dx.doi.org/10.1016/j.sbspro.2014.01.704.

[9] K. Leppik, P. Lippus, and E. L. Asu, 'The production of Estonian vowels in three quantity degrees by Spanish L1 speakers', no. Proceedings of the 19th International Congress of Phonetic Sciences, 2019.

[10] C. Tejedor-García, D. Escudero-Mancebo, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, 'Assessing Pronunciation Improvement in Students of English Using a Controlled Computer-Assisted Pronunciation Tool', *IEEE Transactions on Learning Technologies*, vol. 13, no. 2, pp. 269–282, 2020, doi: 10.1109/TLT.2020.2980261.

Spanish Lexical Stress Produced by Proficient Mandarin learners of Spanish

Peng Li[1] and Xiaotong Xi[2]
[1] Center for Multilingualism in Society across the Lifespan, University of Oslo, Norway
[2] Department of Translation and Language Sciences, Universitat Pompeu Fabra, Spain

Keywords — vowel, stress, Spanish, Chinese learner of Spanish, second language pronunciation

I. Introduction

Lexical stress as one of the most important prosodic features has received much attention in Spanish as a second language (L2) studies [1]. When perceiving and processing L2 lexical stress, learners may unintentionally transfer the acoustic cues from their native languages (L1) [2]. For example, as a tonal language, Mandarin assigns lexical tones on syllable-level by manipulating pitch. By contrast, stress languages such as Spanish assigns stress on lexical level, with the stressed syllables having a higher pitch, longer duration, and greater intensity [3]. Therefore, compared to native speakers, Mandarin-speaking learners may more likely rely on pitch to perceive lexical stress [4] and even realize lexical stress as a combination of Mandarin lexical tones in speech production (e.g., producing stressed syllables as Tone 2, a rising tone) [5]. However, little research has systematically assessed which acoustic feature(s) (e.g., pitch, duration, intensity) Mandarin learners of Spanish use to produce the Spanish lexical contrast. At the segmental level, lexical stress also impacts vowel quality. Spanish stressed vowels are more open [6] and less centralized [7] than unstressed ones. Therefore, when interacting with lexical stress, the seemingly simple five-vowel inventory /i, e, a, o, u/ may still cause difficulties for Mandarin speakers whose L1 features a quite different vowel inventory /i, y, ə, a, u, (ɚ)/ [8]. However, previous studies (e.g., [9] [10]) did not agree on how Mandarin learners produce the five Spanish vowels in different stress conditions and speech (e.g., isolated words vs. running speech). Therefore, this paper aims to assess how Mandarin speakers produce the Spanish lexical stress with the following research questions:

RQ1: Which phonetic feature(s) do Mandarin learners of Spanish use to mark Spanish lexical stress? Based on previous research [2], we hypothesized that for Mandarin learners of Spanish, pitch would be the most salient feature to make stress contrasts in speech production.

RQ2: To what extent does lexical stress affect Spanish vowel quality produced by L1 and L2 speakers? We hypothesize that lexical stress would have different effects on L1 and L2 vowel quality given the distinct vowel inventories in Spanish and Mandarin.

II. Method

The test materials consisted of 30 Spanish $C_1V_1C_2V_2$ words and a short Spanish text. In the wordlist, 15 of the words were oxytones, while the other 15, paroxytones. The target vowel was always set as V_1, so that half of the vowels were stressed, while the other half, unstressed. C_1 and C_2 were chosen from the three voiceless plosives /p, t, k/, because these three consonants are shared phonemes in Mandarin and Spanish. Each of the three voiceless plosives cooccurred with each of the five vowels twice, once in oxytone, once in paroxytone, resulting in 30 C_1V_1 syllables: 3 plosive /p, t, k/ × 5 vowels /i, e, a, o, u/ × 2 stress conditions (stressed vs. unstressed). We selected real Spanish words to avoid unnatural pronunciation. The final word list contained 128 words, with 98 being fillers, organized in random order. The short text was the standard text for eliciting pronunciation of different languages, "*El viento norte y el sol*" (*North Wind and the Sun*).

Ten Mandarin learners of Spanish (female = 5, M_{age} = 27.30) and six Spanish native speakers (female = 3, M_{age} = 24.83) were recruited from Spain. The Mandarin speakers were late adult learners with advanced proficiency and intensive exposure to Spanish. They read the wordlist twice and the text once at a natural and comfortable speech rate. Their speech outcome was recorded automatically. Finally, we obtained 960 and 2592 vowels from the word-reading task (30 words × 16 participants × 2 repetitions) and the text-reading task (162 vowels × 16 participants). After acoustic analyses, we obtained the following data for each vowel: duration, mean pitch, intensity, the mid-points of the first (F1) and the second formant (F2).

III. Results

We analyzed the data with Linear Mixed-Effects Models. For duration, pitch and intensity, the fixed effects were repetition (only for word-reading), speaker (L1 vs. L2), stress (stressed vs. unstressed), and the two-way interaction of Speaker × Stress. For F1 and F2, the fixed effects also included vowel (a, e, i, o, u) and the three-way interaction of Vowel × Speaker × Stress. Each model featured participant and item as random intercepts with appropriate random slopes.

A. The effects of stress on duration, pitch and intensity in L1 and L2 speech

Duration. The word-reading task revealed a significant interaction of Stress × Speaker for duration ($p < .001$). Post-hoc comparisons showed that both L1 and L2 speakers produced the target vowels with longer duration in stressed syllables than in unstressed ones (all $p < .05$). However, L2 speakers produced significantly longer vowels than natives in unstressed syllables (*p*

= .002). By contrast, the text-reading task did not show such an interaction (p = .397), although both groups of speakers produced longer vowels in stressed syllables than in unstressed ones (all p < .001). These results suggest that L1 and L2 speakers of Spanish marked stress with duration to a similar extent.

Pitch. There was a significant interaction of Stress × Speaker for pitch in both word-reading (p < .001) and text-reading (p < .001) tasks. Post-hoc comparisons showed that L2 speakers produced stressed vowels with a higher pitch than unstressed ones when reading words and text, but L1 speakers revealed the same pattern only when reading isolated words (all p < .05). That is, although pitch is relevant to mark Spanish stress, L2 speakers used this cue to a larger degree than L1 speakers.

Intensity. We only found a significant main effect of stress for intensity in both word-reading (p < .001) and text-reading (p = .001) tasks, with stressed vowels showing greater intensity than unstressed ones. This means that both L1 and L2 speakers made Spanish stress contrast with intensity in a similar way.

B. The effects of lexical stress on vowel quality in L1 and L2 speech

F1 (openness). In the word-reading task, we only found two significant main effects on F1: vowel and stress (both p < .001). The results indicate that in both L1 and L2 speech, the five vowels had different degree of openness and that stressed vowels were more open than unstressed ones. By contrast, in the text-reading task, there was a significant interaction of Stress × Speaker × Vowel (p = .012). Briefly, the L2 speakers produced stressed /e, i, o/ more openly than unstressed ones, while the L1 speakers did so with /a, o/. These results indicate that stress influenced the mouth aperture of Spanish vowels, and its impact varied across speakers (L1 vs. L2) and speech type (isolated words vs. running speech).

F2 (tongue position). The analysis of F2 revealed a significant interaction of Vowel × Speaker in both word-reading (p = .04) and text-reading (p < .001) tasks. However, stress was not significant in either task. Post-hoc results showed that when reading isolated words, L2 speakers produced /u/ with significantly higher F2 than L1 speakers (p = .006); while when reading text, L2 speakers showed near significantly lower F2 values for /o/ than L1 speakers (p = .052). In short, stress did not significantly affect the tongue position in either L1 or L2 speech but the two groups of speakers realized specific vowels with different tongue positions.

IV. Discussion and conclusion

This study compared the production of Spanish lexical stress by Mandarin- and Spanish- speakers in isolated words and running speech. The results showed that although with advanced proficiency and intense exposure to the target language, L2 learners still performed differently from native speakers. Regarding RQ1, our results confirmed that although Mandarin speakers could clearly distinguish stressed from unstressed vowels using various acoustic features, they seemed to rely more on the pitch to make stress contrasts compared to L1 speakers. As for RQ2, our results showed that lexical stress influenced vowel quality differently in L1 and L2 speech, especially in running speech. The main difference lay in the mouth aperture, with stressed vowels being more open. In addition, even for advanced learners, their vowel quality was still different from that of the native speakers. All in all, our findings provided new evidence for the *phonetic approach* [2] on L2 lexical stress production: Even proficient L2 learners still transferred the prosodic features from L1 to L2 in speech production (e.g., prefer pitch more than duration in stress contrast). In teaching practice, at least /o, u/ should receive more attention as even advanced learners still showed different backness of the two vowels compared to native speakers.

References

[1] M. Simonet, "The L2 acquisition of Spanish phonetics and phonology," in *The Handbook of Hispanic Linguistics*, J. I. Hualde, A. Olarrea, and E. O'Rourke, Eds. Blackwell Publishing Ltd, 2012, pp. 747–764.

[2] A. Tremblay, "Second language speech production," in *Second Language Speech Learning*, R. Wayland, Ed. Cambridge University Press, 2021, pp. 175–192.

[3] J. I. Hualde, "Stress and rhythm," in *The Handbook of Hispanic Linguistics*, 1st ed., J. I. Hualde, A. Olarrea, and E. O'Rourke, Eds. Oxford, UK: Blackwell Publishing Ltd, 2012, pp. 153–171.

[4] Y. Zhang, S. L. Nissen, and A. L. Francis, "Acoustic characteristics of English lexical stress produced by native Mandarin speakers," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4498–4513, 2008, doi: 10.1121/1.2902165.

[5] Y. Chen, "From tone to accent: the tonal transfer strategy for Chinese L2 learners of Spanish," in *16th International Congress of Phonetic Sciences*, 2007, no. August, pp. 1645–1648.

[6] 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.

[7] E. W. Willis, "No se comen, pero sí se mascan: Variación de las vocales plenas en la República Dominicana," in *Actas del XV Congreso Internacional de la Asociación de Lingüística y Filología de América Latina (ALFAL)*, 2008, pp. 18–21.

[8] S. Duanmu, *The Phonology of Standard Mandarin*. Oxford, UK: Oxford University Press, 2007.

[9] Q. Xia and F. Shi, "An experimental study on Chinese learners' language transfer in pronouncing Spanish vowels," *Foreign Lang. Teach. Res.*, vol. 39, no. 5, pp. 367–373, 2007.

[10] Y. Cao and A. Rius-Escudé, "Caracterización acústica de las vocales del español hablado por chinos," *Phonica*, vol. 15, pp. 3–22, 2019, doi: 10.1344/test.2019.0.3-22.

Pronunciation instruction in Swedish for immigrants

Annika Norlund Shaswar, Umeå University
Christina Sörvåg, Umeå University

Keywords — Swedish for immigrants, basic literacy education, pronunciation instruction, adult second language learners

I. Introduction

For second language learners, achieving intelligible pronunciation is important in order to be able to communicate. Research has shown that pronunciation instruction can support second language learners in developing their pronunciation [eg 1; 2]. However, many teachers find pronunciation instruction difficult, and this is also true for teachers of Swedish for immigrants (SFI). SFI is a language programme for adult second language learners of Swedish [3]. SFI students are a heterogeneous group of learners in terms of previous education. The aim of the study presented here is to explore how SFI teachers who teach students with very short previous education describe their pronunciation instruction, and specifically how they describe connections between pronunciation instruction and basic literacy teaching. The research questions are:

- Which views do the teachers express concerning the importance of pronunciation instruction?
- Which views do they express concerning the content of pronunciation instruction?
- Which views do the teachers express concerning connections between pronunciation instruction and basic literacy teaching?

A. Theoretical framework

In this study the view on language and second language development starts out from a combination of theoretical perspectives. On the one hand, from a sociocultural perspective on second language development, language is understood as communication and meaning making [4] . On the other hand, language is also understood as form, phonetic and phonematic aspects, which are necessary for the production of intelligible spoken language. In addition, the understanding of connections between spoken language and literacy are in this study seen from a social practices perspective based in the research field New Literacy Studies [5]. From this perspective it is underlined that the development of literacy is integrated and interconnected with the development of spoken language.

B. Methodology

The study was performed as part of a larger action research project, including four SFI schools, and the aim of the project was to explore and develop basic literacy education in SFI. For the present study, five teachers who teach at one SFI school participated in individual qualitative interviews and focus group interviews. For the analysis of the interviews, qualitative content analysis was used [6]. The analysis was also inspired by a social constructivist perspective which sees the teachers' expressed views on pronunciation instruction as connected to a specific context and thereby constructed in relation to norms and understandings which dominate a specific society as a certain point of time [7].

II. Preliminary results

Preliminary results show that the teachers describe pronunciation instruction as important. In terms of the content of their pronunciation instruction, they mainly stress prosodic aspects as important. In relation to connections between pronunciation instruction and basic literacy teaching, they underline the need of starting out from spoken language, and to avoid mixing exercises where students need to read texts and practice pronunciation. They express that because the students still need to practice reading on a basic level, it is too difficult to use written material as a starting point for practicing pronunciation. The teachers say that if they avoid written language and instead start out from spoken language in their teaching, the students can easier learn to produce correct consonant and vowel reductions and omissions. One of the teachers also states that separating spoken language from written language in pronunciation exercises has supported her students in developing metalinguistic awareness, such as the difference between verbs in present tense (läser, reads) and infinite form (att läsa, to read).

III. Conclusion

One of the conclusions drawn from the study is that there is need of more research on and development of effective teaching practices for pronunciation instruction for adult second language learners with short previous education.

REFERENCES

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

[2] R. I. Thomson, R. I., & T. M. Derwing, T. M. "The effectiveness of L2 pronunciation instruction: A narrative review". *Applied Linguistics*, *36*(3), 326-344. 2015.

[3] E. Zetterholm, "Swedish for immigrants. Teachers' opinions on the teaching of pronunciation. in Proceedings of the International Symposium on Monolingual and Bilingual Speech. Institute 2017, pp. 308-312.

[4] M. Swain, P. Kinnear,. & L. Steinman, L. Sociocultural Theory in second language education: an introduction through narratives. (2nd ed.) Bristol: Multilingual Matters. 2015.

[5] D. Barton, *Literacy.* An introduction to the Ecology of Written Language. Oxford: Blackwell Publishing. 2007.

[6] K. Krippendorff Content analysis, an introduction to its methodology, London: Sage publications, 2013

[7] S. V. Hunter, Analysing and representing narrative data: The long and winding road. Current Narratives, vol 2, pp 44-54. 2009.

A comparison of lexical tone effects on VOT in L1 and three groups of L2 speakers of Mandarin Chinese

Chiu-ching Tseng[1] and Rina Yamileth Tseng[2]
[1] Department of English Language, Literature, and Linguistics, Providence University, Taiwan
[2] Spanish and Portuguese Department, Georgetonw University, USA

Keywords — ***Voice Onset Time (VOT), lexical tone, L1 influence, L2 Mandarin, Interlanguage***

I. Introduction

This study focuses on the effect of lexical tone on Voice Onset Time (VOT) in Mandarin speakers from four different L1 backgrounds. It surveys VOT variations between L1 and L2 speakers and demonstrates that not only VOT was affected by lexical tone across all language groups, their VOT values were also significantly different between groups. While previous studies reported different results for the tone effect on VOT in L1 Mandarin, particularly between the tone2 and tone3 pair and the tone1 and tone4 pair of Chinese Mandarin [1] [2] [3] [4] [5] [6] [7] [8] [9] [10], this study finds consistent tone effects where VOTs in tone2 and tone3 were significantly longer than those in tone1 and tone4. Moreover, although it has been suggested that VOT varies because of different lexical tones, a question remains as to whether L2 Mandarin from different language backgrounds also exhibits the same effect. Exploring this can potentially shed light on whether tone effect on VOT is a language specific or universal phenomenon. In particular, we ask whether L2 speakers who feature different native VOT values, such as Spanish (unaspirated) [11], Japanese (weakly-aspirated) [12], and English (aspirated) [13] [14] would show the same, similar, or different tone effect patterns in comparison to Mandarin (highly-aspirated) [2].

The experiments elicited stop-initial words produced with one open-unrounded vowel, three places of articulation (POA), four lexical tones, three different speech rates, and three utterance positions for both L1 and L2 speakers. A series of linear mixed-effects regression models were employed to model the effects of the properties mentioned on VOT duration [15] [16]. We wanted to explore whether or not the mentioned factors affect VOT in native and in non-native speech universally or language-specifically.

Testing 164 participants (68 Taiwanese, 34 Spanish, 40 Japanese, and 22 English speakers of Mandarin), the results reveal that when other factors were kept constant, tone indeed influenced VOT where tone2 and tone3 had significantly longer VOT values than those in tone1 and tone4 in all four groups (figure 1). Our findings suggest that the higher the onset tone pitch of a lexical tone, the shorter the VOT. POA and the speech rate were also found to be highly significant factors. The results also disclose that all non-native groups showed the same effects regardless of their L1. This finding suggests that the tone effect on VOT in Mandarin is a universal tendency due to the physiology of the vocal tract rather than due to language-specific phonology. However, we also found that the Spanish and Japanese groups showed extended VOT values, which were not from their native VOTs; nonetheless, their Mandarin VOTs were still significantly shorter than the native Mandarin speakers. This may due to the deficient learning of the aspiration. Thus, the significant VOT differences between groups indicate some degree of L1 influence, which suggests that L2 VOT delay is probably mediated by language-specific phonological grammar [17].

This study provides empirical evidence that an acoustic property, such as VOT, is not an isolated phenomenon but is involved with other complex phonological categories such as lexical tone. It discusses how the effects operate within phonetic and phonological theories. Additionally, it compares Mandarin learners' L1 and L2 VOT directly by conducting a VOT baseline test to offer more comparative VOT values from the same L2 groups, in some regards (figure 2). This cross-linguistic survey offers insight for L2 performance variations regarding phonetics, which may provide Mandarin instructors with multi-dimensional comparisons and confirmation of the interlanguage process as relevant to Second Language Acquisition [18]. The observed phenomena may aid L2 classrooms insights into Mandarin accent variations for L2 English, Japanese, and Spanish learners of Mandarin.

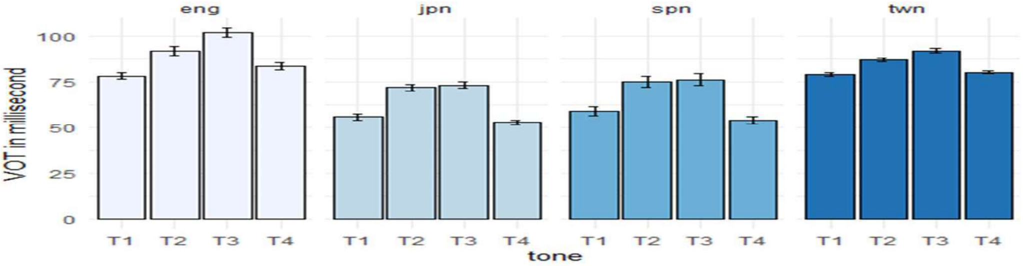


Figure 1. Overall VOT values by lexical tone and by language group

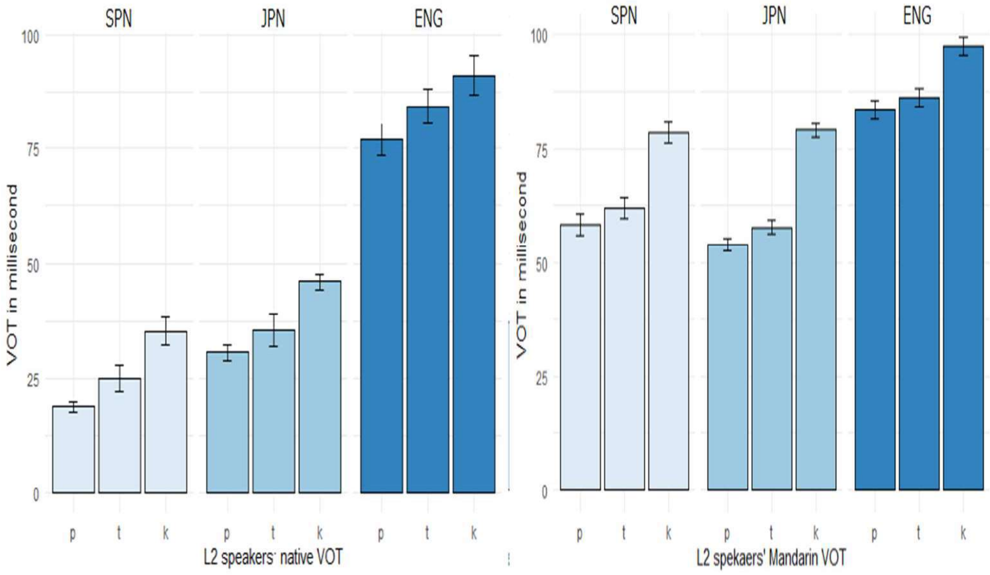


Figure 2. Comparison of L2 groups' native & Mandarin VOTs by language group & by POA

SELETED REFERENCES

[1] Chao, K. Y., Khattab, G., Chen, L. M. (2006, January). *Comparison of VOT patterns in Mandarin Chinese and in English*. In Proceedings of the 4th Annual Hawaii International Conference on Arts and Humanities (Vol.840, p. 859).

[2] Cho, T., & Ladefoged, P. (1999). *Variation and universals in VOT: evidence from 18 languages*. Journal of phonetics, *27*(2), 207-229.

[3] Chen, K., Tsay, J., & Hong, G. (1998). *Duration of initials in Mandarin: Fundamental acoustic research and its clinical significance*. Journal of Speech Language-hearing Association, 13, 138-149.

[4] Chen, L. M., Peng, J. F., & Chao, K. Y. (2009, December). *The effect of lexical tones on voice onset time*. In 2009 11th IEEE International Symposium on Multimedia (pp. 552-557). IEEE.

[5] Lam, C. L. *Effect of tones on voice onset time (VOT) in Cantonese aspirated*. Journal of Phonetics, 27, 207-229.

[6] Li, F. (2013). *The effect of speakers' sex on voice onset time in Mandarin stops*. The Journal of the Acoustical Society of America, 133(2), EL142-EL147.

[7] Liu, H., Ng, M. L., Wan, M., Wang, S., & Zhang, Y. (2008). *The effect of tonal changes on voice onset time in Mandarin esophageal speech*. Journal of Voice, 22(2), 210-218.

[8] Peng, J. F., Chen, L. M., & Lee, C. C. (2014). *Voice onset time of initial stops in Mandarin and Hakka: Effect of gender*. Taiwan Journal of Linguistics, 12(1), 63-79.

[9] Tse, H. (2005). *The Phonetics of VOT and Tone Interaction in Cantonese* (Doctoral dissertation, University of Chicago).

[10] Tseng, C. C. (2018). *The Effect of Lexical Tones on Voice Onset Time in L2 Mandarin Production by English Speakers*. In Proc. ISAPh 2018 International Symposium on Applied Phonetics (pp. 120-125).

[11] Abramson, A. S., & Lisker, L. (1973). *Voice-timing perception in Spanish word-initial stops*. Journal of Phonetics, 1(1), 1-8.

[12] Shimizu, K. (1990). *Cross-Language study of voicing contrasts of stop consonants in Asian languages*. Annexe Thesis Digitisation Project 2016 Block 7.

[13] Lisker, L., & Abramson, A. S. (1964). *A cross-language study of voicing in initial stops: Acoustical measurements.* Word, *20*(3), 384-422.

[14] Lisker, L., & Abramson, A. S. (1967). *Some effects of context on voice onset time in English stops*. Language and Speech, 10(1), 1-28.

[15] R Core Team (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.

[16] Winter, B. (2013). *Linear models and linear mixed-effects models in R: Tutorial 11*. arXiv preprint arXiv: 1308.5499.

[17] Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Routledge.

[18] Gass, S. M. (2013). *Second language acquisition: An introductory course*. Routledge.

The 4th International Symposium on Applied Phonetics (ISAPh 2022) Program (last update: 2022-08-29)

Organization: Lund University

Organizing committee: Shinichiro Ishihara, Nick Kalivoda, Sara Myrberg, Mikael Roll, Frida Splendido, Mechtild Tronnier, Elisabeth Zetterholm

DAY 1: Wednesday, September 14, 2022

Registration		8:00
Opening ceremony		9:00
Keynote lecture 1 Prof. Sonia Frota *Developing prosody in typical and atypical language acquisition*	MAJOR HALL (Piratensalen)	9:30
Coffee break	FOYER (Piratenfoajén)	10:30
Oral session O1A HALL A (Sten Broman)	**Oral session O1B** HALL B (Lukas Sal)	
Kirsty McDougall, Nikolas Pautz, Francis Nolan, Katrin Müller-Johnson, Harriet Smith and Alice Paver *The role of reflection and retention intervals on earwitness performance in voice parades*	Izabelle Grenon, Chris Sheppard and John Archibald *Learning Sounds through Unconscious Associations*	11:00
Nikita Suthar and Peter French *Role of Within Vowel Formants in Forensic Speaker Comparison: A Study Based on Vowels in Marwari as Spoken in Bikaner*	Sabine Gosselke Berthelsen and Mikael Roll *Prosody training aids second language processing*	11:30
Ben Gibb-Reid, Paul Foulkes and Vincent Hughes *And it's just like: The discourse-pragmatic and phonetic variation of just with applications to forensic voice comparison*	Xiaodan Zhang and Joaquin Romero *Explicit Rules or Implicit Imitation: a Comparative Study on the Approaches of Teaching English Prosody*	12:00
Lunch	HOTEL RESTAURANT	12:30
Oral session O2A HALL A (Sten Broman)	**Oral session O2B** HALL B (Lukas Sal)	
Noriko Yamane, Kunyang Sun, Jeremy Perkins and Ian Wilson *Pretest-posttest production and perception results of ultrasound pronunciation training*	Robert Squizzero *The role of dialectology in L2 vowel acquisition; evidence from Mandarin Chinese*	14:00
Anton Malmi, Pärtel Lippus and Einar Meister *Articulatory and temporal properties of Estonian palatalization by Russian L1 speakers*	Naoko Kinoshita and Chris Sheppard *Rapid Adaptation to NNS Japanese Pronunciation*	14:30
Alicia Janz, Simon Wehrle, Simona Sbranna and Martine Grice *The lexical and intonational 1ealization of backchannels is less constrained in spontaneous than task-based conversation*	Katja Haapanen, Antti Saloranta, Kimmo U Peltola, Henna Tamminen and Maija S Peltola *Revitalization and appreciation of local languages and phonetic training of English in Namibia*	15:00
Poster session P1 with coffee	FOYER (Piratenfoajén)	15:30
Keynote lecture 2 Prof. Paul Foulkes *Forensic speech science needs forensic phoneticians*	MAJOR HALL (Piratensalen)	17:00
Reception	GRAND HOTEL	18:30

Day 2: Thursday, September 15, 2022

Keynote lecture 3 MAJOR HALL (Piratensalen) Prof. Viveka Lyberg Åhlander *Speaker's comfort or listening effort? – On the interaction of the speaker, the classroom's sound environment and the students' learning*		9:00
Coffee break FOYER FOYER (Piratenfoajén)		10:00
Oral session O3A HALL A (Sten Broman)	**Oral session O3B** HALL B (Lukas Sal)	
Mehmet Yavas *#sC clusters in Spanish-English bilingual children with phonological disorders*	Åsa Abelin and Elisabeth Zetterholm *Intelligibility of Swedish foreign accented words*	10:30
Katsuya Yokomoto, Aki Tsunemoto and Yui Suzukida *Listening comprehension of World English pronunciation: How effective are awareness-raising activities?*	Heini Kallio, Mikko Kuronen and Liisa Koivusalo *The role of pause location in perceived fluency and proficiency in L2 Finnish*	11:00
Michael Ashby and Patricia Ashby *Practical phonetics in the 21st century*	Trisha Thomas, Gerard Llorach, Clara Martin and Sendy Caffarra *Does accented speech affect attention and information retention?*	11:30
Lunch HOTEL RESTAURANT		12:00
Oral session O4A HALL A (Sten Broman)	**Oral session O4B** HALL B (Lukas Sal)	
Lauren Harrington *A forensic-phonetic analysis of automatic speech transcription errors*	Joaquín Romero *Ceiling effects and the limitations of intelligibility and accentedness ratings for advanced L2 English learners*	13:30
Julio Cesar Cavalcanti, Anders Eriksson and Plinio A. Barbosa *Assessing the speaker discriminatory power asymmetry of different acoustic-phonetic parameters*	Rebeka Campos-Astorkiza *Gradience and L2 Learning of new phonetic categories vs. recategorization: L2 Spanish stops*	14:00
Sascha Schäfer and Paul Foulkes *Individual Voice Recognition Skills in Lay Speaker Identification Tasks*	Pekka Lintunen and Hanna Kivistö de Souza *EFL learners' L2 pronunciation noticing skills in an instructed learning context*	14:30
Poster session P2 with coffee FOYER (Piratenfoajén)		15:00
Keynote lecture 4 MAJOR HALL (Piratensalen) Assoc. Prof. Talia Isaacs *Revisiting second language pronunciation teaching and assessment: Constructs, compatibilities, contradictions, cross-fertilization*		16:15
General discussion MAJOR HALL (Piratensalen)		17:15
Conference dinner GRAND HOTEL		18:30

Day 3: Friday, September 16, 2022

Workshops		
Workshop 1 (part 1) Prof. Yi Xu *Computational tools for studying speech prosody*	MAJOR HALL (Piratensalen)	10:00
Coffee break	FOYER (Piratenfoajén)	11:00
Workshop 1: Prof. Yi Xu (part 2)	MAJOR HALL (Piratensalen)	11:30
Lunch	HOTEL RESTAURANT	12:30
Workshop 2 (part 1) Prof. Jacques Koreman *Creating pronunciation training content for your language of interest – A hands-on workshop*	MAJOR HALL (Piratensalen)	14:00
Coffee break	FOYER (Piratenfoajén)	15:00
Workshop 2: Prof. Jacques Koreman (part 2)	MAJOR HALL (Piratensalen)	15:30
Closing ceremony / Business meeting	MAJOR HALL (Piratensalen)	16:30

Poster session P1:

Atsushi Fujimori, Noriko Yamane, Brian Teaman and Ikuyo Kaneko
Short-term intervention effects on the development of pausing in read speech

Simon Gonzalez
A comprehensive web application for research on vocalic differences in World Englishes (Native and Accented)

Kornélia Juhász and Huba Bartos
The interplay of tone and intonation: f0 contours produced by Hungarian learners of Mandarin

Katrin Leppik, Cristian Tejedor García, Eva Liina Asu and Pärtel Lippus
Improving Spanish L1 learners' perception and production of Estonian vowels with the CAPT tool Estoñol: a pilot study

Peng Li and Xiaotong Xi
Spanish lexical stress produced by proficient Madarin learners of Spanish

Poster session P2:

Andrea Deme and Kornélia Juhász
Transition or insertion? Acoustic analysis of the vocalic section of Mandarin xia *in the production of Hungarian learners of Chinese*

Valéria Krepsz and Anna Huszár
Acoustical cues as boundary markers in a left-headed language

Daniel Lee
Rhythm structure in Thai and Indian Englishes

Annika Norlund Shaswar and Christina Sörvåg
Pronunciation instruction in Swedish for immigrants

Chiu-Ching Tseng
A comparison of lexical tone effect on VOT in L1 and three groups of L2 speakers of Chinese Mandarin