The Linguistic Proceedings Series
Year 2025, Volume 16, Issue 1

E-ISSN: 2529-1092
DOI: 10.36505

Proceedings Phonetics 2025 Hong Kong 16th International Conference on Linguistic Research and Applications

Series Editor
Antonis Botinis

The International Linguistic Society

Phonetics 2025 Hong Kong

Proceedings of 16th International Conference on Phonetic Research and Applications

17-19 September 2025
Hong Kong

International Linguistic Society

Phonetics 2025 Hong Kong
Proceedings of International Conference on Phonetic Research and Applications

Published by The Linguistic Society
Electronic edition
Phonetics 2025 Hong Kong
Athens, Greece
ISSN: 2529-1092
ISBN:
DOI: 10.36505/TheLinguisticProceedings/2025/16/01

Foreword

Welcome to the Phonetics 2025 Hong Kong International Thematic Conference on Linguistic Research and Applications. This hybrid event enables us to meet in person, thanks to the generous hospitality of The Hong Kong Polytechnic University, while also welcoming participants from around the world through our online platform. We are delighted to gather in the vibrant and dynamic city of Hong Kong.

As an international forum for linguists across career stages, the Society is collectively devoted to advancing linguistic research and its practical applications across diverse thematic areas. We encourage both emerging and established researchers to present and discuss current developments in phonetic research and related disciplines.

The origins of our conference series date back to 2006 in Athens, where the first Workshop on Experimental Linguistics (ExLing) was held under the auspices of ESCA (European Speech Communication Association), now ISCA (International Speech Communication Association). In subsequent years, the workshop was hosted in cities including Paris, Saint Petersburg, and Lisbon. In 2009, the ExLing Workshop evolved into the annual ExLing Conference with the establishment of the International Society of Experimental Linguistics (ExLing Society).

In 2024, on the occasion of ExLing 2024 Paris, the ExLing Society was restructured as the International Linguistic Society. This transformation reflects our vision of hosting multiple conference series throughout the year, ensuring continuous opportunities for scholars to submit their work and participate in Society events.

This volume contains the proceedings of Phonetics 2025 Hong Kong. In line with the conference's scope, the papers address a wide range of topics, including consonants and vowels, prosody and intonation, as well as interdisciplinary and applied aspects of phonetics. We warmly welcome the Hong Kong phonetics community and recognise its growing presence within the international research landscape.

We extend our sincere thanks to all participants of Phonetics 2025 Hong Kong, to our keynote speakers — Janet Fletcher, Niels Schiller, and Yi Xu — and to colleagues from the International Advisory Committee, the Review Committee, and the Organising Committee for their invaluable contributions to the success of this conference.

Antonis Botinis
The Linguistic Society

Contents

Speech rate perception and interlocutor identification in human-directed vs. device-directed speech

Yahya Aldholmi, May Al-Sager, Arwa Alsahafi, Reema Alshiddi
King Saud University, Saudi Arabia

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/001/000661

Abstract

This study investigates how listeners perceive differences between human-directed and device-directed speech, focusing on speech rate and interlocutor identification. Seventy-eight native Arabic speakers (aged 19–22; M = 20.46, SD = 1.11) participated in two tasks: rating the speed of 30 short recordings and determining whether each sample was directed toward a person or a device. The results showed that device-directed speech was consistently perceived as faster, while human-directed speech enabled more accurate interlocutor identification. Statistical analyses confirmed that these differences were significant, with moderate effect sizes. The findings suggest that devices produce speech efficiently but lack the natural variability that characterizes human communication. Incorporating more dynamic and expressive features into voice systems could improve user engagement. Future research should consider cultural differences and emotional tone in shaping speech perception.

Keywords: speech perception, human-directed speech, device-directed speech, speech rate, interlocutor identification

Introduction

Although voice assistants are direct and efficient, users still perceive key differences between device and human speech. Devices often sound faster and more precise, lacking the warmth and adaptability of human voices (Vonessen et al., 2024). Speech rate is a crucial characteristic in such perception; faster speech is efficient but less personal, whereas slower speech improves understanding and engagement (Huiyang & Min, 2022). While device speech rate has been studied in languages such as English (Jones et al., 2007) and Arabic (Aldholmi et al., 2021), the rate of speech directed to devices in Arabic has not. Interlocutor identification, which relies on subtle acoustic variations, is also more reliable with human speech (Zellou et al., 2023). Historically, since humans only spoke to other humans, listener identification was not a critical research topic; the emergence of AI-based systems, however, has opened new avenues for inquiry. Thus, this study explores how Arabic listeners evaluate speech rate in human- versus device-directed speech and their ability to identify the intended

interlocutor (human or device). The findings will offer insights for creating more human-like and engaging voice systems.

Methodology

A within-subjects experimental design was employed. Seventy-eight native Arabic speakers aged 19 to 22 (M = 20.46, SD = 1.11) participated and were evenly divided into two counterbalanced groups, each exposed to a different order of stimulus presentation to control for order effects. Each participant completed a total of 60 trials. In the first task, listeners rated 30 recordings (15 human-directed and 15 device-directed) on a 7-point Likert scale ranging from "very fast" to "very slow." In the second task, the participants listened to another 30 recordings and indicated whether each sample was directed toward a human or a device. The stimuli were standardized to control for potential confounds. Each recording contained 6- to 10-word utterances (M = 7.8 words, SD = 1.2) spanning 13–36 syllables (M = 24.5 syllables, SD = 5.1) and lasting 2.2–6.3 seconds (M = 4.25 seconds, SD = 1.1) and was produced by the same speaker to ensure consistency in voice characteristics. For example, /wɒt ɪz ðə məʊst əˈtræktɪv ˈkʌlə tuː juː/ ("What is the most attractive color to you?"), a sentence in Modern Standard Arabic, was typical of the recordings used. Data were collected electronically under controlled listening conditions using noise-canceling headphones. The Wilcoxon signed-rank test was applied to assess differences between conditions, and effect sizes were computed to evaluate practical significance.

Results

The findings revealed apparent differences in how listeners perceived the two types of speech. Device-directed speech was generally judged to be faster, with an average rating of 4.09 (SD = 1.07) compared to 3.56 (SD = 1.04) for human-directed speech. This difference proved to be statistically significant (Z = –12.40, $p < .001$) and showed a medium effect size (r = .45). These results confirm a significant difference in perceived speed, supporting the first hypothesis—that device-directed speech would be perceived as faster than human-directed speech.

Regarding interlocutor identification, participants were more accurate when recognising human-directed speech. The average accuracy was 0.83 (SD = 0.12), whereas device-directed speech was identified less accurately at 0.78 (SD = 0.15). This difference was significant (Z = –9.23, $p < .001$) and had a medium effect size (r = .39).

These results support the study's hypotheses that listeners perceive device-directed speech as faster and achieve more accurate interlocutor identification with human-directed speech.

As illustrated in Figure 1, the ratings distribution further emphasizes these differences. Participants' judgments of device speech speed clustered tightly

around higher values, with a median near 5.0, indicating strong consensus that it was faster. By contrast, ratings for human speech centered around 4.0 and displayed slightly greater variability, suggesting less uniformity in perception. The boxplot comparison clearly shows that device-directed speech was consistently perceived as faster across participants, strengthening the statistical findings.

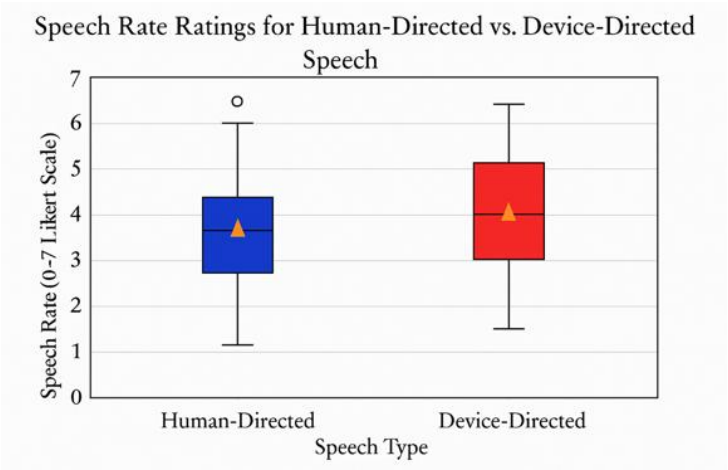


Figure 1. Boxplot of speech rate ratings for Human-Directed vs. Device-Directed Speech.

Discussion

Even though the same interlocutor produced all the recordings, listeners consistently rated device-directed speech as faster than human-directed speech. This finding suggests that people instinctively speak differently depending on the listener. When talking to a device, they tend to speed up and flatten their tone, perhaps aiming for clarity. While this delivery style resembles synthetic voices such as Google TTS (Aldholmi et al., 2021), human speech still carries minor variations that make it sound more natural to human listeners. These small cues matter. In the device-directed samples, participants had greater difficulty determining the intended recipient of the speech; however, they better understood the interlocutor's message since human-directed speech featured more tonal and rhythmic variations. These results remind us that communication requires connection as much as clarity. Voice technology must sound like humans to feel more human, not just to convey information rapidly.

Conclusion

This study illustrated that people's speech patterns naturally vary according to who or what they speak to. The difference was evident even when the same

interlocutor was used. Speech directed at devices was perceived as faster, while speech meant for humans felt more expressive and easier to connect with. That difference had real effects. Listeners could more accurately ascertain whom human-directed speech was meant for, demonstrating that tone and rhythm matter as much as clarity. These results highlight that communication is not just about being understood. Achieving the ideal mix between natural expression and efficiency will be crucial as speech devices advance. In addition, systems should mimic human speech patterns. Future studies can build on these findings by examining how these speech patterns appear in real conversations, across languages, and in emotionally rich situations.

References

Aldholmi, Y., Aldhafyan, R., Alqahtani, A. 2021. Perception of Standard Arabic synthetic speech rate. Interspeech 2021, 1704–1707. https://doi.org/10.21437/Interspeech.2021-39

Huiyang, S., Min, W. 2022. Improving interaction experience through lexical convergence: The prosocial effect of lexical alignment in human-human and human-computer interactions. International Journal of Human-Computer Interaction, 38(1), 28–41. https://doi.org/10.1080/10447318.2021.1921367

Jones, C., Berry, L., Stevens, C. 2007. Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners. Computer Speech & Language, 21(4), 641–651. https://doi.org/10.1016/j.csl.2007.03.001

Vonessen, J., Aoki, N. B., Cohn, M., Zellou, G. 2024. Comparing perception of L1 and L2 English by human listeners and machines: Effect of interlocutor adaptations. Journal of the Acoustical Society of America, 155(5), 3060–3070. https://doi.org/10.1121/10.0025930

Zellou, G., Cohn, M., Pycha, A. 2023. Listener beliefs and perceptual learning: Differences between device and human guises. Language, 99(4), 692–725. https://dx.doi.org/10.1353/lan.2023.a914191

Speech rate of short vs. long interrogative sentences in human-directed vs. device-directed dialectal Arabic speech

Yahya Aldholmi, Arwa Alsahafi, Reema Alshiddi, May Al-Sager
King Saud University, Kingdom of Saudi Arabia

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/002/000662

Abstract

This paper examines the speech pattern adjustments made by female dialectal Arabic speakers when addressing an AI voice assistant compared to a close interlocutor and a stranger across different age groups. Each participant was presented with a set of ten interrogative sentences, controlling for syllabic length (short vs. long). They directed each set to the three addressees, after which the speech rate was measured. The results showed that the overall speech rate averaged 5.41 syllables per second (SpS), with adults speaking slightly faster than teenagers. Speech was fastest with a familiar partner (SpS = 5.54), followed by an unfamiliar one (SpS = 5.48), and slowest with the AI (SpS = 5.21). Linguistic complexity, namely, utterance length, matters, with shorter utterances being articulated more rapidly than longer counterparts. These findings call for future research into additional acoustic features and gender-related differences.

Keywords: device-directed speech; human-directed speech; dialectal Arabic; speech rate; age differences

Introduction

Individuals across languages and dialectal varieties adapt their speaking styles on the basis of several social and contextual factors (e.g., Cohn et al., 2021). Speakers make distinct acoustic–phonetic adjustments for speech directed to machines vs. to humans. Various articulatory features, such as speech rate, pitch, intensity, and duration, have been examined to differentiate between device-directed speech (DDS) and human-directed speech (HDS) (Cohn and Zellou 2021; Cohn et al. 2021; Cohn et al. 2022; Song et al. 2022; Christenson et al. 2023; Cohn et al. 2024a, 2024b).

Studies have amply documented that both speech rate and prosody shape human–AI interactions. Speech rate, articulation clarity, and pitch adjustments are evident and are further influenced by other factors such as context and age, with children exemplifying exaggerated prosody more than adults, likely to guarantee intelligibility (Cohn et al., 2024b). To date, however, research has predominantly considered English varieties (e.g., Cohn et al., 2024a), with far less attention to non-Western varieties. This paper thereby addresses this gap, examining adjustments in speech rate in HDS vs. DDS by native speakers of Najdi Arabic (NA) and whether differences are statistically significant by age

group. Accordingly, we hypothesized that NA speakers will produce slower utterances when directed to an AI voice assistant.

Methodology

The current study tested speech rate under three conditions (ChatGPT–DDS, familiar and unfamiliar human–HDS) and compared two age groups (adolescents and adults). Across these interlocutors, participants produced identical sets of prompts of interrogative utterances.

Participants

A total of $N = 20$ NA participants were recruited through the native Najdi society. Recruitment criteria included dialect nativeness (i.e., including both familiar "experimenter" and unfamiliar interlocutors) and aged between 13–18 (M = 15.5, SD = 1.71) and 30–45 (M = 37.5, SD = 4.61) years old for teenagers and adults, respectively. All the participants were female, had no speech impairments, and were familiar with AI usage.

Procedure

Participants completed the experiment in person with a familiar "experimenter" and remotely with unfamiliar human interlocutors. First, participants were asked to introduce themselves. Then, they asked 10 interrogative sentences of varying length—short (7/8 syllables, M = 7.5, SD = 0.5) and long (14/15 syllables, M = 14.5, SD = 0.5)—per addressee in a counterbalanced order. For each interrogative utterance, speech rate was measured (mean number of syllables per second: SpS) via *Praat* (version 6.4.27), and the extracted data were then analysed in a linear mixed effects model.

Results

The aggregate results revealed an average speech rate of 5.41 SpS. Relative to ChatGPT–DDS, the speech rate was faster in familiar–HDS [$coef = 0.299$, $p < 0.049$]. Although there was no statistically significant effect of age [$coef = -0.248$, $p > 0.269$], there was an effect of sentence length; over the course of the experiment, participants tended to accelerate speech rate when producing short interrogatives [$coef = 0.484$, $p < 0.001$].

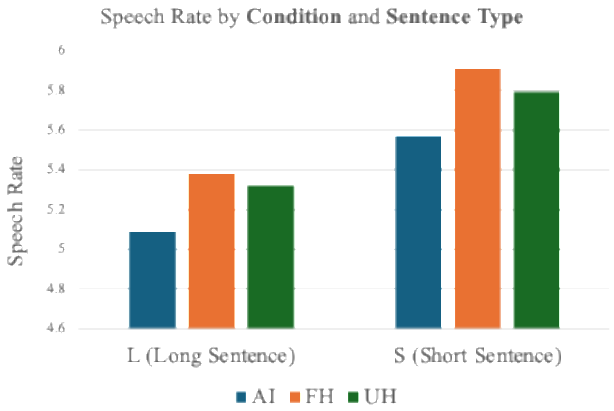


Figure 1. Condition and Sentence Type

Figure 1. Condition and sentence type.

With respect to interaction effects, the model showed that none achieved statistical significance [all $p > 0.05$]. As illustrated in Fig. 1, the fastest speech rate occurred for short sentences (S) under all conditions. The largest speed gap was identified in the familiar human (FH) condition (5.9 vs. 5.4 SpS) and the smallest in AI (5.55 vs. 5.1 SpS), with the unfamiliar human (UH) condition falling in between. Adults outpaced teenagers in sentence length. That is, adults articulated at 5.0 and 5.6 SpS, whereas teenagers articulated at 4.8 and 5.35 SpS for long and short sentences, respectively. However, differences by age were not statistically significant.

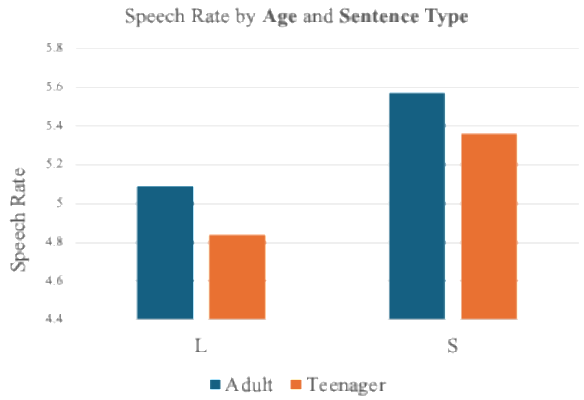


Figure 2. Age and Sentence Type

Figure 2. Age and sentence type.

Discussion and conclusion

This study tested adjustments that emerged when NA speakers interacted with a voice assistant compared to familiar and unfamiliar human interlocutors. There were consistent adaptations for the voice assistant: Speech directed toward the device was slower, which postulates a style shift in comparison to human interlocutors. In related work, a faster speech rate has been observed for an unfamiliar addressee compared to a familiar human (Cohn et al., 2024b). In the current study, however, speakers produced faster speech when talking to a familiar human. In contrast to Cohn et al. (2024a), who identified longer and higher-pitched utterances among children overall compared to adults, this study identified nearly no statistically significant age-based differences within teenagers and adults. This experiment has several limitations that can be addressed in future work. First, the present study examined one variety of dialectal Arabic, Najdi Arabic, but many other related varieties are comparably rare-to-unattested. Second, the current examined acoustic feature was speech rate, aiming to probe the speaking speed within speech directed to devices and NA interlocutors. However, NA exhibits distinct features, such as pitch variation and duration, that speakers might alter when addressing voice assistants.

References

Cohn, M., Zellou, G. 2021. Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments. Frontiers in Communication 6.

Cohn, M., Liang, K.-H., Sarian, M., Zellou, G., Yu, Z. 2021. Speech Rate Adjustments in Conversations with an Amazon Alexa Socialbot. Frontiers in Communication 6.

Cohn, M., Segedin, B. F., Zellou, G. 2022. Acoustic-phonetic properties of Siri- and human-directed speech. Journal of Phonetics 90.

Cohn, M., Mengesha, Z., Lahav, M., Heldreth, C. 2024a. African American English speakers' pitch variation and rate adjustments for imagined technological and human addressees. JASA Express Letters 4.

Cohn, M., Barreda, S., Graf Estes, K., Yu, Z., Zellou, G. 2024b. Children and adults produce distinct technology- and human-directed speech. Scientific Reports 14.

Christenson, B., Ringler, C., Sirianni, N. J. 2023. Speaking fast and slow: How speech rate of digital assistants affects likelihood to use. Journal of Business Research 163.

Song, J. Y., Pycha, A., Culleton, T. 2022. Interactions between voice-activated AI assistants and human speakers and their implications for second-language acquisition. Frontiers in Communication 7.

Bridging the gap: ensuring synthetic phonics continues from kindergarten into primary school in Hong Kong

Geeta Gobindram Bhavnani
The Hong Kong Polytechnic University, Hong Kong

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/003/000663

Abstract

This study investigates the continuity of phonics instruction from kindergarten to primary school in Hong Kong. A mixed-methods approach was employed, comprising a survey of 114 teachers and parents, supplemented by qualitative open-ended responses. Quantitative results indicate a significant perceived disconnect: 46.2% of respondents believe children retain only some phonics knowledge, and 42.9% report phonics is not taught systematically in primary school. Key challenges include a busy primary curriculum (64.8%) and differing teaching approaches (57.1%). Qualitative analysis reveals two primary enablers for continuity: enhanced teacher training and the implementation of a clear, cross-level phonics curriculum. The findings suggest that without structured policy and collaborative planning, the benefits of early phonics instruction are diluted, hindering literacy development in the critical early primary years.

Keywords: phonics continuity, teacher training, curriculum alignment

Introduction

Phonics, a method for teaching reading and spelling by correlating sounds with letters or letter groups, is a cornerstone of early literacy instruction in many English-speaking contexts (Harris & Hodges, 1995). This study focuses specifically on Systematic Synthetic Phonics (SSP), an approach where learners are directly taught the relationship between graphemes (letters) and phonemes (sounds) in a clearly defined, incremental sequence. The efficacy of SSP is well-documented, with studies showing that systematic phonics instruction significantly improves reading outcomes and phonemic awareness compared to non-phonics approaches (Connelly, Johnston & Thompson, 2001; Mann & Wimmer, 2002).

In Hong Kong, the educational landscape is unique, promoting biliteracy and trilingualism. Historically, English instruction relied on "Look and Say" or whole-word methods, which led to lower phonological awareness among learners compared to their peers from alphabetic L1 backgrounds (Holm & Dodd, 1996; Jackson et al., 1999). While phonics has gained popularity in local kindergartens

and learning centres since the 2000s, often as a promotional tool, its adoption is not mandated by the Education Bureau (EDB). This raises a critical question: what happens to children who receive phonics instruction in kindergarten when they transition to primary school? This study aims to bridge this knowledge gap by exploring the perceptions of teachers and parents on the continuity of phonics instruction and identifying the key barriers and enablers in the Hong Kong context.

Methodology

A convergent mixed-methods design was employed to provide a comprehensive understanding of the research problem.

Participants and Procedure

An online survey was distributed via Google Forms to approximately 1,000 former participants of a "Teaching Phonics to Young Learners" course offered at the Hong Kong Polytechnic University with over 16 cohorts of students, from July 14 to August 15 2025. 114 responses were received. The primary inclusion criterion was being a teacher, tutor, or parent of a child who learned phonics in kindergarten and is now in Primary 1-3; 91 respondents (79.8%) met this criterion, and their data forms the core of this analysis.

Results

Quantitative findings: a picture of disconnect

A majority of respondents (85.7%) were actively involved in phonics support. However, perceptions of phonics continuity are concerning. Nearly half (46.2%) reported that children's phonics knowledge erodes after kindergarten, and a combined 63.8% indicated that phonics is either not taught systematically or has effectively ceased in Primary 1-2. The challenges most frequently cited are not isolated but point to deep-seated, systemic issues. The overwhelming concern of a "Busy primary curriculum" (64.8%) highlights a fundamental issue of curricular prioritization, where phonics is seemingly squeezed out by other competing demands. This is compounded by a significant pedagogical misalignment, as over half of the respondents (57.1%) noted stark "Differences in teaching approaches" between kindergarten and primary levels, suggesting that the play-based, multi-sensory methods of kindergarten are not bridged to the more formal, text-focused primary environment. Underpinning these problems is a potential gap in teacher preparedness, with 53.8% identifying that "Primary teachers may not have received focused phonics training." This trifecta of challenges is both reflected in and exacerbated by the finding that over a third of settings (37.3%) use "No set programme," meaning there is no mandated, school-wide systematic synthetic phonics (SSP) scheme such as *Jolly Phonics* or

Letterland. Instead, instruction is likely ad-hoc, dependent on individual teacher initiative or relegated to incidental inclusion within broader textbook units. This underscores a widespread lack of a coherent, structured approach to phonics instruction across this critical educational transition.

Qualitative findings: enablers for continuity

Analysis of the open-ended responses revealed two overarching solutions to the disconnect, richly illustrated by participant suggestions.

Teacher training & teacher practice

The most frequent suggestion (20 responses) centered on the critical need for professional development. Respondents universally emphasized that "Primary teachers need to be trained to use phonics consistently to support their students in reading and spelling," highlighting a perceived skills gap. The core objective of such training, as one respondent noted, should be to ensure "Primary Teachers know where to pick up and start when the young learners begin P1." Participants advocated for "Adaptations of primary teaching materials to align with the phonics approach in kindergarten". The overarching call was for pedagogical shifts, making instruction more interactive, with one respondent succinctly recommending to "make phonics more interactive and fun for kids."

Planning a clear phonics curriculum

The second theme (13 responses) stressed the necessity for systemic, top-down structural change. The predominant view was that a systematic phonics program should "continue to Primary years," ensuring continuity. Respondents explicitly called for "a structured way Phonics is taught and reviewed in Primary school." To operationalize this, they proposed concrete strategies such as including Phonics as a core component of the English syllabus with allocated time and resources, on par with areas like grammar or reading comprehension. Another key proposal was conducting "an initial screening test at the beginning of the year for P1" to identify the diverse skill levels of incoming students. This would allow teachers to effectively move from fundamental knowledge of Letter-Sound Correspondences to blending and segmenting. The ultimate goal, as summarized by one participant, is that "primary school phonics teaching should be systematically designed to allow students to learn step by step," following a clear developmental sequence from simple to complex phonetic patterns, "not just solely based on textbook materials," which often represents an "ad-hoc instruction" approach where phonics is only addressed reactively when it appears in a reading passage or spelling list, rather than being proactively and systematically taught.

Discussion

The findings paint a consistent picture of a significant gap in phonics instruction during the kindergarten-to-primary transition in Hong Kong. The quantitative data confirms a widespread perception that systematic phonics instruction often ceases or becomes fragmented after kindergarten, leading to a regression in children's skills. This echoes historical concerns raised by Holm & Dodd (1996) about the lack of a strong alphabetic foundation in Hong Kong students. The identified challenges—curricular pressure, pedagogical misalignment, and insufficient teacher training—constitute a complex barrier that undermines the investment in early phonics. The qualitative data provides a clear path forward. The strong emphasis on the need for **Teacher Training** suggests that equipping primary teachers with the skills and knowledge to continue phonics instruction is paramount. Simultaneously, the call for a **Clear Phonics Curriculum** highlights the need for systemic, top-down solutions, including cross-level curriculum planning and policy support, to ensure coherence and consistency.

This study has limitations, including a modest sample size and reliance on self-reported perceptions. Future research should involve direct assessment of children's phonics skills and observational studies of classroom practices.

Acknowledgements

The author thanks all the teachers and parents who participated in this survey. Special thanks to Ms Winnie Pang, chief academic officer at pea phonics (info@peaphonics.com), for collecting all the data from the teachers and parents for this paper.

References

Connelly, V., Johnston, R., Thompson, G.B. 2001. The effect of phonics instruction on the reading comprehension of beginning readers. Reading and Writing 14, 423-457.

Harris, T.L., Hodges, R.E. (eds.) 1995. The Literacy Dictionary. International Reading Association.

Holm, A., Dodd, B. 1996. The effect of first written language on the acquisition of English literacy. Cognition 59, 119-147.

Jackson, N.E., Chen, H., Goldsberry, L., Kim, A., Vanderwerff, C. 1999. Effects of variations in reading method on reading achievement. National Institute of Education, Singapore.

Mann, V., Wimmer, H. 2002. Phoneme awareness and pathways into literacy: A comparison of German and American children. Reading and Writing 15, 653-682.

McBride-Chang, C., Treiman, R. 2003. Hong Kong Chinese kindergartners learn to read English analytically. Psychological Science 14(2), 138-143.

Effects of consonant reduction on adjacent vowels in Meru dialects

Conceição Cunha[1], Franziska Muck[1], Fridah Kanana[3]

[1]Ludwig Maximilian University of Munich, Germany
[2]Kenyatta University, Kenya

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/004/000664

Abstract

This paper investigates the influence of fricative reduction on adjacent vowels in three Meru dialects. The voiced fricative /β/ is retained in Chuka but deleted in Imenti, creating vowel sequences (hiatus). Since Bantu languages generally avoid hiatus, three repair strategies have been described for these vowel sequences in Imenti: (1) merger of two vowels with the same quality, (2) shortening of the prefix vowel, and (3) lengthening of the stem vowel to compensate for fricative deletion. Audio data were recorded from 75 speakers across three dialects: Imenti, Chuka, and Tiania, an unstudied dialect. The results show consistent /β/ deletion in Imenti and Tiania, as well as some evidence of compensatory lengthening of the merged vowels in both dialects compared to Chuka.

Keywords: Bantu languages, dialects, sound change, deletion, compensatory lengthening

Introduction

The Ameru people are a Bantu-speaking group in Kenya who reside on the northeastern slopes of Mount Kenya (Fadiman, 1973; Guthrie 1967-71). The dialects within the Meru group belong to the larger genealogical Bantu language family, which traces back to Proto-Bantu. Linguistically, the Ameru are composed of nine sub-ethnic groups who speak closely related and mutually intelligible dialects (Kanana 2011a, 2011b, Cunha et al., 2023). Speakers of these varieties are aware of the subtle differences among the dialects they speak but identify as belonging to a related group.

In a morpho-phonological and lexical analysis, Kanana (2011a, 2011b, 2015) classified these dialects as lying on opposite ends of a spectrum of innovativeness, with Chuka being more conservative and Imenti showing greater morpho-phonological change in its consonantal system. Among other differences, the Meru dialects Chuka and Imenti diverge in the presence or absence of the bilabial fricative /β/ (Kanana, 2015). Since Bantu languages tend to avoid hiatus, the following repair strategies has been suggested and will be investigated:

1. Vowel merger when the prefix vowel is identical to the first stem vowel, or when deletion occurs stem-medially and both stem vowels merge, e.g. /ɣo.taβa, ɣo.taː/, *to draw water* (Kanana, 2015, p. 156).

2. Shortening of the prefix vowel (V1) and lengthening of the stem vowel / koβanda, kwaːnda, *to plant* (Kanana, 2015, p. 157). In a few cases, the vowel does not lengthen.
3. Lengthening of the first stem vowel (V1), e.g. /ɣotuβa, ɣotuːa/ (Kanana, 2015, p. 156).

The third dialect, Tiania, is included in the analysis to examine the position of Tiania within the continuum.

Methods

The recordings were carried out in the villages of residence, where two interviews were recorded simultaneously in two separate rooms, one interview being conducted by a native speaker of Imenti and the other by a team of a native speaker of the local dialect and a trained phonetician. The participants were tasked with producing local dialectal equivalents of words presented on a computer monitor in both English and Swahili. 96 words were presented randomized and repeated two to three times. The speech was recorded with a Beyerdynamic TG H54c head-mounted microphone at 44.1 kHz onto a Tascam US-2x2 interface connected to a laptop using SpeechRecorder 3.12.0 (Draxler and Jänsch, 2004).

Audio data and statistical analysis

The speech signals were forced-aligned using the WebMAUS (Jochim et al., 2017) application for German, as it yielded the best results. These results were corrected manually by trained phoneticians and student assistants using Praat (Boersma and Weenink, 2025). Making use of the emuR-package (Jochim et al., 2023) in R, the segment lists for the segments of interest were extracted from a Emu-SDMS-database (Winkelmann et al., 20179. Extracted data included words together with their /β/ - adjacent vowels in Chuka and the vowels in hiatus after the deletion in Imenti and Tiania. The final corpus consisted of tokens of the 13 target lexemes:

merger: /ɣota(β)a/, /ko(β)ɔra/, /ko(β)ɔria/

prefix V1: /ro(β)ɛni/, /ko(β)ɔria/, /ko(β)anda/, /ŋko(β)anda/, /ko(β)iða/, /ko(β)iŋga/

stem V1: /ɣotu(β)a/, /ɣɔku(β)e/, /ɛŋ|ŋku(β)e/, /kaβɛβɔ|kapiɔ/, /mβɛβɔ|mpiɔ/.

initial deletion: /ro(β)ɛni/, /ko(β)ɔria/, /ko(β)anda/, /ŋko(β)anda/, /ko(β)iða/, /ko(β)iŋga/

Mixed effects linear regression models were fitted for this analysis. They allow to factor out possible variation introduced into the data through random factors (Winter, 2020). These random factors were speakers and lexemes. The models were carried out on the duration of vowels as the dependent variable.

Results

The bilabial voiced fricative /β/ was completely deleted in the analysed contexts in Imenti and Tiania, but not in Chuka, as previously reported (Kanana 2011a, 2011b, 2015). Because all conditions predict lengthening of the stem vowels to compensate for the fricative deletion, Figure 1 shows the $\log_{10}$-transformed duration of the two vowels, normalized by word duration, so that both V1 and V2 can be compared across all analysed lexemes in the three dialects. The general prediction is that this interval should be greater in Imenti than in Chuka.

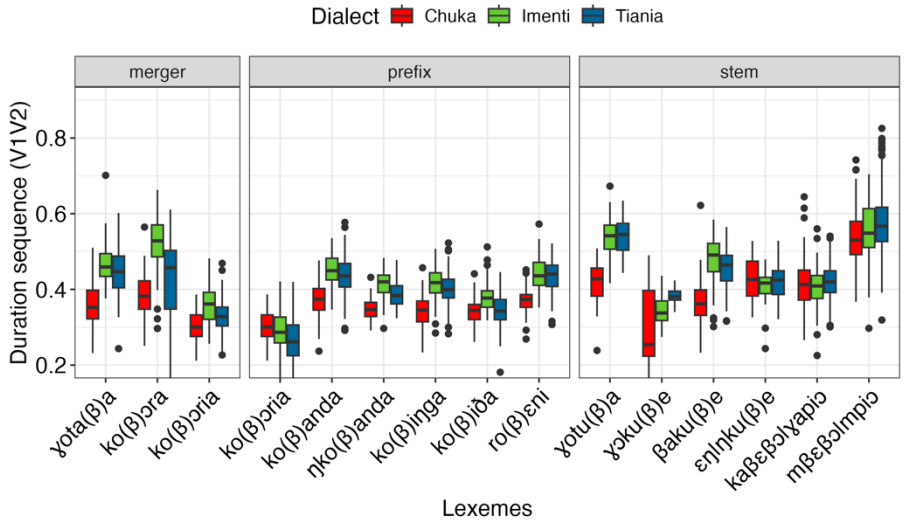


Figure 1. Duration of the sequence between beginning of V1 and end of V2.

Overall, the vowel sequences following the deletion of /β/ are longer in the merger and prefix conditions for Imenti/Tiania than the combined duration of V1 and V2 in Chuka. The form /ko(β)ɔria/ showed two possible realisations: a single long vowel (merger) or a glide followed by an open vowel (prefix V1). In the stem-V1 condition, only the sequences in /ɣotuβa ~ ɣotuːa/ and /βakuβe ~ βakuːe/ were longer, whereas the remaining words did not show this pattern.

The statistical model revealed a significant effect of dialect ($F[2,16.80] = 13.51$, $p < 0.001$) and of condition ($F[2,24.30] = 61.25$, $p < 0.001$), as well as a significant interaction between the two factors ($F[4,33.50] = 94.79$, $p < 0.001$). Pairwise post hoc tests with Bonferroni correction indicated that the duration difference was significant between Chuka and Imenti ($p < 0.001$) and between Chuka and Tiania ($p < 0.001$), but not between Imenti and Tiania.

Discussion

This paper investigates the effects of consonant retention versus deletion in three related dialects. Tiania patterns with Imenti, exhibiting more innovation than Chuka. For reasons of space, only one analysis is presented here. In examining the duration of the interval containing V1 and V2, some evidence for compensatory lengthening was found. The mixed results for stem V1 may have morphological explanations (verb vs. noun), which warrant further study.

Acknowledgements

The research was funded by the project SoundAct which has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101053194).

References

Cunha, C, Kanana, F.E., Harrington, J. 2023. Variation and palatalisation in the production of the plural prefixes in Meru: A study of three dialects, in Radek Skarnitzl, R. and Violín, J. (ed.), Proceedings of the 20th International Congress of Phonetic Sciences: Guarant International, 3397–3401.

Draxler C., Jänsch, K. 2004. SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software," Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04. [Online]. Available: https://aclanthology.org/L04-1125/

Fadiman, J.A. 1973. Early History of the Meru of Mt Kenya, The Journal of African History 14: 1, 9–27, doi: 10.1017/S0021853700012147.

Guthrie, M. 1967-71. Comparative Bantu: an introduction to the comparative linguistics and pre-history of the Bantu languages. Farnborough, Gregg.

Jochim, M., Winkelmann, R., Jaensch, K. Cassidy, S., Harrington, J. 2023. emuR: Main Package of the EMU Speech Database Management System, 2023. [Online]. Available: https://cran.r-project.org/package=emuR

Kanana, F.E. 2011a. Dialect Convergence and Divergence: A Case of Chuka and Imenti, in Selected proceedings of the 40. Annual conference on African linguistics: African languages and linguistics today, 190–205.

Kanana, F.E. 2011b. Meru Dialects: The Linguistic Evidence, NJAS 20: 4, 28

Kanana, F. E. 2015. Lexico-Phonological Comparative Analysis of Selected Dialects of the Meru-Tharaka Group, [Online]. Available: https:/www.peterlang.com/document/1044341.

Kisler, T, Reichel, U., Schiel, F. 2017. Multilingual processing of speech via web services, Computer Speech & Language 45, 326–347. doi:10.1016/j.csl.2017.01.005.

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

Winter, B. 2020. Statistics for linguists: An introduction using R. New York, London: Routledge Taylor & Francis Group.

Temporal dynamics of acoustic emotion encoding

Yuxin Fan[1], Yufeng Wu[2]
[1]Southeast University, Chine
[2]City University of Hong Kong, Hong Kong

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/005/000665

Abstract

Static analyses of speech emotion often overlook temporal dependencies. This study examines how the Valence, Arousal, and Dominance (VAD) of a preceding utterance moderate the relationship between acoustic features and the VAD of the preceding utterance. This study fitted linear mixed-effects models to 5,221 utterances from the IEMOCAP corpus. Results showed that the lagged VAD was the strongest predictor among all dimensions, demonstrating significant emotional inertia. Furthermore, the association between acoustic parameters and subsequent VAD was significantly moderated by lagged VAD. These findings confirm that acoustic-emotion associations are dynamic and context-dependent, challenging static models and highlighting the need to incorporate temporal dynamics in emotion recognition systems.

Keywords: speech emotion recognition (SER), affective computing, acoustic features

Introduction

The voice carries a wealth of information that extends beyond linguistic content to convey details of emotions. As a crucial component of interpersonal communication, this paralinguistic channel holds huge research value. One of the most influential patterns for modelling emotion is the Valence-Arousal-Dominance (VAD) model ((Fontaine et al., 2007)), which insists that emotions can be divided into three fundamental dimensions. Valence (V) describes the direction of an emotion, indicating whether it is positive or negative. Arousal (A) describes the intensity of an emotion, referring to the level of physiological and psychological activation. Dominance (D) is a distinct dimension that describes the sense of control experienced during the emotion.

Existing researches has confirmed that acoustic features—such as F0, intensity and spectral slope—can predict VAD scores. However, these analyses share a critical limitation: they are static. While static analysis can capture the relationship between emotions and acoustic parameters at a specific point in time, it fails to capture the dynamics of emotion as it evolves over time, ignoring the contributions of context and temporal sequencing. An individual's emotional is strongly influenced by the preceding conversational content and emotional states. This study shows how the VAD score of a preceding utterance influences the dynamic relationship between acoustic parameters (prosodic, spectral, and

voice quality measures; e.g., F0, intensity, spectral slope, HNR) and the VAD of the subsequent utterance in spoken English dialogues.

Methodology

This research used the speech corpus from The Interactive Emotional Dyadic Motion Capture (IEMOCAP). (Busso et al., 2008) IEMOCAP is a widely-used database in affective computing that features dyadic interactions between actors. This study focused on the script sessions to extract clear emotional dynamics.

Based on the GeMAPS feature set (Eyben et al., 2016), this study extracted a total of 26 acoustic parameters across five dimensions: F0, intensity, duration, voice quality parameters (spectral balance), and voice quality parameters (variability). The features were extracted using the parselmouth library (Jadoul et al., 2018) in Python. The VAD scores were sourced from the IEMOCAP database, provided by expert annotators.

Analytical framework

Data processing

The raw dataset consisted of 5255 sentences in total. An initial correlation analysis revealed high multicollinearity among the 26 extracted acoustic parameters. To reduce complexity and improve interpretability, this research used Principal Component Analysis (PCA) (Schuller, 2012).

All acoustic parameters were z-scored and adjusted for speaker effects before PCA. The first 10 components (explaining >80% of variance) were retained and descriptively labelled by their strongest loadings (>.5).

Statistical modelling

To account for these data dependencies, this research used a linear mixed-effects (LME) model approach. The LME framework was chosen for its ability to simultaneously address two sources of non-independence: the clustered nature of the data (multiple sentences per speaker) and temporal dependency.

The model's fixed effect structure included the main effects of the acoustic factors, the main effect of the lagged VAD predictor, and their crucial two-way interactions. The random effect structure was specified to account for by-speaker variation in both baseline VAD levels and sensitivity to the temporal dependency effect.

Results

The LME results revealed highly dynamic associations between emotion and acoustics. First, the VAD of the preceding utterance (lagged VAD) was the

strongest predictor across all dimensions, demonstrating strong emotional continuity.

Crucially, the models confirmed that lagged VAD significantly moderated the associations between subsequent acoustic features and VAD.
Acoustic features themselves remained important predictors even under this strong emotional inertia. After controlling for lagged VAD, multiple acoustic factors were still significantly associated with the three VAD dimensions.

Taken together, our results reveal that the association between acoustic features and VAD is highly contingent on the prior emotional state. For instance, the prior state was found to strengthen, weaken, or in some cases, reverse the direction of a feature's influence. This highlights the complex, dynamic nature of emotional encoding. The following figure 1. provides a clear example of this moderating effect.

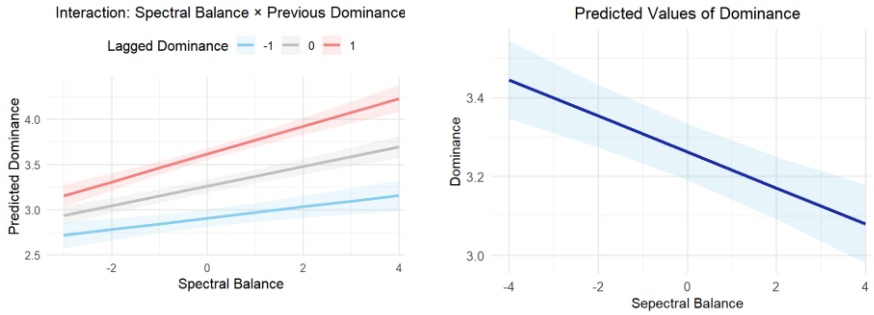


Figure 1. Example of moderating effect.

Discussion

Overall, the results confirm that the perception of emotion in an individual utterance depends not only on its immediate acoustic cues but is also significantly shaped by an "emotional inertia" from the preceding utterance. These findings highlight the limitations of context-independent models and demonstrate that incorporating emotional inertia is essential for future speech emotion recognition systems to capture the continuous and dynamic nature of emotion in dialogue with greater accuracy and human-likeness.

References

Fontaine, J.R.J., Scherer, K.R., Roesch, E.B., Ellsworth, P.C. 2007. The world of emotions is not two-dimensional. Psychological Science 18, 1050-1057.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation 42, 335–359.

Martijn, G. Klaus, S. 2010.Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. Journal of the Acoustical Society of America, 128(3), 1322–1336.

Schuller, B.W. 2012. The computational paralinguistics challenge. IEEE Signal Processing Magazine 29, 97-101.

Eyben, F., et al. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing 7, 190-202.

Jadoul, Y., Thompson, B., de Boer, B. 2018. Introducing Parselmouth: a Python interface to Praat. Journal of Phonetics 71, 1-15.

Cross-linguistic influence on mid vowels of late Salento Italian-French bilinguals

Marie Fongaro[1], Barbara Gili Fivela[2], Maud Pélissier[3]
[1]University of South Bohemia in České Budějovice., Czech Republic
[2]University of Salento, Italy
[3]Université Paris Cité, France

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/006/000666

Abstract

This study focuses on open and close mid vowels in native language (L1), here the Italian variety spoken in Salento, and second language (L2), here French, of 15 Italians from Salento living in the Paris region (hereinafter bilinguals - B). We investigated the phonetic L1-L2 and L2-L1 influence in their L1 and L2 speech production, by comparing it with the L1 speech production of 15 Italian and 15 French control speakers. Acoustic analysis of formant values shows (1) no L2-L1 influence, and (2) an L1 influence on the B L2 /ɛ/, /e/, /o/, and /ɔ/ as for F1 and/or F2.

Keywords: mid vowels, cross-linguistic influence, Italian, French

Introduction

Today, many people live abroad, i.e., in a non-native language (L1) country, where they daily use a second language (L2). Those who acquired the L2 after the age of six are for Grosjean (2013) "late bilinguals". Authors have studied the mutual influence of languages known by late bilinguals (i.e., Cross-Linguistic Influence – CLI, see Jarvis & Pavlenko 2008) in their speech at phonetic level. However, they over-focused on late bilinguals whose L1 or L2 was English and rarely examined so-called "bidirectional CLI", i.e., the influence of L1 on L2 and that of L2 on L1 in the same group of late bilinguals.

This work is one of the few examining bidirectional CLI in a group of late bilinguals for which neither L1 nor L2 is English, specifically Salento Italian-French Bilinguals (hereinafter referred to as B). The B were born and/or grew up in Salento (southern Italy), moved to France as adults, living there for a variable amount of time (here Length Of Residence - LOR), and resided in the Paris region at the time of study. Their L1, i.e. the Italian variety spoken in Salento (hereinafter Salento Italian), and their L2, French, differ in terms of mid vowels: There is no phonological distinction between /ɛ/ and /e/, and /ɔ/ and /o/ in Salento Italian, but there is in French as it is commonly spoken in the Paris region (cf. Grimaldi & Calabrese 2018, Munot 2002). This difference between languages is core when studying CLI, as according to the Speech Learning Model and its revised version (Flege & Bohn 2021; SLM and SLM-r), the L1 and L2 sound

systems of bilinguals coexist in a shared phonetic space, where they interact and can influence each other. Namely, bilinguals may incorrectly classify an L2 sound into their L1 category leading to its inaccurate pronunciation, which can be improved with their increasing L2 experience as they create a new category for this sound. The classification of L2 sound into L1 categories and the creation of a new L2 phonetic category vary between individuals, as it depends on many factors related to the specific speaker.

Goals and hypotheses

This study aims to investigate CLI in open and close mid vowels of B, i.e. Salento Italian-French Bilinguals, by comparing their L1 and L2 speech production with L1 speech production of Italian (hereinafter I) and French (hereinafter F) control speakers. We hypothesise that:

1. The B will produce French /e/, /ɛ/, /o/ and /ɔ/ inaccurately, i.e., not as F controls, since their L1 front mid vowel will influence their L2 /e/ and /ɛ/, just as their L1 back mid vowel will influence their L2 /o/ and /ɔ/.
2. The B will differ in production of Salento Italian mid vowels from I controls, because their L2 front mid vowels will influence their L1 front mid vowel, just as their L2 back mid vowels will influence their L1 back mid vowel.
3. The higher a B speaker's *LOR*, the more accurate his/her pronunciation of French mid vowels will be, and the less his/her pronunciation of Salento Italian mid vowels will resemble that of I controls.

Method

We recorded the L1 and L2 speech of 15 B (9M, 6F; mean age = 41.13 y.o.; SD = 10.39, *LOR* [year]: 1, 4, 6, 8, 8, 10, 12, 14, 14, 15, 18, 22, 24, 27, 33) and the L1 speech of 15 I and 15 F controls, matched as much as possible for age, sex and education level. I and F controls were born and/or grew up in Salento and in the Paris region, respectively, and have lived there for all or most of their lives. Since L2-L1 influence occurs more in spontaneous than in reading-elicited speech (Hevrova 2021), we elicited speech as spontaneous as possible in a controlled setting, using a set of pictures corresponding to Italian and French words with target vowels (/ɛ/, /e/, /ɔ/, /o/) placed in stress-controlled positions. First, a picture of a word X was shown on the PC screen and the speaker produced the carrier sentence 'I say X'. Second, more pictures were shown on the PC screen and the speaker described the way the experimenter was moving them by using the carrier sentence 'I put X next to Y. I moved X'. Orthographically transcribed recordings were automatically labelled and segmented, and manually corrected in PRAAT (Boersma & Weenink 2025), where F1 and F2 of the target vowel were automatically measured from the middle third of each target. A total of 3928 vowels in 3928 words were analysed. Statistical analyses were performed in R

using various packages. Hypothesis 1 and 2 were tested by (1) normalising F1 and F2 values by the Lobanov (1971) method, and (2) building a set of linear mixed-effects models (LME), one for each vowel and each formant. As fixed effects, we entered *group* of speakers (B, I, F) and *language* (French *vs* Italian) with an interaction term into models. As random effects, we used intercepts for speaker and word. *p*-value were obtained by performing pairwise post-hoc tests with Tukey method of *p*-value adjustment for comparing a family of 4 estimates. Hypothesis 3 was tested by (1) computing average F1 and F2 values per speaker, vowel and language for the B, and (2) building a set of linear regressions, one for each formant of each vowel of each language, with *LOR* as independent variable.

Results

Results reveals that (1) B significantly differ from F controls in their L2 production as for the F1 of /ɛ/ (β=-0.18, *SE*=0.07, *t*=-2.80, *p*=0.037), /e/ (β=0.19, *SE*=0.06, *t*=3.21, *p*=0.012), /ɔ/ (β=-0.18, *SE*=0.07, *t*=-2.69, *p*=0.048) and /o/ (β=0.35, *SE*=0.05, *t*=6.58, *p*<.0001), and F2 of /e/ (β=-0.18, *SE*=0.05, *t*=-3.63, *p*=0.004), /ɔ/ (β=-0.37, *SE*=0.05, *t*=-7.79, *p*<.0001) and /o/ (β=0.12, *SE*=0.04, *t*=2.80, *p*=0.037), (2) B do not differ in their L1 production from I controls as for F1 and F2 of any studied vowel (see Figure 1), (3) LOR does not significantly predict F1 or F2 of any of the B L1 and L2 mid vowels.

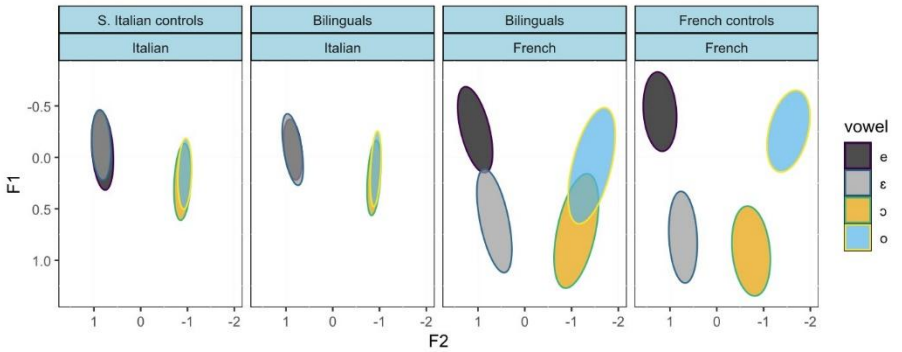


Figure 1. Normalised F1 and F2 values of Salento Italian and French mid vowels produced by the B, I and F controls. The ellipses represent 68% of data.

Discussion and conclusions

This study explored (1) the bidirectional CLI on the B L1 and L2 open and close mid vowels, by comparing the B production with that of I and F controls, and (2) whether *LOR* predicts these CLI. Hypothesis 1 concerned the influence of the B L1 front mid vowel on their L2 /e/ and /ɛ/, and the influence of their back mid vowel on their L2 /o/ and /ɔ/ leading to their inaccurate production

of French open and close mid vowels. This hypothesis was almost entirely confirmed as results show that the B differ from F controls in the F1 and F2 of all their L2 studied vowels, except for the F2 of /ɛ/. This result, in line with the SLM supposition that L2 learners attempt to classify an L2 sound into an L1 category wherever possible, is interpreted here as that many B have assimilated French /ɛ/ and /e/ into their L1 category of Salento Italian front mid vowel, and French /ɔ/ and /o/ into their L1 category of Salento Italian back mid vowel, with a consequent inaccurate production of these L2 sounds. Hypothesis 2, concerning an L2 influence on the B L1 mid vowels, was not confirmed as no significant difference was found between vowels produced by B and I controls, meaning that there is no L2 influence on the B L1 mid vowels. Because the B are heterogeneous speakers in terms of *LOR*, L1 and L2 use, L2 proficiency, age of arrival in France and L2 acquisition onset, we assume that L2-L1 influence might be found if only speakers with very strong L2 experience would be included among the B group. We also hypothesised that the *LOR* would predict the CLI in the B L1 and L2 speech, which was not confirmed. To conclude, as a group, Salento Italians living in the Paris region do not differ in their L1 mid vowels from I controls, but they differ in their L2 open and close mid vowels from F controls, indicating L1-L2 influence. The inter-speaker variability of the B has to be analysed in a follow up study as many other factors potentially affect the speaker accuracy in L2.

Acknowledgements

This study was supported by the MŠMT of the Czech Republic (Programme OP JAC), no. CZ.02.01.01./00/22_010/0008126. Co-founded by the EU.

References

Boersma, P., Weenink, D. 2025. Praat: doing phonetics by computer. Computer program. Version 6.4.27.

Flege, J. E., Bohn, O.-S., The revised speech learning model. In Wayland, R. (ed) 2021, Second Language Speech Learning. 3–83, Cambridge, Cambridge University.

Grimaldi, M., Calabrese, A. Metaphony in Southern Salento. In D'Alessandro, R., Pescarini, D. (eds.) 2018, Advances in Italian Dialectology. 253–291, BRILL.

Grosjean, F. 2013. Bilingualism: A Short Introduction. Hoboken, Wiley, 5–25.

Hevrova, M. 2021. Phonetic Attrition and Cross-Linguistic Influence in L1 Speech of Late Czech-French Bilinguals. PhD thesis.

Jarvis, S., Pavlenko, A. 2008. Crosslinguistic Influence in Language and Cognition. New York, Routledge.

Munot, P., Nève F.-X. 2002. Une introduction à la phonétique. Liège, CEFAL.

Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. The Journal of the Acoustical Society of America, vol. 49, no. 2B, 606–608.

Downtrend in Sylheti phrasal tones

Tulika Gogoi, Amalesh Gope
Tezpur University, India

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/007/000667

Abstract

This study examines the phonological nature of the fundamental frequency (f0) downtrend in Sylheti, an Indo-Aryan tonal language exhibiting both lexical and phrasal tones. Speech data from five native speakers were analysed using Praat, ProsodyPro, and statistical modelling in R and Python to investigate the behavior of phrasal tones within Accentual Phrases (APs) across Intonational Phrases (IPs). The results reveal a consistent stepwise lowering of f0 peaks, independent of sentence length, indicating a phonological rather than purely phonetic process. Mathematical modelling based on Liberman and Pierrehumbert's (1984) downstep and final-lowering equations accurately predicted observed f0 patterns ($R^2 = 0.98$). These findings confirm that Sylheti exhibits a systematic, phonologically governed downtrend across utterances.

Keywords: sylheti, downtrend, f0 modelling, phrasal tones, intonation

Introduction

The gradual downward movement of fundamental frequency (f0) or pitch during the production of utterances, known as 'downtrends,' is well documented across languages and can be phonetic or phonological in nature (Connell, 2001; Gussenhoven, 2004; Gogoi et al, 2024). This study explores the f0 downtrend in Sylheti, an Indo-Aryan tonal language exhibiting both lexical and phrasal tones (Gope, 2016, 2018, 2021, 2025; Gope & Mahanta, 2014–2016; Gogoi, 2024; Gogoi & Gope, 2023; Mahanta & Gope, 2018). The focus is on whether Sylheti's f0 downtrend reflects a phonological process through how phrasal tones marking Accentual Phrases (APs) behave across Intonational Phrases (IPs). Two central questions guide this research: (i) how the overarching f0 downtrend across IPs shapes the surface realization of phrasal tones, and (ii) whether the peak-by-peak descent can be systematically modelled.

Experimental procedure

Five native Sylheti speakers (three males, two females; aged 18-33) from Dharamnagar, Tripura, recorded 11 scripted neutral declarative sentences (8-11 syllables) five times each to analyze f0 downtrend patterns. Sentences incorporated varied tonal sequences to assess lexical tone effects on f0. Recordings (44.1 kHz, 32-bit) were manually annotated at the syllable level using Praat (Boersma & Weenink 2012), and syllable-wise mean and time-normalized

f0 values were extracted with ProsodyPro (Xu 2013). Pitch contours were visually inspected. Statistical analysis involved one-way repeated measures ANOVA (R 4.2.3) for f0 differences, and downward slopes were modelled using Origin 8.1 and Python's linregress function, with model accuracy evaluated via R^2 values from sklearn.metrics.

Results and discussion

Downstep in Sylheti manifests as a gradual lowering of successive f0 peaks. Neutral declaratives are structured into Accentual Phrases (APs) marked by L* pitch accents and Ha boundary tones, with downstep examined through Ha tone peaks. Speakers typically divide Intonational Phrases (IPs) into two or three APs, forming recursive structures that group lexical and functional items (Gogoi 2024).

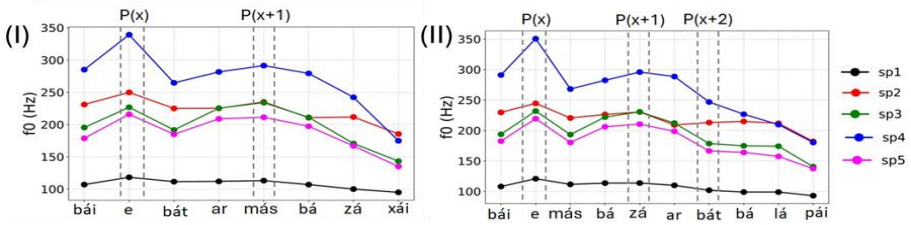


Figure 1. The average f0 contours for five speakers (sp1-5) across different sentences: (I) [bái-e bát ar más bázá xái] '(my) brother eats rice and fried fish,' and (II) [bái-e más bázá ar bát bá lá pái], '(my) brother likes fish fry and rice.'

Visual inspection (Figure 1) reveals that female speakers have higher peaks and wider pitch ranges than male speakers, but all maintain consistent contour shapes. Sentences feature two to three peaks, with P(x) highest and subsequent peaks (P(x+1), P(x+2)) progressively lower. Underlying lexical tones affect only the scaling of L* pitch accents and do not impact peak positions. Scaling changes from IP medial H tone roots adjust pitch height by interpolating L and H tones, but do not alter phrasal tone specification.

Figure 2 illustrates the average f0 contours for selected sentences, highlighting key peaks P(1), P(2), and P(3) with stars denoting the highest tonal points. Linear fits applied to final f0 points using the equation $y = m*x + c$ (where m is the slope indicating the f0 change per syllable and c is the intercept) model overall pitch movement across sentences. The negative slopes confirm a consistent downtrend, with f0 gradually declining toward the end of the sentence.

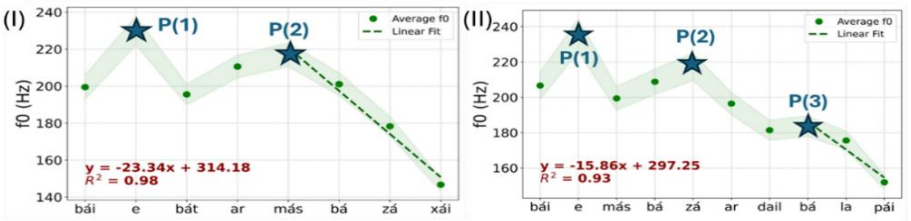


Figure 2: The average f0 contours for two different sentences (I) [bái-e bát ar más bázá xái] '(my) brother eats rice and fried fish,' and (II) [bái e máś bá zá ar dáil bá la pái] '(my) brother likes fish fry and lentils.'

Phonological nature of the downstep of f0 peaks in Sylheti

The stepwise lowering of f0 peaks reflects a phonological pattern rather than a phonetic effect. The consistent and predictable f0 decline across sentences of varying lengths and syllable counts (Figures 1-2) supports this interpretation (Liberman & Pierrehumbert, 1984), with initial and final f0 values and the rise to the first peak remaining similar across utterances. This stability indicates the downtrend in peak height is a phonological property, not an articulation artifact. Following Liberman and Pierrehumbert's (1984) exponential model, the downstep ratio between consecutive peaks was calculated as r = (P(x+1) – R) / (P(x) – R), where R = (Mean f0_last peak + Mean f0_minimum)/2. Each successive peak was predicted as P(x+1) _predicted = R + [r × (P(x) – R)]. For two-peak utterances, the average downstep ratio was r = 0.69 ± 0.09, showing stability independent of intervening syllables. For three-peak utterances, the ratio between the third and second peaks dropped to r = 0.27 ± 0.17, likely due to final lowering, a pattern reported also in American English and Mexican Spanish (Liberman & Pierrehumbert, 1984; Prieto et al., 1996).

To model this, the final lowering constant 1 (Liberman and Pierrehumbert 1984) was applied using P = R + l * (P(down) - R), where *P* is the height of the last peak, P(down) is the peak height predicted by the downstep rule, R is the reference line, and l is the final lowering constant. The ratio determines the value of l, computed as l = (P(obs) - R) / (P(down) - R), where P(obs) is the observed peak height. Once known, l predicts the next peak height using the formula: P(x+2) _predicted = R + l* (P(x+1) - R), where the P(x+2) _predicted is the predicted f0 value. A comparison of predicted and observed f0 values shows that the final lowering model achieves an R^2 of 0.98, surpassing the basic downstep model's R^2 of 0.75. These confirm that Sylheti's downstepped f0 peaks are mathematically modelled and phonological in nature. Moreover, in sentences with more than two peaks, final lowering significantly influences the last peak, as effectively described by Liberman and Pierrehumbert's lowering constant.

Conclusion

The analysis confirms that Sylheti exhibits a consistent, predictable f0 downtrend across utterances, indicative of a phonological rather than a phonetic process. Mathematical modeling accurately captures both the stepwise and final lowering of f0 peaks. These findings deepen our understanding of Sylheti prosody and provide a valuable foundation for comparative research on tonal organization in South Asian languages.

References

Boersma. P., Weenik D. 2012. Praat: doing phonetics by computer (Version 5.3.04_win64) [Computer program].

Connell, B. 2001. Downdrift, Downstep, and Declination. Typology of African Prosodic Systems Workshop, Bielefeld University, Germany.

Gogoi, T., Gope, A. 2023. The Phonetics of Prosodic Marking of Focus in Sylheti. Radek Skarnitzl, Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences. Guarant International, Prague, pp. 1638-1642.

Gogoi, T., Tetseo, S., Gope, A. 2024. A Mathematical and Statistical Approach to Explore the Downtrend Properties in Chokri. Forum for Linguistic Studies (FLS). 6(5), 664–677.

Gogoi, T. 2024. The Phonetics and Phonology of Tone-Intonation Interaction with reference to Sylheti and Chokri. Ph.D. Dissertation. Tezpur University, India.

Gope, A. 2016. The Phonetics and Phonology of Sylheti Tonogenesis. Ph.D. Dissertation. Indian Institute of Technology, Guwahati, India.

Gope, A. 2018. The Phoneme Inventory of Sylheti: Acoustic Evidences. Journal of Advanced Linguistic Studies 7: 7–37. Bahri Publications.

Gope, A. 2021. The Phonetics of Tone and Voice Quality Interactions in Sylheti. Languages 6: 154.

Gope, A. 2025. Image-Based Texture Analysis of Vowel Spectrograms in Sylheti using Random Forest Classifier. Procedia Computer Science, Vol 260, pp. 399-405.

Gope, A., Mahanta, S. 2016a. Perception of Lexical Tones in Sylheti. In DiCanio, C. et. al (Eds). Proceedings of the TAL-2016. Buffalo, NY.

Gope, A., Mahanta, S. 2015. An Acoustic Analysis of Sylheti Phonemes. In The Scottish Consortium for ICPhS 2015 (Eds). Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK.

Gope, A., Mahanta, S. 2014. Lexical Tone in Sylheti. In C. Gussenhoven, Y. Chen, & D. Dediu (Eds). The 4th International Symposium on Tonal Aspects of Languages. Nijmegen, The Netherlands, (pp. 10-14).

Gussenhoven, C. 2004. The Phonology of Tone and Intonation. Cambridge: Cambridge University Press.

Liberman, M., Pierrehumbert J. 1984. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle (Eds). Language Sound Structure, pages 157–233. MIT Press, Cambridge, Massachusetts.

Mahanta, S., Gope, A. 2018. Tonal Polarity in Sylheti in the Context of Noun Faithfulness. Language Sciences. 69, 80-97.

Xu, Yi. 2013. ProsodyPro - A tool for large-scale systematic prosody analysis. Proceedings of the TRASP Conference. 7-10.

The effect of speaker L2 English accent on hiring decisions in China

Yuqing He, Ksenia Gnevsheva
Australian National University, Australia

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/008/000668

Abstract

The present paper investigates the perceived employability of L2 speakers of English in China. In a perception experiment, 90 China-based L1 Mandarin-speaking participants assessed the employability of speakers represented by audio stimuli on five 5-point Likert scales. Two men and two women were included for each of the three L2 English accents (German, Korean, and Mandarin Chinese). Results revealed that speakers with the German accent were rated higher than speakers with the Korean accent. However, speakers with the Mandarin accent were not rated differently from the other L2 accent groups, supporting the prejudice explanation over the comprehensibility one.

Keywords: accents, employment, linguistics discrimination, language attitudes

Introduction

Accent is a unique way of sound production that reflects where a speaker comes from geographically and socially. It has been demonstrated to serve as a source of stereotyping in employment, with people who speak with a second language (L2) accent, in particular, often rated to be less employable than first language (L1) speakers (e.g. Hosoda et al. 2012). Meanwhile, different L2 accents are not rated in the same way; instead, they display a clear hierarchy. For instance, speakers with Asian accents in English, such as Cantonese and Mandarin (Carlson & McHenry 2006), may receive lower employability ratings and be considered more suitable for low-status, non-customer-facing positions compared to European-accented speakers.

The two main theoretical accounts explain the relative ranking of foreign accents through their processing fluency and listeners' pre-existing stereotypes (e.g., Dragojevic et al. 2020). Processing fluency describes the relative ease or difficulty with which people process information. When a person's speech requires greater cognitive effort, listeners are more likely to evaluate it negatively, which is especially evident in non-standard or unfamiliar accents. Alternatively, listeners may use a speaker's accent to infer the speaker's social group and activate stereotypes associated with that group. Consequently, speakers with non-standard accents may be marked as "other" and devalued.

Although numerous studies have demonstrated the effects of accent bias in the workplace in L1-English-speaking countries, little such research has been

conducted in China. In this paper, we investigate how L2 English accents affect employment-related decisions in the Chinese context.

Methodology

Stimuli

The stimuli came from the Speech Accent Archive (Weinberger 2015), in which speakers read the 'Please call Stella' standardized passage in English. The target stimuli included German-, Korean-, and Mandarin Chinese-accented speakers; L1-English-accented stimuli were also included to create a range of accents but do not form the focus on this study. The stimuli comprised four speakers per accent, with an equal number of male and female speakers. Their ages ranged from 25 to 35.

Listeners

Listeners were 90 adult L1 Mandarin Chinese speakers residing in China.

Procedures

The survey was created and deployed using Qualtrics (Qualtrics 2020). In the survey, listeners heard and evaluated each speaker on five 5-point Likert scales that were designed to elicit judgments of the candidate's potential to become a manager, the candidate's likeability, the likelihood that the candidate would be recommended for hiring, satisfaction with the candidate's responses, and the candidate's suitability for an office job in foreign-invested companies.

Data analysis

We took the average of the ratings on the five scales, as these were highly correlated. Linear mixed-effects models (Baayen et al. 2008) were fitted in R (R Core Team 2024) to test for significance with the employability rating as the dependent variable, speaker accent and gender as the independent variables, and listener and speaker as random effects. Contrasts between accents were checked in a follow-up analysis on the linear model.

Results

Figure 1 presents descriptive statistics of the raw data, showing the mean (M) and standard deviation (SD) of the composite employability rating for each accent. The German-accented speakers (M = 3.39, SD = 0.88) received the highest employability rating, followed by the Mandarin-accented speakers (M = 2.74, SD = 1.01), while the Korean-accented speakers (M = 2.45, SD = 0.99) received the lowest rating. Variation was similar across all groups, with the Mandarin-accented speakers showing a relatively higher SD.

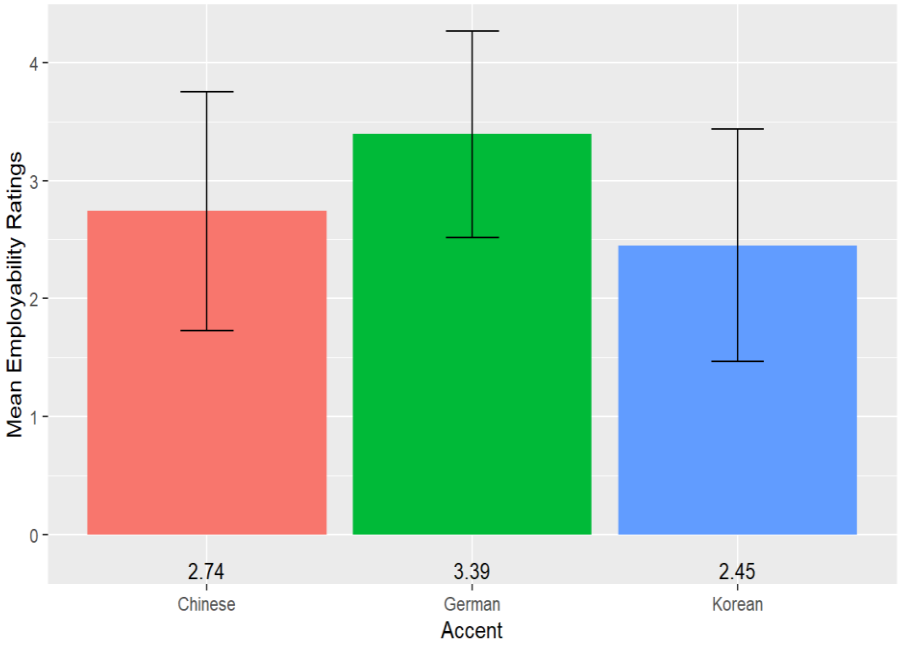


Figure 1. Mean employability evaluations and standard deviation by accent.

Table 1 presents the results of the statistical analysis. A significant difference was found between the German and Korean accents, with the Korean-accented speakers perceived to be significantly less employable. In comparison, ratings for the Mandarin-accented speakers did not differ significantly from those of the other two accent groups. Speaker gender was not a significant predictor.

Table 1. Model-derived Pairwise Comparisons Between Accents.

	estimate	SE	df	t.ratio	p.value
Chinese-German	-0.653	0.287	18.000	-2.277	0.253
Chinese-Korean	0.289	0.287	18.000	1.010	0.909
German-Korean	0.942	0.287	18.000	3.287	0.040

Discussion

The hierarchical pattern of the German and Korean accents in our results aligns with previous findings in L1 English-speaking countries, suggesting that Chinese listeners reflect L1 listener evaluations (Carlson & McHenry 2006). It is possible that phonetic similarities between German and English may enhance the perceived comprehensibility of German-accented English, ultimately leading to more positive judgments. Alternatively, listeners may rely on their ethnicity-based

stereotypes as Asian speakers are often evaluated more negatively compared to speakers from Western backgrounds. However, the data in this study were insufficient to disentangle these two explanations, which requires further investigation. When comparing the Mandarin-accented speakers with the other accent groups, familiarity with the accent did not appear to promote higher ratings despite the fact that both the speakers and the listeners spoke Mandarin as their L1. This rating pattern aligns more closely with the stereotyping explanation (cf. Spence et al. 2024). We conclude that there is an accent-based employment bias in China, thereby contributing to a broader understanding of language attitudes in English as a Foreign Language context.

References

Baayen, R.H., Davidson, D.J., Bates, D.M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59(4), 390-412.

Carlson, H.K., McHenry, M.A. 2006. Effect of accent and dialect on employability. Journal of Employment Counseling 43(2), 70–83.

Dragojevic, M., Fasoli, F., Cramer, J., Rakić, T. 2020. Toward a Century of Language Attitudes Research: Looking Back and Moving Forward. Journal of Language and Social Psychology 40(1), 60-79.

Hosoda, M., Nguyen, L.T., Stone-Romero, E. F. 2012. The effect of Hispanic accents on employment decisions. Journal of Managerial Psychology 27(4), 347–364.

Qualtrics 2020. Qualtrics XM [Software].

R Core Team 2024. The R Project for Statistical Computing (version 4.4.1) [Software].

Spence, J.L., Hornsey, M.J., Stephenson, E.M., Imuta, K. 2024. Is Your Accent Right for the Job? A Meta-Analysis on Accent Bias in Hiring Decisions. Personality and Social Psychology Bulletin 50(3), 371–386.

Weinberger, S. 2015. Speech Accent Archive. George Mason University.

Tense-lax vowels in Tibeto-Burman languages: a phonetic nalysis of Lahu

Ying Hong[1], Yingyi Zhou[2]
[1]The Hong Kong Polytechnic University, Hong Kong
[2]Southwest Minzu University, China

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/009/000669

Abstract

This study examines the phonetic nature of tense-lax vowel contrasts in Lahu, a Tibeto-Burman language. By employing acoustic analysis (duration, formants, fundamental frequency, phonation) of minimal pairs produced by four native speakers, we test the hypothesis that the contrast is a syllable-level phenomenon. Results show significant differences in onset consonant duration, vowel quality, tonal duration, and phonation type between tense and lax syllables, while F0 contours overlap. The findings confirm that the tense-lax distinction in Lahu is a coherent, syllable-level property rooted in laryngeal articulation, rather than a feature confined to the vowel.

Keywords: Lahu, tense-lax vowels, Tibeto-Burman languages, acoustic analysis

Introduction

In Tibeto-Burman languages, the tense–lax vowel contrast is a key phonetic feature. Although this topic has received sustained scholarly attention over several decades, the specific phonetic correlates of "tense" and "lax" vowels remain controversial. Early descriptions said tense vowels involved laryngeal constriction, while lax vowels lacked such tension. However, the tense–lax distinction cannot be simply reduced to the degree of laryngeal tension during articulation. In fact, tense and lax vowels in different language show significant differences in phonetic characteristics (Shi, Zhou 2005). Moreover, the so-called term "tense-lax" has not yet had a strictly defined and widely accepted definition in the field of phonetics (Zhu et al. 2011).

Lahu belongs to the Yi Branch of the Tibeto-Burman language family. A salient feature that has sparked sustained academic debate is its "tense-lax vowel" distinction. This contrast's nature has been controversial; some have suggested that it is a non-contrastive phonetic variation, or a mislabeled feature that is better explained by phonation or tonal coarticulation (Matisoff 1982, Zhu et al. 2011, Liu et al. 2024). Lahu has seven tones, among which two tones are associated with tense vowels, and the other five tones are associated with lax vowels (see the Table 1). Matisoff (1982) transcribed the syllables of Tone 5 and Tone 6 as having a glottal stop coda.

Phonetics 2025 Hong Kong: Proceedings International Conference on Phonetic Research and Applications

Table 1. The seven tones of Lahu.

Lax	Tone value	Words	Tense	Tone value	Words
Tone 1	33	ma^{33} woman			
Tone 2	21	ma^{21} classifier	Tone 5	21	maʔ ^{21} army
Tone 3	54	ma^{54} many	Tone 6	54	maʔ^{54} dream
Tone 4	45	ma^{45}son-in-law			
Tone 7	11/112	ma^{11} teach			

Recent research has shown that phonation is a significant correlate, with modal voice accompanying "lax" vowels and non-modal phonation accompanying "tense" vowels (e.g., creaky voice) (Zhu et al. 2011, Liu et al. 2024). In this study, we use a laryngeal model (for more details, see Thurgood 2007) that combines pitch (F0), vowel quality, and voice quality (phonation) to explain these differences. Our hypothesis is that the Lahu tense-lax distinction is not limited to vowel but rather appears at the syllable level, which includes the onset consonant, vowel, tone, and phonation. By analysing the acoustic correlates throughout the entire syllable, this study seeks to test this hypothesis.

Methodology

The investigation focuses on the Lahu Na dialect, and the recorded materials discussed are based on Zhu et al. (2011). This study carries out a comprehensive acoustic analysis from the aspects of onsets, rime, tones, and phonation. Four speakers (two males and two females) were selected, and the acoustic parameters of 90 pairs of tense-lax contrastive words were analysed and measured using Praat (Boersma & Weenink 2025) and Voicesauce (Shue et al. 2009). The measured parameters include the first and second formants of monophthong rime, duration of onsets and tones, fundamental frequency, harmonic energy difference, harmonic-to-noise ratio, and cepstral peak prominence (CPP).

Results

Acoustic analysis revealed consistent, significant differences between tense and lax syllables across all measured components except for F0 contour. Tense syllables had consistently shorter onset consonants and shorter tonal duration compared to their lax counterparts. For example, the average tone duration for lax syllables was 360ms, while for tense syllables it was only 146ms. Tense-lax status had a significant effect on both F1 and F2. Generally, tense vowels exhibited a lower F1, indicating a higher or more constricted articulation. For low and mid vowels, tense vowels were also more fronted. Furthermore, tense mid vowels showed a consistent tendency toward diphthongization, unlike their lax monophthongal counterparts. After duration normalization (as shown in Figure 1), there was no significant main effect of tense-lax status on F0 contour

(T2 vs T5: p = 0.256 > 0.05; T3 vs T6: p = 0.087 > 0.05). The F0 shapes of tense tones and their corresponding lax tones were nearly identical. This indicates that pitch contour is not the primary acoustic cue distinguishing these syllable types; rather, duration and phonation are the key contrastive features.

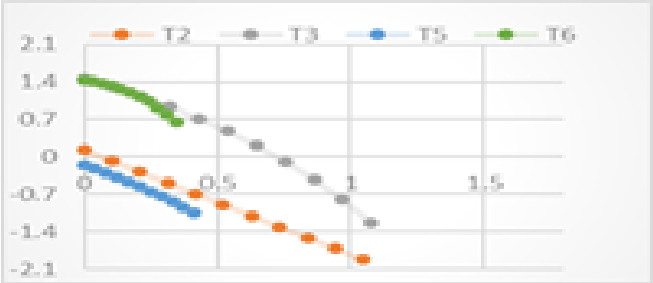


Figure 1. F0 contours of tense-lax vowels in minimal pairs.

All phonation parameters showed highly significant effects. Tense syllables were characterized by creaky voice, while lax syllables used modal voice. For most of the speakers, their lax vowels had positive H1-H2 values, indicating a kind of typical of modal voice. On the other hand, their tense vowels had strongly negative H1-H2 values, which is a classic indicator of creaky or laryngealized phonation. Generally, lax vowels had significantly higher HNR and CPP values, indicating a clearer harmonic structure with less noise. Tense vowels had lower HNR and CPP values, consistent with the aperiodicity of creaky voice.

Discussion and conclusion

The results support our hypothesis that the tense-lax contrast in Lahu is a syllable-level feature. The tense-lax distinction is not confined to the vowel but is realized through a set of coherent phonetic cues in the entire syllable, as summarized in the Table 2.

Table 2. Summary of acoustic differences between lax and tense syllables.

Syllable	Lax (Modal)	Tense (Creaky)
Onset Consonant	Longer duration	Shorter duration
Vowel Quality	Lower F1 and F2, monophthong	Higher F1 and F2, diphthongized
Tone	Longer duration	Shorter duration
Phonation (H1-H2)	Positive (modal)	Negative (creaky)
Phonation (HNR/CPP)	High (less noise)	Low (more noise)

This pattern is consistent with Thurgood's laryngeal models, which show that laryngeal articulation, can influence the whole syllable, including vowel quality,

consonant and tone duration, and voice quality (phonation). It causes the "tense" setting in Lahu, which results in higher F1, a creaky voice, and a shorter duration. This constriction is absent in the "lax" setting, which leads to longer segmental durations and modal voice. In conclusion, this study provides acoustic evidence that the Lahu tense–lax contrast is a phonetically based syllable-level phenomenon. The results show that an integrated study of consonants, vowels, tones, and phonation types is necessary for a thorough comprehension of Lahu. Acoustic parameters such as duration, formants, harmonics, and noise-to-harmonic ratio can distinguish most tense and lax syllables, though there are individual differences. The tense–lax distinction is a property of the entire syllable, rather than a contrast confined to a single segment. Lastly, the findings presented in this study are preliminary. Subsequent research will necessitate a substantially larger sample, coupled with rigorous statistical validation.

Acknowledgements

The first author acknowledges a grant (LC-2024-343(J)) from the College of Professional and Continuing Education, The Hong Kong Polytechnic University. During the preparation of this work the authors used Copilot (Microsoft) to reword and rephrase text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Boersma, P., Weenink, D. 2025. Praat: Doing Phonetics by Computer [Computer program]. Version 6.4.46. Retrieved 26 October 2023 from https://praat.org/.

Kong, J. 2001. 论语言发声 Lun Yuyan Fasheng [On Language Phonation]. Beijing: Central University for Nationalities Press.

Liu, Y., Wei, Y., Luo, Y. 2024. Tense and lax vowels in the Lahu dialect of Yunshan: A laboratory phonological study. Journal of Chinese Linguistics 52(2), 318-335.

Matisoff, J. 1982. The Grammar of Lahu, University of Califomia Press, Berkeley and Los Angeles.

Shi, F., Zhou, D. 2005. 南部彝语松紧元音的声学表现 Nanbu Yiyu songjin yuanyin de shengxue biaoxian [Acoustic performance of tense and lax vowels in Southern Yi language]. Yuyan Yanjiu [Language Research], 25(1), 60–65.

Shue, Y. L., Keating, P., Vicenik, C., Yu, K. 2009. Voicesauce [Computer program]. Program available online at http://www.seas.ucla.edu/spapl/voicesauce/. UCLA.

Thurgood, G., 2007. Tonogenesis revisited: Revising the model and the analysis. Studies in Tai and Southeast Asian Linguistics, pp.263-291.

Zhu, X., Liu, J., Hong, Y. 2011. 拉祜语紧元音：从嘎裂声到喉塞尾 Lahu yu jin yuanyin: Cong galiesheng dao housaiwei [Tense vowels in Lahu: From creaky voice to glottal stop coda]. Minzu Yuwen [Minority Languages], (3), 6-16.

An articulatory study of prenuclear glides in Southwestern Mandarin

Jing Huang[1], Feng-fan Hsieh[2]
[1]City University of Macau
[2]National Tsing Hua University

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/010/000670

Abstract

This study investigates the articulatory characteristics of prenuclear glides in Southwestern Mandarin (SWM) using electromagnetic articulography (EMA). Analyzing gestural timing between onset consonants and glides in six speakers in their twenties, we measured onset plateau (C1) duration and Consonant-Glide (C-G) lag, fitting least-squares regression lines to the data. Results for /pʲen/, /tʲen/, and /tʷan/ show flat regression slopes, indicating in-phase coordination characteristic of complex segments. The /kʷan/ pattern was less definitive but did not exhibit typical anti-phase coordination. These findings suggest that SWM prenuclear glides should be analyzed as secondary articulations of the onset consonant, forming complex segments rather than segmental sequences.

Keywords: glides, Southwestern Mandarin, EMA, articulatory timing.

Introduction

The phonological status of prenuclear glides (*jièyīn*, or "medials") remains a central question in Chinese phonology. Traditional accounts have mainly relied on phonotactic or corpus-internal evidence, whereas recent studies emphasize gestural timing as a more revealing diagnostic. Shaw et al. (2021), for example, argue that complex segments exhibit in-phase gestural coordination, with onset and glide gestures co-articulated synchronously, as in Russian palatalized stops /pj-/. By contrast, segment sequences display anti-phase coordination, where the consonant and glide are executed sequentially, as in American English /pj-/. This distinction parallels the long-standing debate over the status of prenuclear glides in Mandarin Chinese. Chao (1970), among others, treats the *jièyīn* as part of the rhyme, whereas Bao (1990) analyzes it as belonging to the onset cluster. More recent proposals by Duanmu (2007) and Lin (2007) view prenuclear glides as secondary articulations, comparable to complex segments. The present study revisits this issue by examining the timing relationships of prenuclear glides in Southwestern Mandarin using electromagnetic articulography (EMA), a technique that provides fine-grained measurements of articulatory coordination.

Phonetics 2025 Hong Kong: Proceedings International Conference on Phonetic Research and Applications

The study focuses on Southwestern Mandarin (SWM), an understudied dialect that shares a similar prenuclear glide inventory with Beijing Mandarin (i.e., /j/, /w/ and /ɥ/). By examining the temporal coordination between onset and glide gestures, we determine whether these glides function as complex segments or segmental sequences.

Method

This study examined prenuclear glides in SWM, focusing on the Chengdu–Chongqing dialect group. Six native speakers in their twenties participated in the EMA experiment. All participants were born and raised in Hubei and reported no history of speech or hearing disorders. Articulatory data were recorded using the NDI Wave system, which tracks lip and tongue movements in real time at 200 Hz. Target stimuli consisted of CGVN syllables: /pjen/, /tjen/, /twan/, and /kwan/. Each target was embedded in the carrier phrase "__, *pa* __ *pa*" ('__, give __ [sentence-final particle]') and produced in randomized order at a normal speaking rate.

For each token, six repetitions were recorded, and the second occurrence of the target stimulus was analyzed. The corresponding gestures for each target item are as follows: /pjen/ (C: LA vs. G: TBz), /tjen/ (C: TTz vs. G: TBz), /twan/ (C: TTz vs. G: LA), and /kwan/ (C: TDz vs. G: LA). Abbreviations denote: C = onset consonant, G = glide, LA = lip aperture, TT = tongue tip, TB = tongue body, TD = tongue dorsum, x = front-back dimension, z = up-down dimension (Table 1). EMA data were analyzed with the help of MView (Tiede, 2005).

Table 1. The corresponding gestures for onsets and glides.

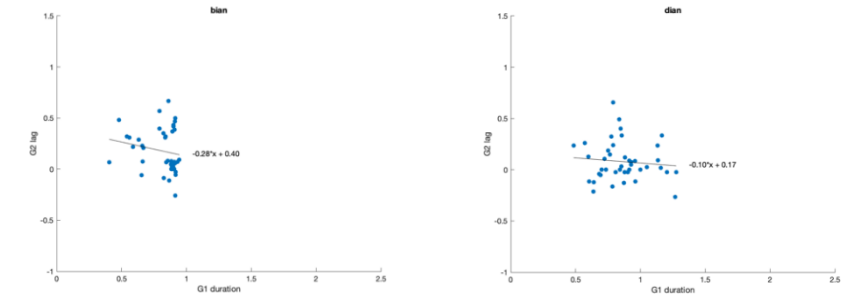
Target syllables	Onset Consonants	Glides
/pjen/	LA	TBz
/tjen/	TTz	TBz
/twan/	TTz	LA
/kwan/	TDz	LA

Results

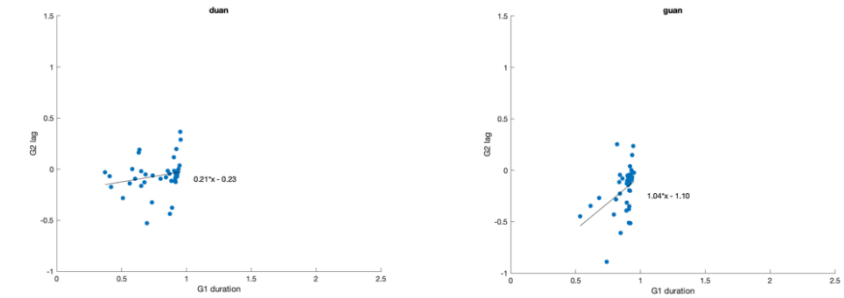
To assess the coordination between consonant and glide gestures, we measured the onset plateau duration (C1) and the consonant–glide (C-G) lag, defined as the temporal offset between the glide and consonant gestures (Shaw et al. 2021). Both measures were normalized using the reference token *pa* from the second occurrence of each trial. For each consonant–glide (CG) combination, the data were plotted as scatterplots, and least-squares regression lines were fitted to examine the relationship between C1 duration (x-axis) and C-G lag (y-axis).

As shown in Figure 1, /pjen/, /tjen/, and /twan/ all exhibit flat regression slopes, indicating that the onset and glide gestures are produced synchronously.

This pattern reflects in-phase gestural coordination, consistent with the behavior of complex segments. In contrast, /kwan/ displays a modest positive slope, suggesting a slightly different coordination pattern. However, the data points are tightly clustered, and the apparent trend is not statistically reliable. Thus, /kwan/ does not show the anti-phase (sequential) timing characteristic of independent segmental sequences.



/pjen/ /tjen/



/twan/ /kwan/

Figure 1. Regression analysis of consonant–glide coordination in Southwestern Mandarin (n = 6).

Discussion

The results reveal two distinct coordination patterns across the examined consonant–glide (CG) combinations. For /pjen/, /tjen/, and /twan/, the regression lines are essentially flat, indicating that onset and glide gestures are produced synchronously. This pattern reflects in-phase gestural coordination, characteristic of complex segments. In contrast, /kwan/ shows a mildly positive slope, suggesting a slightly different temporal relationship. However, given the tight clustering of data points, this apparent trend does not reflect a statistically robust correlation. Crucially, /kwan/ does not exhibit the anti-phase coordination that typifies independent consonant–glide sequences.

Taken together, these findings demonstrate that prenuclear glides in SWM generally pattern with their preceding consonants as complex segments rather than as independent sequential units. This conclusion supports analyses treating glides in SWM as secondary articulations of the onset consonant, paralleling proposals for Beijing Mandarin (Duanmu, 2007; Lin, 2007). These findings parallel Chen et al.'s (2024) results on Hong Kong Cantonese, which similarly show in-phase coordination. Accordingly, the relevant syllables may be transcribed as [pʲen], [tʲen], [tʷan], and [kʷan].

Conclusion

This study investigated the articulatory characteristics of prenuclear glides in Southwestern Mandarin using electromagnetic articulography. The results demonstrate that onset consonants and glides exhibit in-phase gestural coordination, characteristic of complex segments. These findings provide articulatory support for analyses treating SWM prenuclear glides as secondary articulations of the onset consonant, paralleling theoretical proposals for Beijing Mandarin. The present evidence supports the interpretation that prenuclear glides form part of the onset rather than the rime.

Acknowledgements

This study was partially supported by the Ministry of Science and Technology (MOST 109-2410-H-007-061) and was approved by the Research Ethics Committee of National Tsing Hua University (10712HS104 and 10612HS089).

References

Bao, Z. 1990. Fanqie Languages and Reduplication, Linguistic Inquiry 21(3): 317–50.

Chao, Y.-R. 1970[1968]. A grammar of spoken Chinese. Berkeley: University of California Press.

Chen, P.-R., Hsieh, F.-F., Chang, Y.-C. 2024. C-G vs. C-V Timing Differences in Hong Kong Cantonese. In C. Fougeron and P. Pierre (eds.), the Proceedings of the 13th International Speech Production Seminar, 35-38. doi: 10.21437/issp.2024-10

Duanmu, S. 2007. The phonology of standard Chinese. New York: Oxford University Press.

Lin, Y.-H. 2007. The sounds of Chinese. New York: Cambridge University Press.

Shaw, J. A., Oh, S., Durvasula, K., Kochetov, A. 2021. Articulatory coordination distinguishes complex segments from segment sequences. Phonology 38(3), 437-77.

Tiede, M. 2005. MVIEW: software for visualization and analysis of concurrently recorded movement data. New Haven, CT: Haskins Laboratories.

Language shift leading to phonemic shift in Pakistan: a case study of Pakistani Punjabi

Sundar Huma[1], Wali Muhammad Anjum[2]
Lahore College for Women University, Lahore, Pakistan
University of Sargodha, Pakistan

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/011/000671

Abstract

This paper aims to unveil the patterns of language shift, resulting in phonemic shift in Pakistani Punjabi. The risk of endangerment of language and cultural heritage, presupposes risks of exclusive identities. The disturbed identities in particular, the exclusive Punjabi identities will be taken as sample for the study. The individuals for sample have been selected through convenient sampling technique. The Open-ended interviews were conducted to find out phonemic shift of Punjabi phonemes and their effect on individual identity. The results indicated a major shift in the vowel sounds of Punjabi language. The nasalized sounds of Punjabi also deviate, leading to major vowel shift of /æ/. The study will be significant in understanding the phonemic shift in Pakistani, Punjabi language.

Keywords: language shift, phonemic shift, Punjabi language, English, exclusive identities

Introduction

Language is the way through which individuals all around the world communicate. They convey their ideas, emotions and desires (Saussare, 1916). The language system represents the psychological and cultural items of an individual (Kramsch, 2014). The cultures in various regions interact and borrow the patterns from each other. The linguistic items in particular face various changes. One word from the language gets borrowed and the other word gets changed (Kasymova & Lei, 2019). The cultural assimilation sometimes leads to language shift. Language shift refers to a situation where people of a specific community start to use another language instead of their mother language. That new language becomes their vernacular. The vernacular functions as prestigious language (Mufwene, 2020).

Language shift to be the cause of language death has been found out to be the case of colonization. The languages have died owing to the superiority of new language. The individuals do not find the situation to speak their own vernacular (Mufwene, 2020). The historical, cultural, economic, social and psychological factors become the motivation for language shift. Inferiority complex owing to colonization, social privileged conditions imposed by the colonizers, master and

slave relation between the colonizers and the colonized, posed threats to local languages (Nawaz, et al., 2012).

Whorf (1956) maintained language to be the mirror of culture. The linguistic items of a language represent the culture of particular region. The shift of language similarly harbors alienation to particular language. The phonological utterances become distant category of the native language. The shift of distinct /l/ and /n/ sound of Punjabi have been found out to be the cause of shift of mother language from Punjabi to Urdu (Arslan, et al., 2021). The present study focuses on finding out the language shift from Punjabi to English at phonemic level. It administers the concept of language shift, particularly to English language affecting phonemes of native language, particularly of Punjabi language.

Literature review

Language shift leads to language death eventually (Mufwene, 2020). Phoneme being the smallest unit of sound signal bears changes in the native language in the whole process. Punjabi language in Pakistan has seen a major shift in language at sound level. /n/ and /l/ sounds in Punjabi have seen a marked difference owing to prevalence of Urdu language (Arslan, et al., 2021).

The phonological change occurs when the language speaker changes the phoneme distribution of a language. The addition, change or insertion of a sound at a particular place in a word adds to the phonological change. Assimilation, epenthesis, deletion and metathesis have been found out as the major processes responsible for phonological change in a language (Diani, & Azwandi, 2021).

Borrowing has also been considered as one of the reasons of phonological shift leading to language shift. The prestigious language is spoken by the individuals for economic, social and cultural reasons causing them to be averting from their mother language (Kasymova & Lei, 2019).

Verticalization model of language shift has been searched out as the contributing factor to language and identity loss. The loss of local control on local institutions and their interconnectedness causes a shift in the structure of community. (Salmons, 2022). The loss or death of local language is considered as a loss of culture. The otherness creates an identity that is exclusive of native culture (Ulfa, et al., 2018).

The language shift of Punjabi language to English language however, has remained a case to be discussed. The phonological events that possess challenges to Punjabi language and cultural identity lack data. The psychological processes affiliated with language particularly of sense of belongingness to a particular culture especially Punjabi, lags the language shift phenomenon. This paper aims to find out the phonological shifts in Punjabi language.

Pronunciation	English phonemes		Punjabi Pho
ں کالج دا سٹوڈنٹ آں تے مینوں پینٹنگ دا شوق وا	**Equivalent** /æ/	**Realized** /wə/	اے
میں ٹیک کمپنی وچ جاب کرنا چاہندی واں ۔	**Equivalent** /ə/	**Realized** /wə/	آں
کمپیوٹر سائنس	**Equivalent** /ɾ/	**Realized** /ə/	ر

Methodology

The research implies qualitative research paradigm. The sample taken for the study involves 6 participants divided into two groups. One group belonged to Punjab, Pakistan. The other group belonged to Canada, America. Both the groups had their birth place in Punjab, Pakistan. Open ended interviews were conducted to search out the phonological differences. The phonological utterances were analysed instrumenting Yule's consonant and transcription chart.

Results and discussion:

The results indicated a major shift in Punjabi phonemes uttered by the American speakers. The equivalent /**æ/ sound of Punjabi was realized as**/wə/ **sound. The both the sounds in Punjabi language connote the different words. Similarly,** آں equivalent in English **/ə/ was realized as** /wə/**. The** ر consonant sound of Punjabi was realized as schwa **/ə/** sound of English. Such shift in the utterances of the group 2 (i.e. American) was found owing to the effect of English language. Particularly, the similar sound patterns of vowels of English and Punjabi language caused a merger effect of the sounds.

The present study's acoustic and perceptual data reveal substantial phonemic divergence in the Punjabi speech of "Group 2" participants (American-raised Punjabi speakers), compared to native Punjabi speakers in Pakistan. Notably:

1. The Punjabi vowel traditionally realized as /æ/ by native speakers was frequently produced by many Group 2 speakers as a reduced central vowel, perceptually akin to /ə/ or a schwa-onset sequence, sometimes with a slight glide (perceived as /wə/).

2. Similarly, Punjabi segments corresponding to schwa (inherent or epenthetic vowel in Indo-Aryan abugida writing) were often realized as English-style /ə/ rather than a "Punjabi" vowel.

Conclusion

The phonemic deviations observed among American-raised Punjabi speakers in this study offer a compelling instance of how language contact, bilingualism, and dominant-language influence can reshape the phonology of a heritage language. While existing literature on contact phonology and heritage-language maintenance supports the plausibility of such shifts, your data appear to push the boundaries, suggesting potentially novel contact-induced phonetic/phonological restructuring. Given the sociolinguistic significance of these results — for language preservation, identity, and pedagogy — they warrant further empirical investigation with acoustic analysis and native-speaker perceptual evaluation.

References

Whorf, B.L. 1956. Language, Thought and Reality. Selected Writing. (Ed.). Carroll, J.B. MIT and John Wiley & Sons, Inc. NewYork: London.

Saussare, F.D. 1916. Course in General Linguistics. (Baskin, W. Trans.). In Meisel, P. & Saussay, H. (Ed.), Course in General Linguistics Ferdinand de Saussare.

Kasymova, O., Lei, G. 2019. Linguistic and Cultural assimilation of borrowing in the Russian language: Case Study of Sinicisms. Arts and Humanities Open Access Journal, 3(1), 69- 73.

Mufwene, S.S. 2020. Language Shift. The International Encyclopedia of Linguistic Anthropology.

Arslan, M.F., Mahmood, M.A., Haroon, H. 2021. Highlighting the Sound Shift in Punjabi Language: A Corpus- Based Descriptive Study. Linguistic Forum, 3(1), 1-5.

Nawaz, S., Umer, A., Anjum, F., Ramzan, M. 2012. Language Shift: An Analysis of Factors Involved in Language Shift. Global Journal of Human Social Science Linguistics & Education, 12(10).

Ulfa, M., Isda, I. D., Purwati. 2018. The Shift of Acehnese Language: A Sociolinguistic Study to Preserve Regional Languages. Studies in English Language and Education, 5(2), 161- 174.

Diani, I., Awandi. 2021. Phonological Change Processes of English and Indonesian Language. Journal of Applied Linguistics and Literature, 6(1), 133- 148.

Salmons, J. 2022. The Last Satages of Language Shift and Verticalization: Comparative Upper Midwestern Data. In Selected Proceedings of the 11th Workshop on Immigrant Languages in the Americas (WILA 11), ed. Kelly Biers and Joshua R. Brown, 71-78.

Geminates in Libyan Arabic: investigating articulatory correlates

Amel Issa
University of Gharyan, Libya, University of Leeds, UK.

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/012/000672

Abstract

This EPG study examines articulatory correlates of singleton and geminate sonorants (/l, n, r/) in Tripolitanian Libyan Arabic (TLA). Alveolar contact (rows R1–R3) was quantified by Amount of Contact (AoC) and Centre of Gravity (CoG); mean palatograms were inspected visually. Despite prior acoustic evidence showing minimal strengthening for TLA sonorants (Issa 2017), geminates exhibit greater linguopalatal contact, more posterior contact and longer articulatory durations than singletons. Findings indicate that contrastive structure is preserved articulatorily even when acoustic strengthening is absent, supporting multimodal approaches to gemination.

Keywords: Libyan Arabic, gemination, sonorants, EPG, articulatory correlates

Introduction and motivation

Duration is the most robust correlate of gemination cross-linguistically (Khattab, Al-Tamimi 2008, Arvaniti 1999), but non-durational articulatory indices — tongue shape, contact region and contact area — also differentiate singletons and geminates (Local, Simpson 1988, Payne 2006). EPG work reports laminal vs apical contact and increased contact area for geminates in several languages (Payne 2006; Kraehenmann, Lahiri 2007, 2008, Ridouane 2007). In TLA, gemination is pervasive and contrastive, yet articulatory evidence is lacking; acoustics show similar formant structure and intensity for singleton and geminate sonorants (Issa 2017). This study asks whether articulatory measures reveal systematic differences between singleton and geminate sonorants in TLA.

Method

Participant and design

One native male TLA speaker (34 y), born and raised in Tripoli, participated; he reported no speech or hearing deficits, acquired TLA as a first language and had L2 English. Resource and recruitment constraints limited the study to a single informant

Phonetics 2025 Hong Kong: Proceedings International Conference on Phonetic Research and Applications

EPG data collection

EPG recordings used WinEPG with an Articulate-style custom palate (62 electrodes; eight rows). Audio was sampled at 22.05 kHz; EPG at 100 frames s^{-1}. The participant read randomized trisyllabic minimal/near-minimal pairs containing medial intervocalic /l, n, r/ in singleton and geminate forms, elicited in the carrier ma tguliːʃ _____ taːni. Fillers were interspersed.

Data analysis

Data were annotated in Articulate Assistant v1.18 (waveform, spectrogram and palate contact displayed concurrently). Analyses targeted the anterior three electrode rows (R1–R3). Measures: Amount of Contact (AoC; proportion of electrodes active), Centre of Gravity (CoG; electrode-index weighted mean), and mean palatograms averaged over the constriction interval. Statistical tests were ANOVAs with factors phonological status (singleton/geminate) and sound category (/l, /n, /r/); post-hoc LSD tests applied where relevant.

Results

Amount of Contact (AoC)

AoC patterns differ reliably by sound and phonological status. Mid-frame ANOVA: sound category significant ($F(2,5)=54.61$, $p<0.001$); phonological status non-significant ($F(3,5)=0.10$, $p=0.955$); interaction significant ($F(5,22)=3.72$, $p<0.05$). Max-frame ANOVA: phonological status significant ($F(3,5)=6.19$, $p<0.05$); sound category significant ($F(2,5)=80.83$, $p<0.001$); interaction non-significant ($F(5,22)=2.19$, $p=0.092$). Interpretation: geminates show greater maximum AoC and longer articulatory durations than singletons (Fig. 1).

Centre of gravity (CoG)

CoG is systematically lower (more posterior) for geminates than for singletons. Mid-frame ANOVA: sound category significant ($F(2,5)=11.91$, $p<0.05$); interaction significant ($F(5,22)=30.12$, $p<0.001$); phonological status non-significant ($F(3,5)=1.61$, $p=0.299$). Post-hoc tests show CoG(singleton) > CoG(geminate), $p<0.001$. Max-frame CoG patterns mirror mid-frame results. These CoG shifts are consistent with more posterior/laminal contact in geminates and more anterior/apical contact in singletons (Fig. 2).

Visual palatograms

Mean palatograms corroborate quantitative measures. For /l/ geminates show increased contact at R2–R3 and occasional posterior extension to R4; singletons concentrate at R1. For /n/ geminates occlude across R1–R3 (apico-laminal),

while singletons concentrate on R1. The rhotic /r/ likewise shows greater contact area in geminates (Fig. 3). Visual inspection therefore supports AoC and CoG findings.

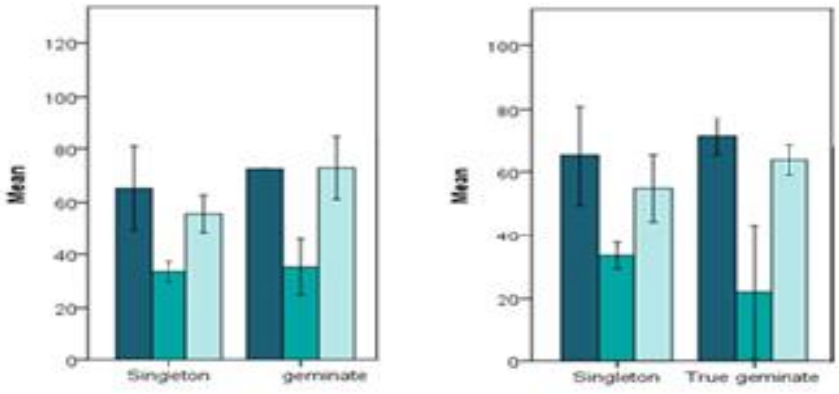


Figure 1. AoC Max-Frame (left) and Mid-frame (right).

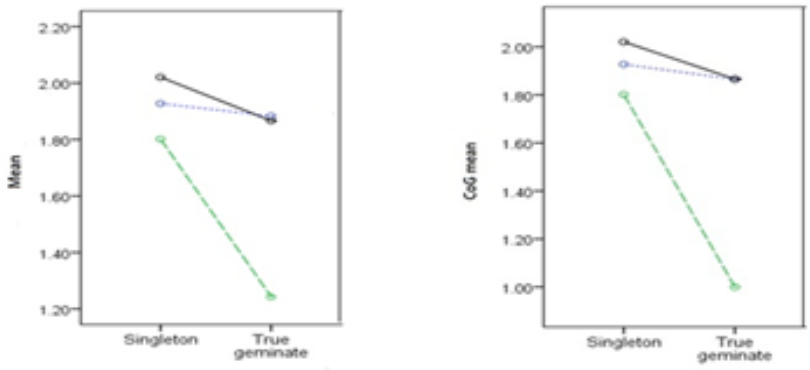


Figure 2. CoG Max-Frame (left) and Mid-frame (right).

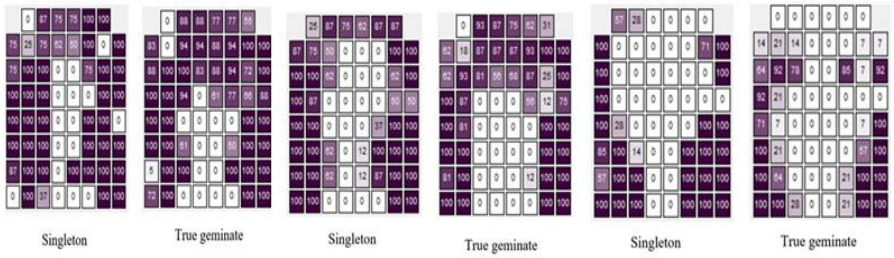


Figure 3. Palatograms of the singleton and geminates/l/ (left), /n/ (middle), and /r/ (right).

Discussion and conclusion

The AoC, CoG and palatographic inspection consistently distinguish singleton and geminate sonorants in TLA: geminates involve greater and more posterior linguopalatal contact and longer articulatory durations. These spatial and temporal enhancements parallel reports for Italian and other languages (Payne 2006, Kraehenmann, Lahiri 2008, Ridouane 2007) and suggest articulatory fortition. However, Issa (2017) found no acoustic strengthening for TLA sonorants (similar formant structure and intensity for singletons and geminates), therefore, the present articulatory strengthening may be primarily an articulatory phenomenon not reflected in the acoustic signal. This dissociation has implications for phonological representation and articulatory planning of geminates and supports the use of multimodal (acoustic + articulatory) evidence in gemination research. This study provides the first EPG documentation of gemination in a Libyan Arabic dialect and contributes new data to the typology of geminate consonants.

References

Arvaniti, A. 1999. Effects of speaking rate on the timing of single and geminate sonorants. Proceedings of the XIVth International Congress of Phonetic Sciences 599-602. San Francisco, CA.

Issa, A. 2015. On the phonetic variation of intervocalic geminates in Libyan Arabic. Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, UK.

Issa, A. 2017. Acoustic cues to the singleton-geminate contrast: the case of Libyan Arabic sonorants. Proceedings of the 18th conference of the international speech communication2988-2992. Interspeech 2017, Stockholm.

Kraehenmann, A., Lahiri, A. 2007. Non-neutralizing quantity in word-initial consonants: artriculatory evidence. In J. Trouvain and W. Barry (eds.), Proceedings of the 16th International Congress of Phonetic Sciences 465–468, Saarbrücken.

Kraehenmann, A., Lahiri, A. 2008. Duration differences in the articulation and acoustics of Swiss German word-initial geminate and singleton stops. Journal of the Acoustical Society of America 123(6), 4446–4455.

Khattab, G., Al-Tamimi, J. 2008. Durational cues for gemination in Lebanese Arabic. Language and Linguistics 22, 39-55.

Lavoie, L. 2001. Consonant strength: phonological patterns and phonetic manifestations. Garland Publishing, Inc.

Local, J., Simpson, A. 1988. The domain of gemination in Malayalam. In D. Bradley., E. J. A. Henderson., M. Mazaudon, (Eds). Prosodic Analysis and Asian Linguistics: To Honour R. k. Sprigg. Pacific Linguistics, C-104, 33-42.

Payne, E.M. 2006. Non-durational indices in Italian geminate consonants. Journal of the International Phonetic Association, 36(1), 83-95.

Ridouane, R. 2007. Gemination in Tashlhiyt Berber: an acoustic and articulatory study. Journal of the International Phonetic Association, 37(2), 119-142.

Enhancing post-secondary language majors' accentual awareness through video analysis and reflection

Wience Wing-sze Lai
The Hong Kong Polytechnic University, Hong Kong

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/013/000673

Abstract

This study examines the impact of video analysis and reflective practice on accentual awareness in phonetic education for language majors. Twenty-two students completed assignments using authentic YouTube videos featuring diverse English accents, identifying segmental and suprasegmental features, transcribing words in IPA, and providing written analyses and reflections. Results showed strong performance in understanding and explanation, relevance and originality, and analyses, but weaker reflection, indicating challenges in metacognitive skills. Correlation analysis revealed understanding and explanation, relevance and originality as key predictors of success, while reflection had minimal influence. Students excelled in segmental recognition but struggled with suprasegmental features. Findings highlight the need for explicit instruction and scaffolded reflection to enhance analytical depth and self-awareness in phonetic education.

Keywords: phonetic education, accentual awareness, video-based learning, reflective practice, suprasegmental analysis

Introduction

English is spoken worldwide with diverse accents reflecting regional, social, and cultural variation. British Received Pronunciation (RP) has traditionally served as the standard in phonetic education and pronunciation teaching (Roach, 2004; Wells, 1982). Conventional instruction prioritises segmental features within RP or General American (GA), offering limited exposure to authentic speech from other varieties (Jenkins, 2000). This narrow focus can hinder learners' ability to interpret diverse accents in academic and professional contexts. Recent research highlights technology-enhanced approaches to address these limitations. Video-based tasks improve awareness of suprasegmental features such as rhythm and intonation, while shadowing enhances pronunciation and attitudes towards authentic input (Phan et al., 2024). Structured reflection supports metacognitive growth, enabling learners to monitor progress and refine strategies (Paterson, 2022). Exposure to multiple accents positively influences awareness and attitudes towards global Englishes (Chiu & Lin, 2024). Although learner-produced videos are increasingly used (Lam & Yunus, 2022), their application in phonetic

education, especially for fostering metacognitive engagement, remains limited. Instruction typically relies on scripted teaching videos rather than authentic speech, and systematic engagement with learner-generated video analyses and unscripted multimedia resources is rare (Zhang et al., 2022; Wedlock & Binnie, 2023; Mukhtarova, 2024).

To address these gaps, this study implements a two-tiered video task structure within phonetic education. First, students analyse authentic YouTube videos featuring global English varieties, identifying segmental and suprasegmental features. Second, they create self-recorded explanatory videos articulating phonetic differences. This approach discourages reliance on pre-existing IPA transcriptions and fosters deeper engagement through analysis and explanation. Guided reflection further promotes metacognitive awareness, enabling learners to evaluate understanding and identify challenges. Against this backdrop, the study explores two research questions: (1) How does engaging in authentic video analysis and explanatory video creation influence students' ability to perceive and articulate accent differences beyond RP? (2) How do these tasks, combined with guided reflection, impact phonetic analytical skills and metacognitive awareness?

Methodology

Twenty-two post-secondary language majors enrolled in a phonetics course participated. The assignment was graded as part of coursework to ensure authentic engagement. Students provided informed consent through a written statement embedded in the assignment template, confirming agreement to anonymised reporting and data destruction within two years. The assignment comprised three components: (1) a 2–3 minute MS Teams video analysing a YouTube accent sample, explaining segmental and suprasegmental features, (2) IPA transcription of five words differing from RP and a 150–200 word analysis supported by two sources, and (3) a 50–100 word reflection on accent familiarity and challenges. Strict technical requirements ensured academic integrity. Students were required to transcribe authentic speech rather than copy IPA from dictionaries, promoting practical transcription skills.

Quantitative scores were collected for each assessment criterion. Correlation analysis was conducted to examine how scores for "Understanding and Explanation", "Relevance and Originality", "IPA Transcription", "Analyses", and "Depth of Thoughts in Reflection" were related to the overall "Assignment Mark" and "Coursework Mark". Qualitative data from student reflections and written analyses were thematically coded to identify patterns in accent awareness and learning challenges. All procedures adhered to institutional ethical standards.

Results

Two-tiered tasks broadened exposure to accents beyond RP, including Hong Kong Cantonese, Mandarin, American, Canadian, Scottish, Liverpool, Italian,

Japanese, Korean, Malaysian, Filipino, Indian, Australian, and New Zealand (Māori). This breadth of engagement was reflected in strong performance in understanding and explanation (average: 79.1/100) and relevance and originality (83.6/100), indicating that learners were able to identify salient phonetic features and articulate distinctions across accents with clarity and creativity. For example, students observed features such as /v/→/b/, /r/ and /l/ → [ɾ], and vowel epenthesis in Japanese English; /θ/→/p/ and vowel insertion after final consonants in Korean English; omission of final consonants, /n/-/l/ confusion, and lack of vowel length distinction in Cantonese English; and retroflexion of /t/ and /d/ along with diphthong monophthongisation in Indian English. Suprasegmental analysis was less consistent. Some students identified mora/syllable-timed rhythm and monotone intonation in Japanese and Korean English, and expressive intonation in Italian and Indian English. However, many analyses lacked depth in describing rhythm and stress patterns, suggesting the need for scaffolding in prosodic analysis. IPA transcription scores were moderate (75/100), indicating technical competence with room for improvement. Most students provided accurate IPA transcriptions for at least five words, demonstrating proficiency in phonemic representation. However, some assignments revealed minor errors, such as missing stress marks or incomplete transcription.

Reflection scores were lowest (65.9/100), indicating challenges in articulating nuanced distinctions and connecting observations to learning strategies. Correlation analysis revealed understanding and explanation strongly predicted assignment and coursework marks ($r = 0.87$ and $r = 0.79$, respectively), followed by relevance and originality ($r = 0.75$ and $r = 0.52$). Technical proficiency in transcription ($r = 0.58$ and 0.65) and analytical writing ($r = 0.62$ and 0.4) exerted a moderate influence. Reflection exhibited weak correlations (ranging from -0.27 to 0.12), suggesting metacognitive skills require targeted support. Strategies such as structured prompts, peer feedback, and exemplar-based modelling could help students move beyond surface-level reflections.

Discussion and conclusion

This study demonstrates the pedagogical value of integrating authentic video analysis and learner-generated explanatory tasks into phonetic education. The approach aligns with calls for technology-enhanced instruction beyond RP-centric models (Roach, 2004; Jenkins, 2000) and supports findings on the benefits of authentic input for accent recognition and positive attitudes towards global Englishes (Phan et al., 2024; Chiu & Lin, 2024). Students showed strong analytical writing and IPA transcription skills (Lam & Yunus, 2022), and correlation analysis confirmed that detailed analyses predicted higher overall performance.

However, challenges persisted in suprasegmental analysis, echoing Jenkins' (2000) observation that prosody remains difficult for learners. Many reflections were descriptive rather than evaluative, indicating that metacognitive growth requires structured prompts and exemplars (Paterson, 2022). Limited progress in rhythm and intonation analysis highlights the need for targeted scaffolding and peer feedback mechanisms.

Overall, multimedia tasks broadened exposure to diverse English accents, moving beyond RP and GA, and enhanced segmental analysis skills. Structured reflection emerged as a distinct area needing explicit support to foster deeper metacognitive engagement. Providing clear examples of suprasegmental analysis and collaborative review can strengthen critical insight and learning strategies. This model offers a practical framework for bridging theory and practice in multilingual phonetic education.

References

Chiu, C., Lin, Y. 2024. An Attitude-Changing Investigation into English Accents with Explicit Instruction on Global Englishes. Journal of Educational Practice and Research 37, 45-62. Taiwan, National Academy for Educational Research.

Jenkins, J. 2000. The Phonology of English as an International Language. Oxford, Oxford University Press.

Lam, N., Yunus, M. 2022. Student-Produced Video for Learning: A Systematic Review. Journal of Language Teaching and Research, 13(4), 780–792.

Mukhtarova, A. 2024. Enhancing Foreign Language Learning through Authentic Video Materials: Theoretical and Practical Perspectives. Journal of Language Pedagogy and Applied Linguistics, 6(1), 45–60.

Paterson, M. 2022. Prompting Metacognitive Reflection to Facilitate Speaking Improvements in Learners of English as a Foreign Language. English Teaching & Learning 46, 1-20. Springer, Berlin.

Phan, T., Nguyen, H., Le, Q. 2024. Effects of Video-Based Shadowing on Suprasegmental Features: EFL Learners' Pronunciation Performance and Attitudes. English Teaching & Learning 48, 55-72. Springer, Berlin.

Roach, P. 2004. British English: Received Pronunciation. Journal of the International Phonetic Association 34, 239-245. Cambridge University Press.

Wedlock, J., Binnie, T. 2023. Selecting and Using Authentic Videos for Intentional Second Language Learning: Nine Considerations. Journal of Language and Literacy Education, 19(2), 101–115.

Wells, J. C. 1982. Accents of English. Cambridge, Cambridge University Press.

Zhang, Y., Li, H., Chen, X. 2022. A Decade of Short Videos for Foreign Language Teaching and Learning: A Review. Journal of Language Teaching and Research, 13(5), 950–965.

Seeing, hearing, and feeling L2 sounds through metaphoric gestures

Enid Lee
Okinawa International University, Japan

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/014/000674

Abstract

This paper presents an embodied approach to L2 pronunciation instruction using metaphoric gestures. Drawing on multimodal communication and embodied cognition, it argues that mapping physical actions onto abstract phonological concepts makes L2 sounds more accessible, tangible, and memorable. A pedagogical framework is introduced featuring thirteen metaphoric gestures designed to address segmental and suprasegmental challenges faced by Japanese EFL learners, such as /r/-/l/ distinction and English rhythm. These techniques foster visualization, self-monitoring, and multisensory engagement. Student feedback and high evaluations over seven years indicate the method's effectiveness and appeal. The paper concludes with cultural and pedagogical considerations and calls for further empirical research.

Keywords: pronunciation, metaphoric gesture, embodied cognition, Japanese EFL

Introduction and theoretical framework

Gesture studies have long established an intrinsic link between speech and gesture (McNeil 2005). Research in educational contexts suggests that animated teaching enhances engagement and effectiveness (Richmond 1996, 2002). In pronunciation pedagogy, kinesthetic or embodied approaches are not new (Acton 1984), but the systematic use of gestures as a core instructional strategy remains underexplored. This study introduces a structured framework of metaphoric gestures to address persistent perceptual and articulatory challenges, with a focus on Japanese EFL learners. Metaphoric gestures, which map physical actions onto abstract concepts (Cienki, Müller 2008), draw on theories of multimodal communication and research on embodied cognition (Glenberg, Kaschak 2002, Morett 2019). By engaging visual, auditory, and tactile modalities, such gestures can make pronunciation more accessible, tangible and memorable. This study demonstrates how these gestures can improve learners' accuracy and intelligibility while creating an engaging classroom environment.

Purpose and methodology

The purpose of the study is to demonstrate a practical framework for using metaphoric gestures to address persistent segmental and suprasegmental challenges for Japanese EFL learners. The techniques were developed and

refined over seven years in a 15-week undergraduate pronunciation course (N=40-60 per class). Each lesson dedicated 10-15 minutes to introducing one or two gestures, involving teacher modelling, chorus repetition, and individual practice. These gestures were then consistently integrated into subsequent lessons as "catchments" (McNeill 2005) for review and corrective feedback. Findings and reflections reported here are based on qualitative data from course evaluations and student feedback.

Gestural framework

The framework consists of 13 catchments designed to be reiterative and metaphoric. They target specific challenges, often leveraging cultural familiarity (e.g., the Japanese "*yubikiri*" for liaison) and providing contrast with L1 phonemes to prevent negative transfer. These techniques enhance visualization, self-monitoring, and retention by activating multiple senses simultaneously (Macedonia & von Kriegstein 2012).

Table 1: Summary of Selected Gesture Catchments

Catchment	Target Feature	Key Gesture Description
#1 Trapped Honeybee	Voiced fricative /z/	😬 Tense smile, feel throat vibration, avoid tongue contact.
#5 Angry Dog	Approximant /ɹ/	😠 Growl with rounded lips, tongue elevated but not touching.
#8 Returning Boomerang	Rhotic vowels (e.g., /ɪɚ/)	↩️ Arm extends for vowel, retracts for /ɹ/ sound.
#11 Knock, Knock	Word Stress & Rhythm	👊 Hard/slow knock on stressed syllables, soft/fast on unstressed.
#12 Pinky Swear (*yubikiri*)	Liaison	🤙 Hook pinkies to link words (e.g., “find‿out”)

Detailed examples

Catchment #1: Trapped Honeybee (/z/)

Students imagine a bee buzzing in their mouth. They maintain a tense smile, focus on vocal cord vibration, and consciously avoid the tongue-tip contact that produces the affricate /ʤ/(common in Japanese). Practice: *zip, zebra, prize*.

Catchment #5: Angry Dog (/ɹ/)

Students create a growling sound with rounded lips, keeping the tongue tip from touching the alveolar ridge. This contrasts explicitly with the Japanese flap /ɾ/, which involves a quick tap. Practice: *red, river, right*.

Catchment #11: Knock, Knock (Stress and Rhythm)

Students knock on a desk, varying force and speed to physically manifest English stress-timing (e.g., *PHO-to-graph* vs. *pho-TO-gra-phy*), contrasting it with Japanese mora-timing. Practice: *AC-ti-vate* vs. *ac-ti-VA-tion*, *NA-tio-nal* vs. *na-tio-NA-li-ty*.

Student feedback and conclusion

Course evaluations from 2018–2024 consistently showed high satisfaction, with mean scores of 4.4–4.9 (out of 5), surpassing university averages by 0.4–0.8 points. Student comments emphasized the clarity, memorability, and enjoyability of gestures—for example, the "Angry Dog" helping distinguish /ɾ/ from /ɾ/. While findings suggest strong engagement and perceived learning, they rely on qualitative feedback. Further empirical research with controlled testing is needed to measure pronunciation gains objectively and assess long-term, cross-linguistic applicability. Teachers must also consider cultural sensitivity, learner age, and proficiency, as gestures carry different meanings across contexts (Kita 2009, Tellier 2008). Gestures should complement rather than replace auditory and analytic instruction.

This study highlights the pedagogical potential of metaphoric gestures in making abstract phonological concepts more concrete and accessible, supporting awareness, monitoring, and retention in L2 pronunciation.

References

Acton, W. 1984. Changing fossilized pronunciation. TESOL Quarterly, 18(1), 71–85.

Cienki, A., Müller, C. 2008. Metaphor, gesture, and thought. In Gibbs, R. W. Jr. (ed.), The Cambridge Handbook of Metaphor and Thought, 483–501. New York, Cambridge University Press.

Glenberg, A.M., Kaschak, M.P. 2002. Grounding language in action. Psychonomic Bulletin & Review, 9(3), 558–565.

Kita, S. 2009. Cross-cultural variation of speech-accompanying gesture: A review. Language and Cognitive Processes, 24(2), 145–167.

Macedonia, M., von Kriegstein, K. 2012. Gestures enhance foreign language learning. Biolinguistics, 6(3–4), 393–416.

McNeill, D. 2005. Gesture and Thought. Chicago, University of Chicago Press.

Morett, L.M. 2019. When hands speak louder than words: The role of gesture in the communication, encoding, and recall of words in a novel second language. The Modern Language Journal, 103(3), 640–655.

Richmond, V.P. 1996. Nonverbal communication in the classroom. Acton, Tapestry Press.

Richmond, V.P. 2002. Socio-communicative style and orientation in instruction. In Chesebro, J.L., McCroskey, J.C. (eds.), Communication for Teachers, 104–115. Boston, Allyn and Bacon.

Tellier, M. 2008. The effect of gestures on second language memorisation by young children. Gesture, 8(2), 219-235.

Pitch relationship and phonation cues in Mandarin tone perception

Ok Joo Lee, Kyungmin Lee
Seoul National University, Korea

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/015/000675

Abstract

This study investigates how pitch relationship and phonation cues shape Mandarin tone perception among native and non-native listeners. Focusing on high and low tones (Tones 1 and 3) in Mandarin, we examined the effects of pitch relationships between syllables and phonation type (modal vs. creaky) across five groups: native Mandarin listeners, and Cantonese and Korean listeners with low or high Mandarin proficiency. Results from a tone identification task with 140 participants show that the pitch relationship was the primary cue, while creaky voice notably facilitated Tone 3 responses. Native listeners were most sensitive to pitch, whereas Cantonese and Korean listeners relied differently on phonation cues depending on L1 background and L2 proficiency, revealing distinct perceptual adaptation strategies.

Keywords: pitch relationship, phonation, Mandarin, tone, perceptual strategies

Introduction

Mandarin has four lexical tones, '55', '35', '214', and '51', in Chao's five-number scale (Chao 1930), conventionally labeled Tones 1 to 4. Tone 3 surfaces as a low tone when followed by Tones 1, 2, or 4, and as Tone 2 before another Tone 3. The low-tone realization predominates in the lexicon and natural speech (Duanmu 2000/2007; Zhang 2010). While pitch height and movement are crucial for tone identification, pitch relationships between syllables and phonation also affect perception: creaky voice, often produced with Tone 3, biases listeners toward Tone 3 identification (Kuang 2017; Huang 2020; Lee and Lee 2022). The goal of this study is to examine how pitch relationship and phonation cues, which are not inherent to the target tone itself, interact in native and non-native perception of Mandarin Tones 1 and 3. A tone identification task was conducted with 140 participants, including native Mandarin listeners as well as Cantonese and Korean listeners with differing levels of Mandarin proficiency, to assess the effects of L1 background (tone vs. non-tone) and L2 proficiency (low vs. high).

Methods

A tone identification experiment was conducted with five listener groups: 32 native Mandarin listeners (Man), 22 high-proficiency Cantonese listeners (CanH),

26 low-proficiency Cantonese listeners (CanL), 30 high-proficiency Korean listeners (KorH), and 30 low-proficiency Korean listeners (KorL). Mandarin proficiency was classified primarily based on *Putonghua Shuiping Ceshi* or *Hanyu Shuiping Kaoshi* scores and the length of Mandarin study.

The stimuli were disyllabic expressions with a level tone on the first syllable and a rising tone on the second, perceived as either T1+T2 or T3+T2 depending on first-syllable pitch height. Nine segmentally identical pairs, each with a plosive onset and high, mid, or low vowel, were used (e.g., *baohan* [pɑo.xan], "to contain; to be filled with"). Recordings by one male and one female Mandarin speaker were resynthesized, manipulating F0 along an 11-level continuum and first-syllable phonation type. The second-syllable onset pitch was fixed at midrange to control contrast effects. A total of 360 stimuli were presented individually via *LabVanced* in eight counterbalanced blocks, and participants identified each as T1+T2 or T3+T2.

The tone identification responses were analyzed using the R package *Hmisc* (Harrell 2022). Fixed factors included listener group (Man, Can, Kor), first-syllable phonation (modal, creaky), ΔF0 (Syllable 1 height minus Syllable 2 onset), onset type, pitch relationship (Syllable 1 < Syllable 2 or Syllable 1> Syllable 2), and speaker gender (M, F), with participants treated as random effects. A mixed-effects logistic regression tested main and interaction effects of listener group with other variables. *Bonferroni* post hoc analyses were conducted, with a 95% confidence interval applied.

Results

Figure 1 summarizes the effects of F0 and phonation on tone perception. All groups except KorL showed categorical perception, with the F0 boundary increasing in the order KorH < Man < CanL < CanH. KorH exhibited a broader Tone 1 category, whereas CanL and CanH showed broader Tone 3 categories. Creaky voice enhanced Tone 3 identification across groups, particularly among Korean listeners. For KorL, however, it did not yield categorical perception but instead enhanced low-pitch perception acoustically, indicating limited reliance on phonation for tonal contrast.

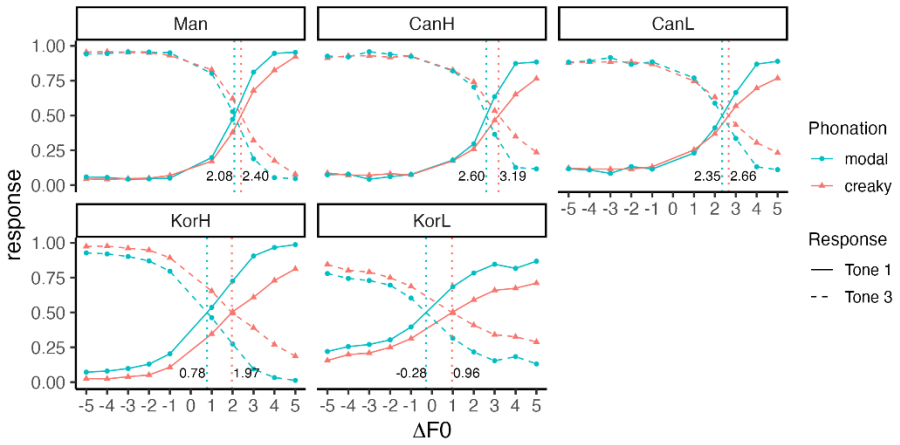


Figure 1. ΔF0 and phonation effects across listener groups

The mixed-effects logistic regression model examined the effects of pitch, phonation, and language experience factors. Significant main effects were found for the S1–S2 pitch relationship, ΔF0, phonation type, and listener group (all $p < .0001$), with pitch relationship and ΔF0 as the strongest predictors (χ^2 = 8,079.04 and 7,659.97). Creaky voice elicited more Tone 3 responses than modal voice, and the S1 < S2 condition produced more Tone 3 responses than S1 > S2. KorL yielded the fewest Tone 3 responses. Significant pitch–phonation and group–cue interactions indicate that pitch effects are modulated by phonation and language experience. *Bonferroni* post-hoc analyses further demonstrated the effects of L1 background and L2 proficiency on both the S1–S2 pitch relationship and phonation cues. In the S1 < S2 condition, Tone 3 identification ranked Man, CanH, CanL, KorH > KorL ($p < .0001$), indicating weaker Tone 3 perception for KorL. In the S1 > S2 condition, Tone 1 identification ranked KorH, KorL > Man > CanH, CanL, reflecting L1 background effects. Under modal voice, Tone 1 identification ranked KorL > KorH > Man, CanH, CanL, whereas under creaky voice, Tone 3 identification ranked Man, CanH, CanL, KorH > KorL, indicating that creaky phonation facilitated Tone 3 identification for KorH. Post-hoc analyses of the pitch–phonation–group interaction revealed distinct Tone 3 identification patterns: in the S1 < S2 condition, KorL showed the lowest Tone 3 perception across phonation types (all $p < .0001$), whereas KorH approached native levels under creaky voice. Mandarin listeners outperformed CanL and KorH under modal but not creaky voice, suggesting their stable reliance on pitch cues. In S1 > S2 conditions, Korean listeners showed the highest Tone 1 identification, which decreased with creaky voice, while Cantonese listeners consistently favored Tone 3 identification.

Conclusion

The findings of the present study highlight the impact of language experience on tone perception. Mandarin and Cantonese listeners, unlike Korean listeners, relied more on contextual pitch relationship cues, interpreting tones relative to adjacent syllables. By contrast, Korean listeners attended mainly to target pitch height, though high-proficiency learners showed greater use of contextual cues, resembling native tone-language patterns. Tone-language listeners also differed notably. Compared with Mandarin listeners, Cantonese listeners showed a stronger bias toward Tone 3, likely due to their L1 tone system, which includes two low-level tones ('22', '33'). Phonation also significantly influenced tone perception across listener groups; creaky voice enhanced Tone 3 identification, especially among high-proficiency Korean listeners, indicating adaptive cue weighting with experience. Unlike tone-language listeners, they relied more on phonation, reflecting the perceptual salience of creaky voice.

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government [NRF-2022S1A5A2A01043267].

References

Chao, Y. R. 1930. A system of tone letters. Le Maître Phonétique 45, 24-27.

Duanmu, S. 2000/2007. The Phonology of Standard Chinese. Oxford, Oxford University Press.

Harrell, F. 2022. Hmisc: Harrell Miscellaneous. R package version 4.7-1. https://CRAN.R-project.org/package=Hmisc

Huang, Y. 2020. Different attributes of creaky voice distinctly affect Mandarin tonal perception. The Journal of the Acoustical Society of America 147(3), 1441.

Kuang, J. 2017. Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. The Journal of the Acoustical Society of America 142, 1693.

Lee, K., Lee, O.J. 2022. Native and non-native perception of Mandarin level tones. Linguistic Research 39(3), 567-601.

Zhang, J. 2010. Issues in the analysis of Chinese tone. Language and Linguistics Compass 4(12), 1137-1153.

Acoustic features of Cantonese speech acts: prosodic evidence from words and sentences

Meixuan Li, Bingxin Liu, Si Chen
The Hong Kong Polytechnic University, Hong Kong

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/016/000676

Abstract

Prosodic cues provide a suprasegmental access to speech-act perception, yet they remain under-explored in tonal languages. The present study analysed the production of six speech acts (Statement, Doubt, Suggestion, Command, Celebration and Complaint) by twenty Cantonese adults. Mean F_0, F_0 range, intensity and duration jointly doubled chance-level classification, despite competition from lexical tone. Doubt and Suggestion exhibited the highest mean F_0 and widest span; Statement and Complaint occupied the lowest pitch region; Command showed the narrowest span and greatest loudness; Complaint was longest in duration. The findings confirm that global intonation, intensity and timing reliably encode pragmatic force in Cantonese.
Keywords: Cantonese, speech act, prosody, production

Introduction

Language can do things. When words are uttered, the intentions of the speaker are implied and inferred by the listener in real time. According to Speech-Act theory (Austin 1962; Searle 1969), this inference relies on illocutionary-force-indicating devices distributed across the lexical, syntactic and prosodic tiers. Among these, prosody is uniquely efficient with its independency. Previous studies have shown that global F_0 height, contour shape and boundary direction map systematically onto speech-act categories (Hellbernd & Sammler 2016). Cantonese complicates the picture because the same F_0 channel is already committed to lexical tone. Production studies on questions, statements and imperatives (Ma et al. 2008) suggest that register shifts and boundary tones can coexist with lexical tone. Yet the empirical base remains fragmentary, limited to isolated contrasts and adult speakers. The present experiment therefore asks: How are six everyday speech acts encoded prosodically in Cantonese, and do global acoustic features differentiate them despite lexical tone?

Hypothesis

Speech acts differ systematically in (i) Mean F_0 & pitch range, (ii) Intensity, (iii) Duration. These cues allow machine classification above chance.

Phonetics 2025 Hong Kong: Proceedings International Conference on Phonetic Research and Applications

Methods

Participants and materials

Twenty Cantonese adults (10 F/10 M; 21–27 yrs; with normal hearing and speech) were recorded. Materials comprised nine disyllabic words covering tone combinations and six carrier sentences "我哋 + Verb-Object". Six acts × two repetitions yielded 180 tokens per speaker. Contextual scenarios were scripted for natural production.

Procedures

Each trial presented a scenario prompt plus an AI partner's turn. The speaker responded with the target utterance in context. Recordings were made in a sound-proof booth with a high-quality microphone.

Acoustic analysis

Manual segmentation was performed by trained phoneticians. F_0 was normalised to semitones per speaker. Extracted measures were Mean F_0, F_0 range, Mean RMS intensity and Duration.

Data analysis

Descriptive statistics (M ± SE) were plotted for each parameter × speech act × utterance type. A jack-knife linear discriminant analysis (LDA) with leave-one-speaker-out cross-validation assessed how well the four-cue prosodic vector predicted speech-act category. Significance against the 16.7 % chance level was evaluated with χ^2 tests.

Results

Descriptive patterns

Descriptive statistics showed intention-specific prosodic patterns in both word and sentence conditions (Table 1&2). Doubt and Suggestion were realized with the highest mean F0 and the widest pitch spans, whereas Statement and Complaint occupied the lowest pitch region and Command exhibited the narrowest span. Command and Celebration were the loudest (≈ 65 dB), Complaint the longest (≈ 0.8 s in words, 1.3 s in sentences), and Command the shortest.

Classification

A jack-knife linear discriminant analysis using only these four cues (mean F0, F0 range, intensity and duration) classified tokens at 35.3 % accuracy for words and 31.8 % for sentences, roughly double the 16.7 % chance level ($\chi^2 > 184$, $p < 10^{-25}$). Commands and Statements were recognized best (≈ 46–54 % correct).

Confusion patterns

Command and Complaint were best recognized. Celebration and Doubt were often confused with others. There was overlap between Doubt–Suggestion–Statement.

Table 1. Acoustic Summary (words).

Speech Act	Duration (ms)	F0 Range (st)	Mean Intensity (dB)	Mean F0 (st)
Statement	646.68	84.21	59.87	172.94
Doubt	694.28	159.16	61.02	211.25
Suggestion	639.10	154.15	59.81	214.99
Command	577.41	105.12	66.22	199.30
Celebration	699.18	120.99	65.10	207.92
Complaint	830.05	99.96	62.04	171.45

Table 2. Acoustic Summary (sentences).

Speech Act	Duration (ms)	F0 Range (st)	Mean Intensity (dB)	Mean F0 (st)
Statement	1070.80	1070.80	1070.80	1070.80
Doubt	81.55	81.55	81.55	81.55
Suggestion	58.79	58.79	58.79	58.79
Command	171.02	171.02	171.02	171.02
Celebration	1164.36	1164.36	1164.36	1164.36
Complaint	159.62	159.62	159.62	159.62

Discussion

Overall, the results show that Cantonese speakers reliably produce pitch height, pitch span, loudness, and timing patterns to convey different speech acts. These global prosodic patterns are strong enough to yield machine classification at twice-chance accuracy, yet not distinctive enough to prevent specific confusions. This is the first attempt to provide a systematic acoustic map of Cantonese speech-act prosody, showing prosodic space available despite the tone system. These findings confirm that global F0, intensity, and durational cues jointly encode basic pragmatic intentions, though not yet at ceiling discriminability.

Limitations

Four global parameters only. Accuracy is modest; dynamic or spectral features may enhance classification.

Conclusions

Six speech acts are reliably differentiated by global prosody. Distinctions are not perfect but systematic. Cantonese prosody extends beyond lexical tone and provides acoustic benchmarks for further study.

Acknowledgements

This study forms part of my PhD project at The Hong Kong Polytechnic University and was supported by the PolyU Research Postgraduate Scholarship (PRPgS). We thank all participants and research assistants for their contribution.

References

Austin, J. 1962. How to Do Things with Words. Oxford University Press.

Green, M. 2000. Illocutionary Force and Semantic Content. Linguistics and Philosophy 23, 435–473.

Hellbernd, N., Sammler, D. 2016. Prosody conveys speaker's intentions: Acoustic cues for speech act perception. Journal of Memory and Language 88, 70–86.

Ma, J. K.-Y., Ciocca, V., Whitehill, T. L. 2008. Acoustic cues for the perception of intonation in Cantonese. Interspeech 2008, 520–523.

Searle, J.R. 1968. Austin on Locutionary and Illocutionary Acts. The Philosophical Review 77(4), 405–424.

Searle, J.R. 1969. Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press.

Searle, J., Vanderveken, D. 1985. Foundations of Illocutionary Logic. Cambridge University Press.

Wu, W.L. 2009. Sentence-final particles in Hong Kong Cantonese: Are they tonal or intonational? Interspeech 2009, 2291–2294.

High rising terminals in first- and second-generation Mandarin- and Anglo-background speakers in Australia

Chengjin Liu, Ksenia Gnevsheva
Australian National University, Australia

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/017/000677

Abstract

This study examines High Rising Terminals (HRTs)—rising pitch on declaratives—among Anglo-Celtic, first-generation (Gen 1), and second-generation (Gen 2) Mandarin-background women in Australia. Of 7,204 intonation units, 1,724 were HRTs, with higher use in Gen 1 (26.5%) and Gen 2 (29.7%) than in Anglo speakers (19.4%), though only Gen 2 differed significantly. Acoustic analysis showed similar rise alignment across groups, with Gen 2 having smaller (3.09 ERB) and Gen 1 larger (4.53 ERB) excursions than Anglos (3.80 ERB). These results indicate that Mandarin-background speakers use the mainstream Australian English HRT patterns but at higher rates, suggesting convergence with and potential leadership in ongoing prosodic change.

Keywords: high rising terminals, prosody, ethnolinguistic variation, acoustic analysis, Australian English.

Introduction

Ethnolects, the distinct linguistic varieties emerging through contact between heritage and dominant languages (Clyne, 2000), often show generational differences: first-generation (Gen 1) speakers retain L1 features while second-generation (Gen 2) may converge toward mainstream norms. Hoffman and Walker's (2010) interpretation of ethnolectal formation predicts that Gen 2 speech can either resemble Gen 1 (Carlock & Wölck, 1981) or show convergence to mainstream norms or innovation (Gnevsheva, 2020; Hoffman & Walker, 2010). While segmental and lexical features of ethnolects are well studied, prosody remains underexplored.

This study focuses on High Rising Terminals (HRTs)—rising pitch contours on declarative utterances—as a potential marker of ethnolinguistic group membership in Australia. HRTs are common in Australian English and other English varieties and vary in pitch alignment and excursion (Guy et al., 1986; Levon, 2020; Ritchart & Arvaniti, 2014). Socially, they occur more frequently among younger speakers, women, and certain ethnic groups (Britain, 1992; Guy et al., 1986; Levon, 2016, 2020). This study compares the frequency and phonetic realisation of HRTs across different generations of Mandarin-background speakers to explore prosodic variation within multilingual Australia.

Phonetics 2025 Hong Kong: Proceedings International Conference on Phonetic Research and Applications

Methodology

Data came from three corpora: the Sydney Speaks Corpus (Travis, 2024) for Anglo-Celtic speakers (born 1993–1998), the Second-Generation Chinese-Australian Corpus (Zhang, 2015) for Gen 2 Mandarin-background Australians (born 1994–1997, native-born or arrived before 5), and the AusESL Corpus (Gnevsheva & Travis, 2024) for Gen 1 Mandarin-background Australians (born 1984–1991, arrived after 17, lived in Australia for more than 5 years). The study analysed 24 female speakers (eight per group).

Declarative Intonation Units (IUs) were identified, and HRTs were coded via a two-stage auditory procedure (Levon, 2016): each IU was labelled "definitely HRT," "definitely not HRT," or "not sure," with uncertain cases re-coded by two linguists and disagreements resolved acoustically (pitch excursion (F0max – F0start) over 40% defined HRTs). HRT rates were analysed using logistic mixed-effects models, with participant group and centred year of birth as fixed effects, and speaker as a random intercept. Year of birth was not significant and was pruned from the model.

In addition, a random sample of ten HRT tokens per speaker was analysed in Praat to examine rise alignment and excursion. Rise alignment was segmented at the word-level, while excursion was presented in both Equivalent Rectangular Bandwidths (ERB) and percentage of excursion, enabling direct comparison with previous studies (Guy et al., 1986; Levon, 2020; Ritchart & Arvaniti, 2014).

Results

Across all groups, 1,724 HRTs were identified (23.93% of 7,204 declarative IUs). Anglo-Celtic speakers showed the lowest usage (19.36%, *SD = 7.16*), while Gen 2 Mandarin-background speakers showed the highest (29.66%, *SD = 12.24*), significantly higher than Anglos ($\beta = 0.557$, $SE = 0.271$, $z = 2.06$, $p = .040$). Gen 1 speakers (26.49%, *SD = 13.55*) also exceeded Anglos, but not significantly ($\beta = -0.444$, $SE = 0.235$, $z = -1.893$, $p = 0.058$) (Figure 1). The greater variability among Gen 1 speakers likely reflects differences in exposure and interactional experience. Acoustic analysis showed similar rise alignment across groups, with rises beginning on the nuclear syllable—typically the final stressed syllable—regardless of ethnicity. However, rise excursion differed: Gen 1 speakers had the largest mean excursion (4.53 ERB; 136%), followed by Anglos (3.80 ERB; 84%) and Gen 2 speakers (3.09 ERB; 82%).

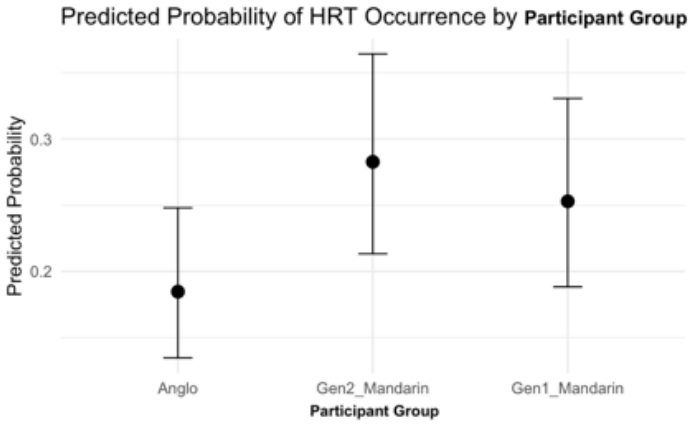


Figure 1. Predicted Probability of HRT Occurrence by Participant Group Based on a Binomial Logistic Mixed-Effects Model.

Discussion

Our findings show a rise in HRT use among Anglo-Celtic Australians compared to previous studies (19.36% vs. 1% in Guy et al., 1986), indicating ongoing language change. Anglo speakers typically anchored rises to the nuclear syllable, aligning with Guy et al. (1986), and showed larger mean excursions (3.80 ERB; 84%) than reported for other English varieties (1.25–1.57 ERB) (Levon, 2020; Ritchart & Arvaniti, 2014). Overall, results indicate increasingly dynamic and distinctive HRT patterns in contemporary Australian English.

Gen 2 speakers used HRTs significantly more than Anglos, with similar rise alignment and slightly smaller excursions (3.09 vs. 3.80 ERB), suggesting mainstream-like phonetic patterns coupled with higher rates of use (cf. Hoffman and Walker, 2010). Unlike ethnic minorities in Britain, whose HRTs remain low in frequency (Levon, 2016, 2020), Gen 2 speakers in Australia appear to adopt and extend the feature, contributing to socioprosodic innovation in Australian English. Gen 1 speakers showed comparable rates and slightly larger excursions (4.53 ERB), suggesting successful adaptation to local patterns, with limited evidence of Mandarin transfer. Overall, Gen 1 speakers have assimilated to mainstream norms, while Gen 2 speakers expand HRT use as a socially meaningful resource, illustrating that HRTs among Mandarin-background Australians reflect evolving ethnolectal variation rather than heritage transfer.

Acknowledgements

We thank the Phonetics 2025 HK staff and reviewers for their support and feedback. Our gratitude also goes to Catherine Travis, Rosey Billington, Anton Malko, and Evan Kidd for their guidance, and to Yuqing He and Josie Grundy for their help with reliability

testing and valuable feedback on the drafts. Special thanks to Gerry Docherty for conducting the ASR processing.

References

Britain, D. 1992. Linguistic change in intonation: The use of high rising terminals in New Zealand English. Language Variation and Change, 4(1), 77–104.

Carlock, E., Wölck, W. 1981. A method for isolating diagnostic linguistic variables: The Buffalo ethnolects experiment. Variation Omnibus. Edmonton, Alberta: Linguistic Research (Current Inquiry Into Language, Linguistics and Human Communication, 40, 17–24.

Clyne, M. 2000. Lingua Franca and Ethnolects in Europe and Beyond. Sociolinguistica, 14(1).

Gnevsheva, K. 2020. The role of style in the ethnolect: Style-shifting in the use of ethnolectal features in first- and second-generation speakers. International Journal of Bilingualism, 24(4), 861–880.

Gnevsheva, K., Travis, C. 2024. Corpus of Australian English as a Second Language (AusESL).

Guy, G., Horvath, B., Vonwiller, J., Daisley, E., Rogers, I. 1986. An intonational change in progress in Australian English. Language in Society, 15(1), 23–51.

Hoffman, M. F., Walker, J. A. 2010. Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. Language Variation and Change, 22(1), 37–67.

Levon, E. 2016. Gender, interaction and intonational variation: The discourse functions of High Rising Terminals in London. Journal of Sociolinguistics, 20(2), 133–163.

Levon, E. 2020. Same difference: The phonetic shape of High Rising Terminals in London. English Language & Linguistics, 24(1), 49–73.

Ritchart, A., Arvaniti, A. 2014. The use of High Rise Terminals in Southern Californian English. Proceedings of Meetings on Acoustics, 20.

Travis, C. 2024. Sydney Speaks corpus: An overview. Australian Journal of Linguistics, 0(0), 1–19.

Zhang, S. 2015. Multiple voices under one name: Ethnic Orientation and heritage language in second generation Chinese-Australians. [Master Thesis]. The Australian National University.

Cross-dialectal perspective on the form and meaning relation: the case of Tone 3 sandhi

Yuxin Lu, Yu-Hsiang Tseng, R. Harald Baayen
University of Tübingen, Germany

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/018/000678

Abstract

This study investigates the tonal realization of Tone 3 sandhi in spontaneous speech of Beijing and Taiwan Mandarin, focusing on whether it is influenced by word-specific and semantic factors. Using corpus data and Generalized Additive Mixed Models, we found complete neutralization between T2–T3 and T3–T3 tone patterns in both varieties. Nevertheless, words exhibited distinct "pitch signatures," with some showing similar pitch contours across varieties and others notable differences. Computational modeling using the Discriminative Lexicon Model further revealed that tonal contours can be predicted from contextual meaning with above-chance accuracy. These results suggest that tone realization is shaped not only by phonological structure but also by word-specific semantics and collocational preferences.

Keywords: tone 3 sandhi, GAMMs, form-meaning relation, word-specific tonal realization.

Introduction

Tone 3 sandhi is a phonological process that a Tone 3 becomes Tone 2 when it is followed by another Tone 3. In a recent study by Lu et al. (2025), disyllables with T3-T3 tone patterns were found to be completely neutralized with T2-T3 in spontaneous speech of Taiwan Mandarin. More importantly, words' meaning emerged as an important predictor of tonal realizations. The finding that semantics co-determines pitch realizations has also been found for two-syllable words with other tone patterns in Taiwan Mandarin (Chuang et al., 2025; Lu et al., 2025b). However, little is known about this in other varieties of Mandarin beyond Taiwan Mandarin.

The current study investigates how Tone 3 sandhi is realized in spontaneous speech, in both Beijing Mandarin and Taiwan Mandarin. Three main questions are addressed. First, does word-specific tonal realization also occur in Beijing Mandarin? Second, does Tone 3 sandhi also show complete neutralization in spoken Beijing Mandarin, as in Taiwan Mandarin? Third, if so, can the observed pitch contours be predicted from their meaning in context with above-chance accuracy?

Phonetics 2025 Hong Kong: Proceedings International Conference on Phonetic Research and Applications

Data

The data of Beijing Mandarin come from the Beijing Corpus (Ruan et al., 2018), and the data of Taiwan Mandarin come from the Taiwan Mandarin Spontaneous Speech Corpus (Fon, 2004). Both datasets consist of naturally occurring conversational speech in unstructured interviews. F0 value was estimated over the entire syllables since the focus of the current study is on how pitch is realized on whole onomasiological words. Subsequently, we selected tokens with T2-T3 and T3-T3 tone patterns from both corpora (see Table 1). For more details on data selection, see Lu et al. (2025).

Table 1. Overview of tokens and word types in two datasets.

Place	Tone pattern	Tokens	Word types
Beijing	T2-T3	1291	23
Beijing	T3-T3	1096	34
Taiwan	T2-T3	1828	30
Taiwan	T3-T3	1239	33

Tonal realizations

We made use of Generalized Additive Mixed Models (Wood, 2017) to model pitch contours as a function of *normalized time, tone pattern, neighouring tones, gender, duration, word position, speaker,* and lastly, *word*. Two separate models were fitted to the data of Beijing and Taiwan Mandarin respectively. The GAMM analysis shows that there was no clear evidence for the difference in pitch contours between T2-T3 and T3-T3 for neither Beijing nor Taiwan Mandarin. Consistent with Lu et al. (2025), this suggests that T3-T3 is completely neutralized in both spoken Beijing and Taiwan Mandarin.

However, when we did not control for the effect of *word*, the tone sandhi in our data also appeared incomplete. This highlights the importance of word-level effects, namely, the words with the same tone patterns had their own "pitch signature". As shown in Figure 1, some words showed similar pitch contours across varieties, while others exhibited notable differences. These word-specific components were driven by word's semantics and collocational preferences. For example, *还有* (hai2you3, 'still have') was typically preceded by PAUSE and *然后* (*ran2hou4*, 'then') in both varieties. However, *可以* (*ke3yi3*, 'can'), which shows different pitch contours especially in the beginning of the syllable, was mostly preceded by PAUSE, *也* (*ye3*, 'also'), and *你* (*ni3*, 'you') in Beijing Mandarin, but mostly by PAUSE in Taiwan Mandarin.

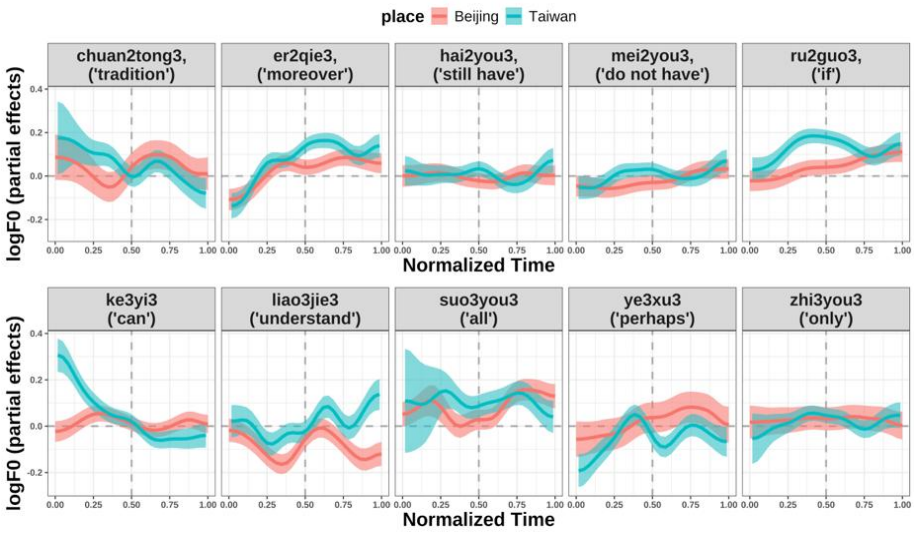


Figure 1. A selection of predicted pitch contours of words, estimated by factor smooth for word. In every panel, pitch contours are color-coded by place.

Computational modelling

To deepen our understanding of how pitch contours may be driven by semantics, we obtained semantic vectors presented by 896-dimensional contextualized embeddings from the Qwen-2.5 large language model. Furthermore, we made use of the Discriminative Lexicon Model (DLM, Heitmeier et al., 2025), a computational model focusing on the relationship between form and meaning. DLM was able to predict the pitch contours of word tokens from their meanings with the mean accuracy of 29.69% for Beijing data and 29.21% for Taiwan data (premuation baseline 10.30% and 10.44%). However, when we permuted the semantic vectors between Beijing and Taiwan Mandarin, the model's mean prediction accuracy dropped to 26.65% and 26.71% respectively.

Conclusion

The present study examined the pitch contours of T2–T3 and T3–T3 sequences in both Beijing and Taiwan Mandarin, and further investigated whether tonal contours could be predicted from words' meanings in context. First, we successfully replicated the findings of complete neutralization and word-specific tonal realizations in Taiwan Mandarin, originally reported by Lu et al. (2025a), using a slightly different dataset. Second, we extended these findings to Beijing Mandarin, showing that the word effect is robust across varieties. Third, we demonstrated that computational modelling can predict tonal contours from contextual

meaning with above-chance accuracy, suggesting a complex and nuanced interaction between tonal realization and semantics.

Note

1. The Hanyu Pinyin annotation of disyllables undergoing tone sandhi still follows the citation tones of each syllable in isolation for the ease of understanding, for example, 有点 is annotated as *you3dian3*.

References

Chuang, Y.-Y., Bell, M. J., Tseng, Y.-H., Baayen, R. H. 2025. Word-specific tonal realizations in Mandarin. Accepted for publication in Language. arXiv:2405.07006v2.

Heitmeier, M., Chuang, Y.-Y., Baayen, R. H. 2025. The Discriminative Lexicon: Theory and implementation in the Julia package JudiLing. Cambridge University Press. in press.

Lu, Y., Chuang, Y.-Y., Baayen, R. H. 2025. Form and meaning co-determine the realization of tone in Taiwan Mandarin spontaneous speech: the case of Tone 3 sandhi. arXiv preprint arXiv:2408.15747. Under Revision for Journal of Chinese Linguistics.

Lu, Y., Chuang, Y.-Y., Baayen, R. H. 2025. The realization of tones in spontaneous spoken Taiwan Mandarin: a corpus-based survey and theory-driven computational modeling. arXiv preprint arXiv:2503.23163. Accepted for publication in Corpus Linguistics and Linguistic Theory.

Ruan, F., Song, Q., Li, K., Hao, Y. 2018. Definition of corpus, scripts, standards and specifications of environment/speaker coverage for Mandarin languages. Technical report, Beijing Haitian Ruisheng Science Technology Ltd.

Wood, S. N. 2017. Generalized additive models: an introduction with R. CRC press.

Sociophonetic perception of Dh-Stopping in South Yorkshire English

Bartolomé Díaz Martínez
University of Murcia, Spain

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/019/000679

Abstract

This study investigates covert linguistic attitudes towards the phenomenon known as dh-stopping ([ð]→[d]) in Sheffield English. This variant is especially salient in the pronunciation of Old English second person pronouns (thee, thou) but also appears in function words. We explore how different sociodemographic groups perceive this variation in terms of solidarity, accent, rurality, and perceived age. 111 participants from Sheffield and surrounding areas completed a matched-guise experiment where they rated speakers on several scales based on their pronunciation of the dental fricative [ð] and dental stop [d] variants. Results show that gender, age, and geographic area strongly influenced the perceptions of these variants, and the complex social meanings attached to phonetic variants in Sheffield English.

Keywords: dh-stopping; linguistic attitudes; Sheffield English; solidarity; rurality

Introduction and background

The dh-stopping phenomenon, a local variant of the interdental fricative [ð] realized as a dental stop [d], has sociolinguistic relevance in South Yorkshire English. This phenomenon is closely associated with the way Sheffielders pronounce Old English second person address terms *thee* and *thou*, although this realisation also occurs in function words other than pronouns. The nickname *Dee-Dah's* is still used to describe people from Sheffield by those from nearby places in a pejorative way. On the basis that this feature is noted as being more typical for older males in the area, there does not seem to be a huge amount of evidence that it is in current usage. This study explores attitudes towards this variant focusing on solidarity, accent, acceptance, and rural perceptions among speakers of different gender, age, and towns within South Yorkshire.

Objectives

The main objective of this project was to investigate the covert social attitudes associated with the production of the British consonant /ð/ as either [ð] or [d], focusing on how these attitudes vary according to linguistic context and listener background. Specifically, the study aimed to determine (1) the covert attitudes linked to dh-stopping in second-person relic pronouns (pronominal level), (2) the attitudes associated with dh-stopping in function words other than pronouns

(phonetic level), and (3) whether local speakers from Sheffield perceive and evaluate dh-stopping differently from non-locals in South Yorkshire.

Methodology

Participants from Sheffield and nearby cities completed a matched-guise perception experiment evaluating audio samples with both the standard and dh-stopping variants. This was achieved by 'splicing' audio segments containing specific phonetic variables. The study compared listeners' reactions to instances of *dh*-stopping when it appeared in Old English pronouns (e.g., *thee*, *thou*) versus when it occurred in other function words not derived from Old English pronouns. Responses on solidarity (friendliness and pleasantness), accent acceptance (attractiveness and refinement), and rurality were collected and analysed.

Results and discussion

Boxplots and bar charts illustrate key differences. Regarding the first research question, women tend to attribute lower solidarity to the dh-stopping [d] variant. Also, older participants perceive the dh-stopping variant as more rural and are more tolerant of its usage.

Figure 1 below shows the distribution of solidarity scores attributed to the dh-stopping variant, broken down by gender. Female participants tend to attribute lower solidarity to the use of [d]; whereas Figure 2 represents rurality ratings by age group. Older participants show greater tolerance towards rural associations of the dh-stopping variant.

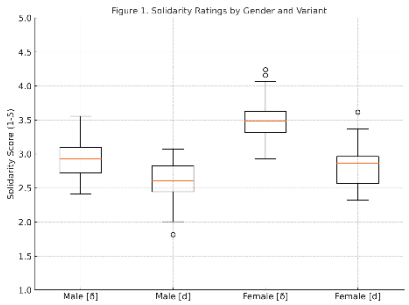


Figure 1. Solidarity scores by gender.

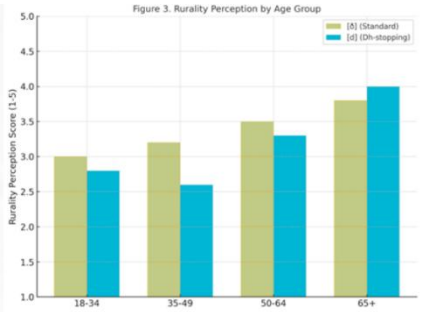


Figure 2. Rurality ratings by age any difference.

As concerns our second research question, whether it is possible to determine any difference between the perceptions of locals (Sheffield) compared to non-locals from other areas of South Yorkshire, Sheffield participants rate the [d]

variant more favourably than those from other areas nearby, as portrayed in Figure 3.

The findings suggest that dh-stopping functions as a sociolinguistic marker influenced by gender, region, and age. Female speakers attribute less solidarity to the variant what suggest that this variant may be socially stigmatized, especially in more formal or public contexts. However, the more positive ratings from older participants and Sheffield residents suggest that dh-stopping retains covert prestige within certain social groups, which implies that regional pride in Sheffield promotes more positive acceptance. Age influences rural perception, reflecting social attitudes towards language variation and identity. Certainly, the results highlight the complex social meanings attached to phonetic variants in Sheffield English.

ANOVA tests confirm these differences as statistically significant. As shown in Figure 4, at the pronominal level, the *t*-test gives a value of $t = -1.785986$ and $p = 0.11192$. This result is not significant at $p < 0.05$, meaning that differences in evaluations of relic pronouns could be due to chance, while at the phonetic level, the *t*-test yields $t = -3.739694$ and $p = 0.00571$. This result is significant at $p < 0.05$, indicating that the observed differences for function words other than pronouns are statistically meaningful.

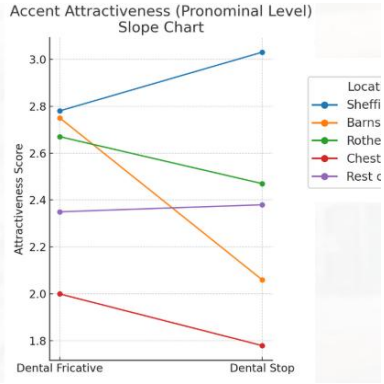


Figure 3. Accent attractiveness by area.

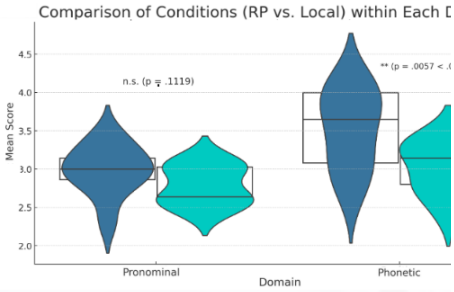


Figure 4. Phonetic vs. pronominal domains.

Conclusion

Overall, the elicited reactions among the inhabitants of Sheffield and South Yorkshire mainly suggest that the Dee-Dah phenomenon has followed a natural evolution, not so much towards a marker or indicator within Sheffield but

towards a Labovian stereotype, circulating by word of mouth among the population of Sheffield and surrounding areas. All in all, it is manifest that the sociolinguistic significance of the dh-stopping variant goes beyond mere phonetic variation, reflecting broader social identities and group memberships in South Yorkshire. These insights contribute to our understanding of sociophonetic variation and its relationship with identity in urban dialects.

References

Cooper, P. 2013. Enregisterment in Historical Contexts: A Framework. Unpublished Ph.D. thesis. University of Sheffield

Eckert, P. 2008. Variation and the indexical field. Journal of Sociolinguistics, 12(4), 453-476.

Labov, W. 1972. Sociolinguistic Patterns. Philadelphia: University of Pennsylvania Press.

Montgomery, C. 2007. An introduction to language and society. Routledge.

Preston, D.R. 1999. A sociolinguistic perspective on dialect perception. In Chambers, J.K., Trudgill, P., Schilling-Estes, N. (Eds.), The Handbook of Language Variation and Change (pp. 559-581). Oxford: Blackwell.

Silverstein, M. 2003. Indexical order and the dialectics of sociolinguistic life. Language & Communication, 23(3-4), 193-229.

Trudgill, P. 1983. On Dialect: Social and Geographical Perspectives. Oxford: Blackwell.

Nasal consonants in Malayalam

Caterine Michael, Reenu Punnoose
[1]Indian Institute of Technology Palakkad, India

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/020/000680

Abstract

Almost all the world's languages have at least one voiced nasal in their phonemic repertoire. Having more than three or four nasal consonants in a language, however, is uncommon. One such exception is Malayalam, a Dravidian language spoken predominantly in the southern state of Kerala in India. Malayalam has six or arguably, seven nasal phones in its inventory. In addition, the nasals in Malayalam are also characterized by singleton-geminate contrasts in the same intervocalic-medial position. Preliminary auditory and acoustic observations suggest that durational differences are important in distinguishing the geminate and singleton contrast. In addition, the nasals that form contrasting subgroups appear to be predominantly distinguished based on secondary articulation cues.

Keywords: nasals, Malayalam, duration, secondary articulation.

Introduction

Malayalam, a language of the Dravidian family predominantly spoken in Kerala, has a rich inventory of nasal consonants. While most scholars who have worked on Malayalam phonology (Chandrasekhar 1953, Krishnamurti 2003) agree that there are six nasal phones- bilabial /m/, dental [n̪], alveolar [n], retroflex /ɳ/, palatal /ɲ/ and velar /ŋ/, other scholars (Mohanan & Mohanan 1984, Namboothiripad, Garellek 2017, Khan 2019) identify a seventh place of articulation. The labels that appear in grammatical sketches about nasals in the language are mainly descriptive and are not based on an empirical analysis of all the nasals in the inventory.

Distribution of nasals in Malayalam

The scholars working on Malayalam phonology agree on the phonotactics of the bilabial, retroflex and velar nasals. The bilabial nasal occurs in all the word positions as singletons and intervocalically as geminates. The retroflex nasal occurs medially in intervocalic singletons and geminates; and in a nasal-plosive cluster. Palatal nasals are attested in Malayalam word initially as singletons and medially in a homorganic nasal-plosive cluster and in an intervocalic position.
Interestingly, the relation between the dental and alveolar nasals have been debated in the literature. While many scholars (Chandrasekhar 1953)opine that the dental-alveolar relationship is allophonic, other scholars (Mohanan, Mohanan 1984, Asher, Kumari 1997) suggest that along with the complementary

distribution, the relationship is also phonemic, where they are contrastive intervocalically as geminates.

Debates also exist with the seventh nasal, the palatovelar, whether it contrasts intervocalically as geminates (Mohanan, Monahan 1984, Namboodiripad, Garellek 2017) or are allophones of the velar nasal (Asher, Kumari 1997).

In addition to the rich nasal inventory in Malayalam, nasal geminates are attested intervocalically in all the places of articulation. The geminates have a phonemic relation with the singletons in the same word position (Local, Simpson 1999). An instance of this is in the minimal pair *kaŋ:i* 'link' and *kaŋi* 'first vision.'

Given the distribution of Malayalam nasals, it is important to note that the seven-way contrast is not present across all the vowel contexts in Malayalam. Excluding the palatovelar nasal, the six-way contrast is only seen in one instance-(a_i). In other vowel contexts, they form subgroups of two/three contrastive pairs.

Methodology

Data was collected from 12 native Malayalam speakers (6 male and 6 female) from the districts of Calicut (north), Trivandrum (south) and Kottayam (south-central) in Kerala within the age group of 55 to 65 years.

A word list of 53 tokens, consisting of intervocalic nasal geminates and singletons, were displayed on the screen. Recordings were done using a Zoom H1n Handy Recorder and an Audio-Technica ATR3350xiS clip-on microphone.

The speech samples were annotated using PRAAT software. The statistical analysis was done using the R (R Core Team) lme4 package. Anova and pairwise analysis (emmeans) was carried out.

Results

The six-way contrast in the data set is confined to the a_i environment, and the acoustic analysis suggests that for geminate-singleton pairs duration is key; and for different nasal places of articulation, the second formant frequency (F2) of the preceding vowel appears to be instrumental in maintaining the contrasts.

Duration

Durational differences have been significant in maintaining the geminate-singleton contrast across various studies (Local & Simpson 1999). In this study, a similar pattern can be observed, where the duration of the geminates is more than twice the length as compared to the contrastive singleton counterparts in the a_i environment.

The alveolar and retroflex singletons were significantly shorter as compared to the geminates: Alveolar (estimate =-102.750, SE =9.091 t value=-11.302, p= <.001), Retroflex (estimate=-115.917, SE= 9.091, t value= -12.750, p= <.001).

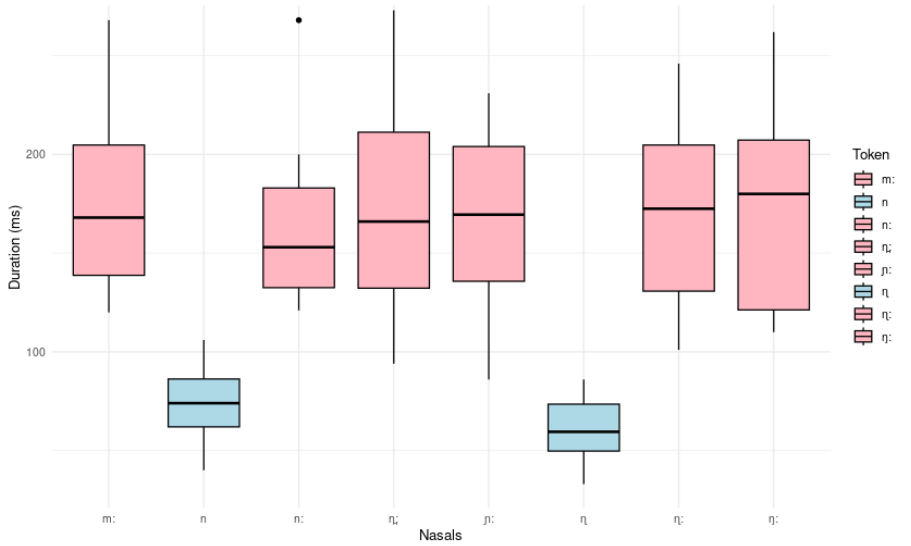


Figure 1. Duration of geminate and singleton nasals in a_i context.

Generally, durational differences among the nasal consonants, from bilabial to retroflex, go on a continuum from longer to shorter (Local, Simpson 1999). However, in the a_i context, that showed maximum nasal place contrast among the geminates, there was no significant pattern.

Formant frequencies

The adjacent vowels and their transition to and from the consonants have been shown to be important cues in identifying place differences among nasals (Khan 2019), and F2, that correlates to the frontness and backness of a vowel, particularly serves as an acoustic cue (Khan 2019). Therefore, the difference in F2 of the preceding and/or following vowel suggests coarticulatory and secondary articulation effects. In the current study, though the effects of the following vowel did not show any statistical significance, the effects of the preceding vowel showed significance. A pairwise comparison showed some pairs to be statistically significant:

Table 1. Statistically significant pairs of nasals in preceding vowel context.

Nasal	Estimate	SE	t-value	p
Alveolar-dental geminates	335.3	83.6	4.011	<0.05
Alveolar-bilabial geminates	-465.6	83.6	-5.572	<0.001
Palatal-dental geminates	-320.6	72.4	-4.428	<0.05
Alveolar-retroflex Singleton	281.7	59.1	4.764	<0.05

In each pair, the F2 value of the preceding vowel of both alveolar and palatal nasals (alveolar/palatal) was higher indicating frontness (palatalization); while the F2 value of bilabial, dental and retroflex nasals was lower indicating backness (velarization). Thus, in addition to the patterns of the alveolar-dental pair, as seen in Khan (2019), the nasals can be grouped into pairs and appear to maintain the distinction by using secondary articulation (palatalization and velarization) as cues.

Conclusion

Nasal duration is therefore significant in understanding the geminate-singleton difference among the speakers in Malayalam. The F2 of the preceding vowel was higher for some nasals when looking at them in terms of pairs, indicating secondary articulation as a potential cue. Analysis of other acoustic properties and a larger sample is required to understand more about how the nasals in Malayalam are maintained.

References

Asher, R.E., Kumari, T.C. 1997. Malayalam. London: Routledge.

Chandrasekhar, A. 1953. Evolution of Malayalam. Poona: Deccan College.

Khan, S. 2019. Palatalization and Velarization in Malayalam nasals: a preliminary acoustic study of the dental–alveolar contrast. In Proceedings of FASAL, 9. Reed College

Krishnamurti, B. 2003. The Dravidian Languages. Cambridge: Cambridge University Press.

Local, J., Simpson, A. P. 1999. Phonetic implementation of geminates in Malayalam nouns. In Proceedings of the 14th ICPHS, San Francisco, 1059-1062

Mohanan, K.P., Mohanan, T. 1984. Lexical phonology of the Consonant System in Malayalam. Linguistic Inquiry, 15 (4), 575-602.

Namboodiripad, S., Garellek, M. 2017. Malayalam (namboodiri dialect). Journal of the International Phonetic Association, 47(1) 109-118.

A gestural approach to Latin /pl, fl, kl/ cluster realizations in Galego-Portuguese and Sardinian

Benjamin Schmeiser
Illinois State University, US

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/021/000681

Abstract

Among Romance languages, Galego-Portuguese and Sardinian are the only languages that replace the lateral with a rhotic from a Latin /pl, fl, kl/ onset cluster, as in (1):
(1) Latin: ECCLESIA English: 'church'
Portuguese: igreja [i.ˈgRe.Za]/Galician: igrexa [i.ˈgRe.Sa]
Sardinian: crexia [ˈkRe.Sia]
The current study greatly adds to the field by addressing this unique environment in Romance languages in two distinct ways: first, it offers a historical analysis, which includes treating apparent exceptions. Second, it discusses the phenomenon using both a gestural approach (i.e. Articulatory Phonology) and Phase Windows for both languages.

Keywords: Galego-Portuguese, Sardinian, consonant clusters, articulatory gestures

Introduction

Romance languages are a group of languages that have evolved from (Vulgar) Latin. There are *at least* fourteen, and they are often classified by the five 'major' languages (Spanish, Italian, French, Portuguese, and Romanian), and the remaining as 'regional' languages. In (2) below:

(2) Romance languages

(a) Spanish	(b) Italian
(c) French	(d) Portuguese
(e) Romanian	(f) Occitan
(g) Provençal	(h) Gallego
(i) Romansch	(j) Catalan
(k) Sardinian	(l) Navarro-Aragonese
(m) Astur-Leonese	(n) Neopolitan

The reader will note that, for the purpose of the current discussion, Gallego (also referred to as 'Galician') and Portuguese are combined here ('Galego-Portuguese'), given that they are from the same family (Galician-Portuguese) historically and evolved similarly for the phenomenon discussed here. They later evolved into separate languages by the 14th century.

Phonetics 2025 Hong Kong: Proceedings International Conference on Phonetic Research and Applications

Sardinian is indeed its own language and not a dialect of Italian. Sardinian is the language spoken on the island of Sardinia, which is Italy's second largest island (Sicily being the largest). Sardinian is known for, among other things, its conservative evolution from Latin.

Word-Initial complex onsets in Latin

The following table shows the obstruent + liquid sequences in Latin for the word-initial environment. On the left side, one observes that Latin had three voiceless stops, three voiceless stops, and the voiceless labiodental fricative. More relevant to our discussion here is the right side, which contains the possible obstruent + lateral sequences; note there are no examples of a dental stop, followed by a lateral. In addition, given the space constraints of here, the current study is confined to the possible voiceless obstruents (/pl/, /kl/, and /fl/) as shaded in blue. Rhotacization does indeed occur in /bl/ sequences in Galego-Portuguese (e.g. branco from Late Latin blancus 'white'), however I did not find any examples of this sequence in Sardinian. The current study then, is devoted to the voiceless environment; future research should consider the /bl/ environment in more detail.

Table 1. Word-initial Latin complex onsets.

For /r/-:			For /l/-:		
Obstruent	Example	English gloss	Obstruent	Example	English gloss
/b/	brācchĭu	'arm'	/b/	blandu	'bland'
/d/	drăco	'dragon'	/d/	-	
/g/	grŏssu	'large'	/g/	glūten	'glue'
/p/	prātu(m)	'meadow'	**/p/**	**plānu**	**'flat'**
/t/	trahĕre	'to bring'	/t/	-	-
/k/	crēdĕre	'to believe'	**/k/**	**clamāre**	**'to call'**
/f/	frŏnte	'front'	**/f/**	**flamma**	**'flame'**

The topic merits further study because rhotacization in obstruent + lateral word-initial clusters is unique to Galego-Portuguese and Sardinian. For example, notice that for a word like Latin *ecclesia* 'church' e[kl]ésia, it evolves into Catalan: es[gl]ésia, French: é[gl]ise, Italian: [kj]esa, but crucially Portuguese: i[gR]eja/ Galician: i[gR]exa and Sardinian: [kR]esia

Rhotacization in Galego-Portuguese and Sardinian

In what follows, we observe in (3) how word-initial /pl/, /fl/, and /kl/ clusters undergo rhotacization during the evolution from Latin to Galego-Portuguese and Sardinian:

(3) Galego-Portuguese

a. Latin [pl]- PLATTUS → b. [pR]ato 'plate'
c. Latin [fl]- FLACCUS → d. [fR]aco 'weak'
e. Latin [kl] ECCLESIA → f. i[gR]eja / i[gR]exa 'church'

Sardinian

g. Latin [pl] PLUS → h. [pR]us 'more'
i. Latin [fl] FLAMMA → j. j. [fj]amma /[fR]ada 'flame'
k. Latin [kl] ECCLESIA → l. [kR]esia 'church'

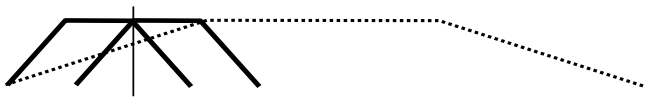
In terms of word transmission, the above word sample, though admittedly small, might suggest that a tendency is for popular and semi-learned words to undergo rhotacization, whereas learned words would likely not undergo this process.

Articulatory phonology

Articulatory Phonology (Browman and Goldstein 1992) is an approach rooted in basic units called gestures (see below) that represent the smallest unit of phonological representation. Major articulators produce constrictions in the vocal tract, varying in their constriction degree and exact location to form a gesture. In the case of consonant clusters, each consonantal gesture has a timing relationship with the other consonantal gesture and also with the underlying vowel (Gafos 2002).

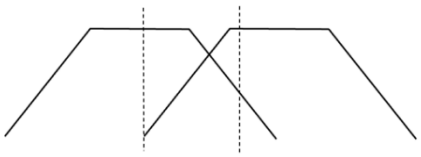
Lexical representation specifies the requisite gestures and specifies which gestures are to be coordinated. Coordination between associated gestures is assumed to be variable but constrained to particular ranges specific to the types of gestures involved (e.g. C to C); it acts to limit the temporal compressibility or disassociation of gestures, which is useful for capturing the timing variability observed in the coordination of these gestures.

With specific regard to Galego-Portuguese and Sardinian, the Phase Window (Byrd 1994, 1996b) limited the cluster in terms of gestural overlap of the two consonant gestures. For the current analysis, the two gestures overlap within the Phase Window, however the gesture representing the lateral was realized with more overlap. For the current analysis, the two gestures overlap within the Phase Window, however the gesture representing the lateral was realized with more overlap. In Gafos's (2002) alignment terms, C2's target aligns with C1's release, as in:



Less distance between the two consonants is allowed because the lateral is markedly different in its articulatory realization: laterals involve an additional gesture whereby the sides of the tongue are lowered as to allow for lateral air release.

Taps lack this additional gesture and are shorter in duration, therefore, diachronically when a lateral loses the additional gesture, decreases in duration, and moves further away from the C1 as noted by the presence of an intrusive vowel, it is reinterpreted as a rhotic tap. That is, in Galego-Portuguese and Sardinian, there was a timing restriction in the Phase Window in which the C2's onset aligns with C1's c-center, as in:



Conclusions

The current study considered how the word was brought into the language by illustrating modifications in gestural timing and their consequent effect on the evolution of popular and semi-learned words in both languages.

References

Browman, C., Goldstein, L. 1992. Articulatory phonology: an overview. Phonetica 49, 155-180.

Byrd, D. 1994. Articulatory timing in english consonant sequence. Ph.D. Dissertation. University of California, Los Angeles, CA.

Byrd, D. 1996b. A phase window framework for articulatory timing. Phonology 13, 139-169.

Gafos, A. 2002. A grammar of gestural coordination. Natural Language and Linguistic Theory 20.2, 269-337.

From belief to behavior: exploring what language research methods truly measure

Aleksandra Siemieniuk
VIZJA University, Poland, The Maria Grzegorzewska University, Poland

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/022/000682

Abstract

Research on language use often relies on self-report questionnaires. While efficient, this method captures beliefs about language rather than the behavior itself. This paper compares three methodological approaches: self-report questionnaires, an elicited-production task based on structured scenarios, and a semi-structured conversational task. Using examples from a study on self-presentation in Polish, I illustrate the type of data each method provides—from trait-level ratings to concrete utterances and co-constructed interaction. I argue that combining these approaches offers a more complete account of language use than any single method alone, providing a general framework for psycholinguistic research.

Keywords: methodology, self-report, elicited production, conversational analysis

Introduction

Many aspects of language use are contextual and only partly conscious. Any empirical study must therefore choose what kind of data will represent the phenomenon.

In psychology and psycholinguistics, the default is self-report questionnaires, in which participants rate statements on a scale to produce trait-like scores. In linguistic research, researchers often use natural conversations, which provide rich detail but less control. Between these are elicited-production tasks, where participants produce language in response to controlled stimuli.

This paper takes a methodological perspective. It compares self-report, elicited production, and conversational data, asking: (i) what questions can each method answer? (ii) what are their main limitations? and (iii) whether we can use them interchangeably? I will use examples from my study on self-directed humour in Poland to illustrate the above-mentioned methods; the conclusions are general for linguistics researchers.

Methodology

Self-report questionnaires

Self-report instruments conceptualize language use as a subjective disposition. They are efficient for large samples and well-suited to questions such as "How do self-perceived communication styles relate to personality?" However, they

operate at a high level of abstraction. Respondents generalize across many episodes ("I often…"), which makes the method vulnerable to well-documented biases (Paulhus, 2002): (i) social desirability, (ii) recall bias, and (iii) limited meta-pragmatic awareness. Self-report provides information about self-construal, but almost no direct evidence of how language is actually used.

Elicited Production from Structured Scenarios

A second family of methods elicits concrete utterances in response to controlled situations. Scenarios can be verbal or visual; in my research, I designed cartoons showing everyday mishaps, but the schema is similar for purely textual vignettes (see Figure 1).

Such tasks are placed somewhere between abstraction and naturalism (Gernsbacher & Foertsch, 1999). Contextual control is higher than in spontaneous data, because all participants face the same scenario. At the same time, behavioural specificity is greater than in self-report, because responses are actual examples of language use.

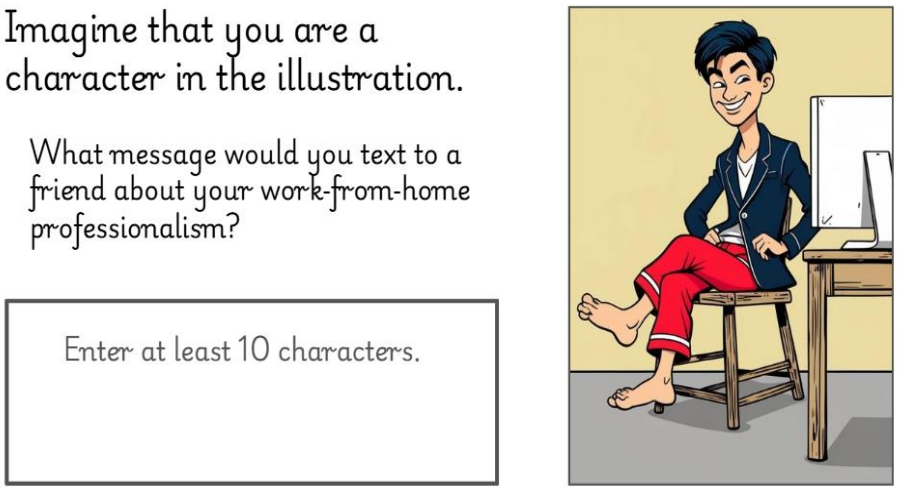


Figure 1. Example of a structured scenario used in the elicited-production task.

For example, in the scenario shown in Figure 1, one participant said, "Business on top, party on the bottom!" Another commented, "I am so bad at being professional when working from home." Although both scored similarly on a self-defeating humour scale, their comments reveal different emotional tones. These differences are only visible through the actual language production. The scenario-based elicitation method allowed me to explore how interlocutors express self-evaluation both lexically and syntactically (e.g., through metaphors and hyperbole). However, we must be aware of limited assumptions about

language use drawn from this data; participants imagine scenarios rather than experiencing them.

Conversational Methods

Conversational methods are based on the foundation that human communication is naturally interactive (Garrod & Pickering, 2007). Researchers might observe these interactions in natural settings during informal conversations or encourage participants to discuss specific topics. In my study, they openly shared their experiences on predetermined topics. Then, I reviewed transcripts and video recordings to identify gestures, facial expressions, and paralinguistic cues. All of them are crucial to analyse non-literal speech (Poggi & Vincze, 2018). Their role involves accurately interpreting the meaning of utterances.

It is illustrated in the following excerpt from the conversation on the topic: "Choose NFZ – because your health deserves the best!" (NFZ, the National Health Fund, is known in Poland for long waiting times and questionable quality). Let us examine the transcript by focusing solely on the verbal data.

> Speaker 1 [SPK_1]: I generally have an outstanding experience with the NFZ, and somehow it amuses me now, but on the other hand, it can only make me despair. I don't remember why, but I completely lost feeling in my hands, like it just cut off.
> Speaker 2 [SPK_2]: In your hands? Only?
> SPK_1: Yes, in my hands. […] So we went to the hospital... we waited, I think, two hours... Doctor saw me. He did something like this and said I was just pretending and that I should leave...
> SPK_2: So I'm dying here, I mean.
> SPK_1: Yes... he did something like that and said, "Well, listen, madam. Generally, your daughter is pretending, so..."

First, SPK_1 sets a complex emotional frame with resignation and a smile, which quickly gives way to a frown as she states, "It can only make me despair." Her facial expressions reflect frustration. She uses a pinch gesture to illustrate her symptoms, making her account more vivid and believable.

A participant in the study emphasizes the doctor's dismissive gesture, in which he states she is 'pretending.' This physical cue shows disbelief and invalidates her experience. Conclusion, 'Generally, your daughter is pretending,' with a detached tone, confirms this invalidation. The contrast between SPK_1's distress and the doctor's dismissive behaviour forms the story's core conflict.

The listener actively co-constructs the narration. Question, "In your hands? Only?", with a leaned-forward posture and focused gaze, signals engagement. When SPK_1 says, "So I'm dying here, I mean," their facial expression shifts to frustration.

This approach reminds us that language is a social phenomenon—a system of meanings created collaboratively. It gives insights into the interlocutor's perspective and communication strategies. This interactive aspect of language is

almost impossible to be accurately studied through self-reports or isolated responses. However, this method reduces experimental control and increases transcription and coding costs.

Conclusion

Table 1 shows that these three methods serve different purposes and are not interchangeable. Self-report is best for examining individual differences, elicited production for analyzing the linguistic structure of utterances in typical contexts, and conversational data are essential for studying interactional use and functions.

Table 1. Comparison of three approaches to studying language use.

Dimension	Questionnaire	Elicited Production	Conversation
Ecological Validity	↓↓	↓↑	↑↑
Researcher Control	↑↑	↓↑	↓↓
Linguistic Detail	↓↓	↓↑	↑↑
Researcher Effort	↓↓	↓↑	↑↑

Acknowledgements

This work was supported by the National Science Center (NCN) in Poland, under Grant PRELUDIUM (2023/49/N/HS2/03693).

References

Paulhus, D.L. 2002. Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, D. E. Wiley (Eds.), The role of constructs in psychological and educational measurement (pp. 49–69). Erlbaum.

Gernsbacher, M.A., Foertsch, A. 1999. Using picture-priming to investigate the structure and growth of discourse representations. Discourse Processes, 27(2), 167–187.

Garrod, S., Pickering, M. J. 2007. Alignment in dialogue. In Gernsbacher M.A. (Ed.), Handbook of psycholinguistics (pp. 305–325).

Poggi, I., Vincze, V. 2018. The role of facial expression and tone of voice in expressing and recognizing irony. Frontiers in Psychology, 9, 2368.

Deep neural networks identify sensitive regions of an acoustic tube

Runhui Song[1,2], Johan Sjons[1,3], Axel Ekström[2,3]
[1]Uppsala University, Sweden
[2]KTH Royal Institute of Technology, Sweden
[3]Stockholm University, Sweden

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/023/000683

Abstract

Tube vocal tract modelling is a staple of 20th century phonetics and speech acoustics research. Here, we apply modern data analysis methods – deep neural networks – to derive from tens of thousands possible configurations of an acoustic tube, all possible relationships between tube perturbations and formant frequencies. Here, we demonstrate the validity of the broader methodological framework, and illustrate how our deep neural network pipeline produce high fit for formant predictions made by a computer program simulating the acoustic properties of a close-to-open tube.

Keywords: tube, machine learning, vocal tract, speech production

Introduction

Relationships between speech production and acoustic outcome have long been a staple of articulatory phonetics research (Fant, 1971). Relevant historically influential theories have proposed "stable" regions of the vocal tract predicting the prevalence of vowel qualities in natural languages (Stevens, 1989; Mrayati et al., 1988); and "distinctive regions", where perturbations of some vocal tract areas are more influential on formant frequencies than perturbations of others. However, investigations targeting such relationships have to date never been performed incorporating recent advancements in big data machine learning techniques.

Methods

Predicting the behavior of an acoustic tube

We designed an experiment where a computational acoustic tube model was set to randomly perturb area function increments, holding lengths of segments constant. The length of the total section was held constant at 16 cm, while the number and areas of segments were varied systematically across three experiments: (1) a four-tube model, (2) an eight-tube model, and (3) a 16-tube model.

The algorithm was derived from Liljencrants and Fant (1975). Using this method, a transfer determinant is recursively computed throughout the tube sequence (i.e., list of segments). Angular frequency ω is defined as:

$$\omega = 2\pi F$$

where F denotes sound wave frequency in Hertz.

Normalized phase angle of the n^{th} segment is computed

$$\theta_n = \frac{\omega L_n}{c}$$

where c is the speed of sound at 35°C and L_n is the length of the n^{th} segment. The ratio of the area of two subsequent segments is represented as

$$k_n = \frac{A_{n+1}}{A_n}$$

and the formula for deriving the transfer determinant is

$$\begin{cases} \Delta_1 = \cos\theta_1 - \frac{\omega L_0}{c}\sin\theta_1, \\ \Delta_n = d_{n-1,n}\Delta_{n-1} - b_{n-1,n}\Delta_{n-2}, \quad \text{when } n \geq 2, \end{cases}$$

where

$$d_{n-1,n} = \cos\theta_n + k_{n-1}\cos\theta_{n-1} \cdot \frac{\sin\theta_n}{\sin\theta_{n-1}},$$

$$b_{n-1,n} = k_{n-1} \cdot \frac{\sin\theta_n}{\sin\theta_{n-1}}.$$

Finally, a quasi-spectral function is constructed, once the determinant for the final segment Δ_M has been computed

$$Y(F) = \cos^2(\arctan(\Delta_M))$$

In addition, an adjustment was also included specifically for transitions where a subsequent segment $L_{n-1} < 0.16 \cdot L_n$. This correction was derived from Ingård (1953) and is specified as:

$$\delta_i \simeq 0.48 \cdot \sqrt{A}(1 - 1.25\xi)$$

This correction was implemented to control for the atypical behavior of narrow-to-open segment transitions.[1]

Deep neural networks

To model area function-formant relationships and appropriately model the complex, non-linear dependencies between input features (area segments of the vocal tract) and the target output (formants), we employed multi-layer perceptrons (MLPs). The network architecture consisted of two hidden layers with 64 and 32 neurons, respectively. To our knowledge, this is the first such modelling effort.

The neural networks were trained on synthetic datasets generated through a tube model of the vocal tract. The model simulated speech by varying cross-sectional areas across different segments of the tract. Each dataset consisted of thousands of data points to ensure a representative sample of possible configurations. To validate the robustness of the models, k-fold cross-validation

was applied (2-fold for the first experiment, 4-fold for the subsequent ones), which helped mitigate overfitting and ensured generalizability.

To address that neural networks are notoriously opaque in terms of interpretability, we used SHapley Additive exPlanations (SHAP) (Shapley, 1953; Lundberg & Lee, 2017) to assess the influence of each input segment in affecting changes to formants.

Results

Our analyses reaffirm several key assumptions about speech production and acoustics, for that opening (i.e., lips) or anterior constrictions (i.e., oral cavity) had dominant roles in shaping F3, in ways that are consistent with both lip rounding and rhoticity. In addition, taken in sum, our observed SHAP values match perfectly previously reported "sensitivity functions" for segments observed for each of F1, F2, and F3 - effectively serving as a sanity check on the appropriateness of our methodology. However, our results also highlight several often under-recognized relationships. For example, our models consistently show a stark influence of constriction on F1 and F2 in the posterior-most segment (corresponding to the glottis or larynx opening).

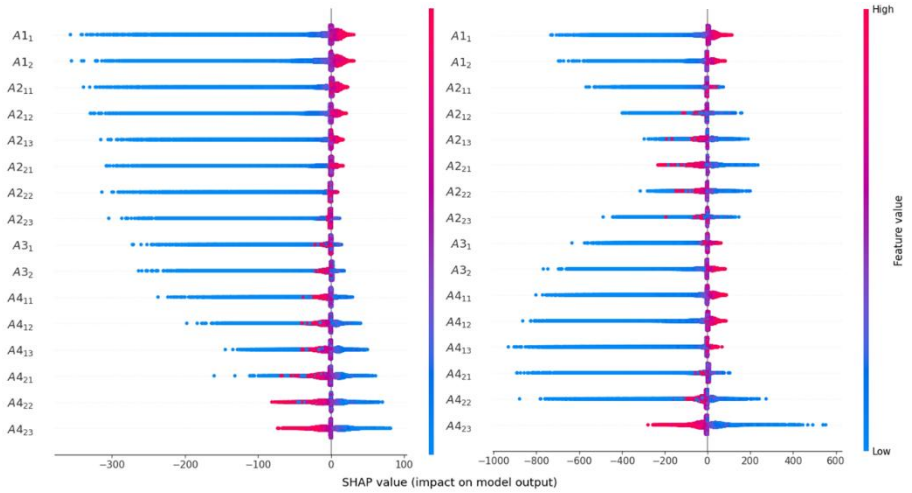


Figure 1. A "sensitive region" can be understood as one where effects of constriction vary significantly.

Discussion

Our work serves the dual purpose of adding to available methods for investigating a long-standing question in phonetic sciences - "how do perturbations of an acoustic tube correspond to speech output?"; and builds on, reaffirms, and nuances earlier attempts to answer the same. Our methodology is blind to any biases possibly imposed by pre-existing theory; yet, it reiterates a basis of phonetic and phonological theory, drawn purely from acoustic theory (Fant, 1971; Carré et al., 2017). Relationships underpinning speech can be derived from the properties of an acoustic tube.

Notes

1. Details pertaining to this correction are also found in Ingard (1953, p. 1041) and Fant (1971, p. 36). The issue is noted in Liljencrants and Fant (1975) but expressly not included, because natural configurations rarely include such rapid transitions.

Acknowledgements

AE was funded through the Swedish Research Council (2025–00209\VR).

References

Carré, R., Divenyi, P., Mrayati, M. 2017. Speech: A dynamic process. De Gruyter. https://doi.org/10.1515/9781501502019

Fant, G. 1971. Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations. Walter de Gruyter.

Ingard, U. 1953. On the theory and design of acoustic resonators. The Journal of the Acoustical Society of America, 25(6), 1037-1061. https://doi.org/10.1121/1.1907235

Lundberg, S.M., Lee, S.I. 2017. A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), Advances in neural information processing systems, 30 (NIPS 2017).

Liljencrants, J., Fant, G. 1975. Computer program for VT-resonance frequency calculations. STL-QPSR, 16, 15-21.

Mrayati, M., Carré, R., Guérin, B. 1988. Distinctive regions and modes: a new theory of speech production. Speech Communication, 7(3), 257-286. https://doi.org/10.1016/0167-6393(88)90073-8

Shapley, L.S 1953/1997. A value for n-person games. In H. W. Kuhn (Ed.), Contributions to the theory of games. Princeton University Press, pp. 307–317.

Stevens, K.N. 1989. On the quantal nature of speech. Journal of Phonetics, 17(1-2), 3-45. https://doi.org/10.1016/S0095-4470(19)31520-7

Dipping-tone contrasts in a multi-dipping-tone system: a case study of Lǔliáng Jìn Chinese

Yang Wei
Hong Kong Shue Yan University, Hong Kong, Shenzhen MSU-BIT University, China

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/024/000684

Abstract

This paper, through an analysis of 22 native Lǔliáng Jìn Chinsese speakers' recordings, demonstrates the contrasting patterns of dipping tones in multi-dipping-tone systems. It is discovered that there are 13 cases where a three-dipping-tone contrast is maintained, while nine cases exhibit a two-dipping-tone contrast. Within a multi-dipping-tone system, the dipping tones may exhibit variations in contours, durations, and phonations, thereby maintaining contrasts with each other.

Keywords: Lǔliáng Jìn Chinsese, dipping tones, multi-dipping-tone system, tonal types

Introduction

A dipping tone has a falling-rising contour. The most familiar case is Shǎngshēng in Běijīng Mandarin, which is transcribed as [214] using Chao tone letters. Dipping tones appear in a number of Chinese dialects. Zhu et al. (2012) first classified seven types of dipping tones under the framework of the "multi-register and four-level" tonal model. In their system, there are four dipping tones in the modal register, which are low-dipping /323/ (低凹調), back-dipping /523/ (後凹調), front-dipping /324/ (前凹調) and double circumflex /3232/ (兩折調); and three in the lower register, which are creaky low-dipping /202/ (嘎咧低凹調), creaky high-dipping /404/ (嘎咧高凹調) and breathy dipping /213/ (弛聲凹調).

Jìn is a major variety of Chinese; it is spoken in Shānxī province (山西省) and neighboring regions (Hou and Wen, 1993). In certain dialects of Jìn Chinese, there are multiple dipping tones, but with subtle variations. The Lǔliáng dialect is a major sub-dialect of Jìn. This study, by analyzing 22 cases of Lǔliáng dialects, finds that there are 13 cases maintaining a three-dipping-tone contrast, and nine cases having a two-dipping-tone contrast.

Methodologies

The "multi-register and four-level" tonal model (Zhu 1999, 2005), and the universal tonal inventory (Zhu 2014, 2018) are the main methodologies applied in this study. 22 native Lǔliáng Jìn Chinese speakers' recordings have been

acoustically and phonologically analyzed. The software Praat was employed to reduce noise on some recordings, to annotate the recorded tokens for tonal measurements, to extract pitch values, and to draw up spectrograms, etc. The method in Zhu (2010) was adopted to measure the tone-bearing part of each token and to normalize the inter- and intra-speaker variations.

Data analyses

This study, conducted on a sample of 22 cases of Lǔliáng dialects, reveals that 13 cases exhibit a three-dipping-tone contrast, while nine cases demonstrate a two-dipping-tone contrast.

Contrastive pattern A: low-dipping vs. back-dipping vs. mid-short low-dipping

There are nine speakers maintaining a "low-dipping vs. back-dipping vs. mid-short low-dipping" contrastive pattern. This is the most common dipping-tone contrastive pattern in Lǔliáng dialects. In this contrastive pattern, T1a is the low-dipping tone and T2 is the back-dipping tone, both of which are long tones; and T4b is the mid-short low-dipping tone. The figure below shows the tonal curves of the nine dialect sites.

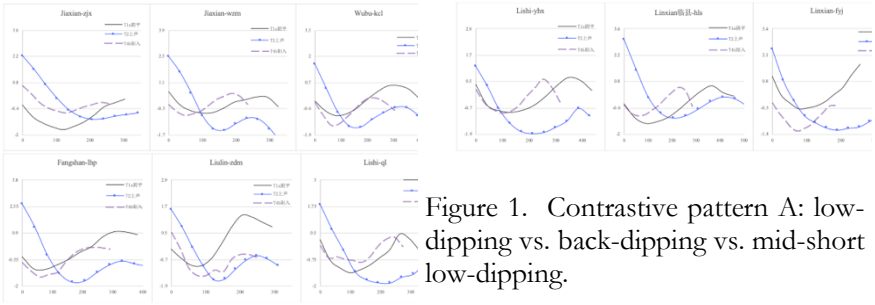


Figure 1. Contrastive pattern A: low-dipping vs. back-dipping vs. mid-short low-dipping.

Contrastive pattern B: low-dipping vs. back-dipping vs. mid-short back-dipping

Among the three-dipping-tone systems, there are three cases displaying a "low-dipping vs. back-dipping vs. mid-short back-dipping" contrastive pattern. The figure below shows the tonal curves of the three cases.

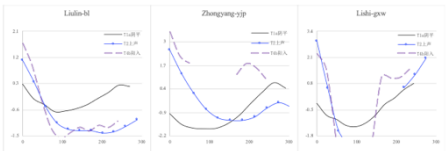


Figure 2. Contrastive pattern B: low-dipping vs. back-dipping vs. mid-short back-dipping.

Similar to the contrastive pattern A discussed above, both T1a and T2 are long tones. Furthermore, T1a is a low-dipping tone and T2 is a back-dipping tone. However, in pattern B, the T4b is a mid-short back-dipping tone, which means that T4b has the same contour as T2, but its duration is shorter than that of T2. The three dipping tones in pattern B are also contrastive along the two dimensions of height and duration.

In the cases we have analyzed in patterns A and B, all the T1as are low-dipping tones, and all the T2s are back-dipping tones. In the following eight cases, we can see that their T1a and T2 have merged into one tone, while the merged dipping tone might be either low-dipping or back-dipping.

Contrastive pattern C: back-dipping vs. mid-short back-dipping

The pattern C has five dialect sites. The neutralized tone of T1a and T2 of these five cases is a back-dipping tone. Among these five cases, the T4b in the case of Shílóu-cgh is a mid-short low-dipping tone, while the T4b in the other four cases is a mid-short back-dipping tone with creaky voice.

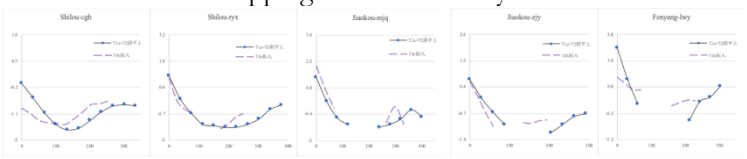


Figure 3. Contrastive pattern C: back-dipping vs. mid-short back-dipping.

Contrastive pattern D: low-dipping vs. mid-short low-dipping

The pattern D has three cases. In contrast to the pattern C, these three cases each have a neutralized low-dipping tone. All the T4bs are mid-short low-dipping tones.

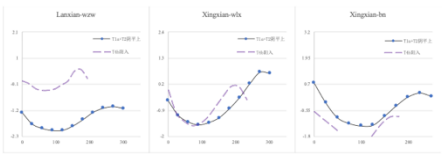


Figure 4. Contrastive pattern D: low-dipping vs. mid-short low-dipping.

Besides the twenty typical cases we have discussed above, the Wúbǔ-lcz and Fényáng-gf cases provide other contrastive possibilities of the dipping tones in Lǔliáng dialect. In the case Fényáng-gf, a "front-dipping vs. back-dipping vs. low-dipping" contrastive pattern has been formed. In the case of Wúbǔ-lcz, T4b prolonged its duration and finally merged into T1a. As a result, in the tonal system of Wúbǔ-lcz, two dipping tones remain.

Conclusions

Among the three-dipping-tone contrast cases, (1) there are nine speakers maintaining a "low-dipping vs. back-dipping vs. mid-short low-dipping" contrastive pattern. This is the most common dipping-tone contrastive pattern in Lǔliáng dialects. In this contrastive pattern, T1a is the low-dipping tone and T2 is the back-dipping tone, both of which are long tones; and T4b is the mid-short low-dipping tone. (2) There are three cases displaying a "low-dipping vs. back-dipping vs. mid-short back-dipping" contrastive pattern, in which T1a is a low-dipping tone, T2 is a back-dipping tone. and T4b is a mid-short back-dipping tone. (3) One case presents that T1a has a front-dipping or rising contour, T2 is a back-dipping tone, and T4b is a low-dipping tone.

Among the two-dipping-tone contrast cases, (1) five speakers display a "back-dipping vs. mid-short back-dipping" contrast; (2) three speakers present a "low-dipping vs. mid-short low-dipping" contrast; (3) and one speaker has a "low-dipping vs. back-dipping" contrast.

Acknowledgements

This work is supported by the Ministry of Education's Humanities and Social Sciences Research Project 教育部人文社会科学研究青年基金项目 titled "Jinfangyan rushengdiao de shengxue xingzhi, leixing ji yanhua yanjiu 晋方言入声调的声学性质、类型及演化研究" (Grant No. 23YJC740074), and the Guangdong Provincial College Early Career Research Grant (Humanities and Social Sciences) 广东省普通高校青年创新人才类项目 titled "Jiyu shiyan yuyinxue de jinyu bingzhoupian shengdiao leixing ji yanhua yanjiu 基于实验语音学的晋语并州片声调类型及演化研究" (Grant No. 2022WQNCX071).

References

Hou, J. Wen, D. (Eds.) 1993. Shanxi Fangyan Diaocha Yanjiu Baogao. Taiyuan: Shanxi College Associated Press.

Zhu, X. 1999. Shanghai Tonetics. Munich: Lincom.

Zhu, X. 2005. Shanghai Shengdiao Shiyan Lu [An experimental study on Shanghai tones]. Shanghai: Shanghai Educational Publishing House.

Zhu, X. 2018. Yuyin D. [On phonetics]. Shanghai: Xuelin Press.

Zhu, X, Yi, L., Zhang, T. 2012. Aodiao de zhonglei [A classification of dipping tones]. Zhongguo Yuwen [Studies of the Chinese language] 5:420-436.

Vowel restructuring under retroflex trill suffixation in JingMen Mandarin

Yue Xie, Roshidah Hassan
University of Malaya, Malaysia

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/025/000685

Abstract

This study employs acoustic and descriptive analyses to investigate how the rhotic trill [r], representing the -zi suffix, influences the monophthongal vowel system of Jingmen Mandarin. The analysis compares monosyllabic words representing the base vowel system with their -zi-suffixed counterparts, based on F1 and F2 formant measurements. The results show that underlying monophthongs tend to shift systematically toward [r], with front and back vowels undergoing resyllabification, while mid vowels exhibit stable shifts without resyllabification. These findings suggest that in Jingmen Mandarin, frontness plays a crucial role in the restructuring of the vowel system, as speakers adjust their articulation to accommodate the low-mid rhotic [r].

Keywords: vowel change, morphophology, rhotic suffix

Introduction

Rhotic suffixation represents a common morphophonological process in Chinese, typically associated with the *-er* suffix meaning 'son'. In the Jingmen context, however, there is no morphological realization of *-er*; instead, the suffix *-zi* ('son') carries a rhotic sound. Both suffixes have played crucial roles in derivational morphophonological processes since Middle Chinese, historically contributing to lexical expansion through the formation of disyllabic and polysyllabic words. In modern Chinese, although these suffixes have largely lost their semantic content and now function as empty morphemes, they remain phonetically significant in expressing pragmatic meanings such as diminutiveness, familiarity, and naturalness in everyday speech.

From a phonetic perspective, the rhotic suffix [r] in Jingmen Mandarin is characterized as an apical, trilled, voiced, and syllabic consonant when suffixed to a noun phrase (Liu 2017). On the one hand, its phonetic properties [+rhotic] and [+apical] correspond more closely to those observed in studies of the *er* suffix (for example, in Beijing Mandarin (Lee 2005) than to the *zi* suffix. On the other hand, its distinct manner of articulation (a trill) and its unspecified vowel quality make it particularly valuable for acoustic comparison with the plain vowel system, in order to explore how rhotic suffixation influences vowel production. Such a comparison allows us to determine whether the phonological processes in Jingmen Mandarin align with those associated with *er* suffixation. Specifically,

this study examines the acoustic characteristics of vowels in monosyllabic contexts (the plain vowel system) and in zi-suffix contexts (the zi system).

Methodology

Speech data were collected from two female native speakers of Jingmen Mandarin. Both speakers were born and raised in Shayang County, Jingmen City, and have lived there throughout their lives.

For the base vowel system, two meaningful monosyllabic words were selected to represent each rime. These words were embedded in a fixed carrier sentence, "把__再说一遍" ("say __ again"). For the *zi*-suffixed system, two disyllabic or trisyllabic phrases corresponding to each base vowel rime were selected and produced in isolation. All recordings were made using a Zoom H1n stereo recorder at a 44.1 kHz sampling rate, with four repetitions per token.

Acoustic analyses were conducted in Praat (version 6.0.19). Each monophthong token was divided into four equal intervals by five timepoints. For plotting, values from the middle timepoints were used, as well as the averaged values between the 75% and 100% timepoints. For the suffix [ɾ], formant values were extracted from the midpoint of the entire syllabic portion, using only tokens with preceding onsetless vowels [i], [u], and [a] to minimize consonantal influence.

Results

Firstly, to understand the relationship between the plain vowel system and the rhotic [ɾ], we examine Figure 1. It shows that [ɾ] aligns with vowels [a], [ɳ], and [ə] in terms of frontness, forming the middle category. The vowel [i] is front, [y] is mid-front, and [u] and [o] are back vowels. Regarding height, [i] and [y] are high vowels, while [u], [o], [ɳ], and [ə] share a mid height. The vowel [a] is noticeably low, with an F1 value exceeding 1100 Hz, whereas [ɾ] falls in the mid-low range. Overall, when comparing the plain vowel system with the rhotic suffix [ɾ], we can identify four degrees of frontness and four degrees of height contrast.

After understanding the feature contrasts between the plain vowel system and the rhotic [ɾ], we can observe two distinct patterns of change when suffixation occurs. As shown in Figure 2 (left panel), the dashed lines, which represent the average formant values from the final 25% of the vowel in F1 and F2, shift more centrally toward [ɾ] compared with the solid lines representing the midpoint values. This indicates a dynamic shifting of vowel quality over time.

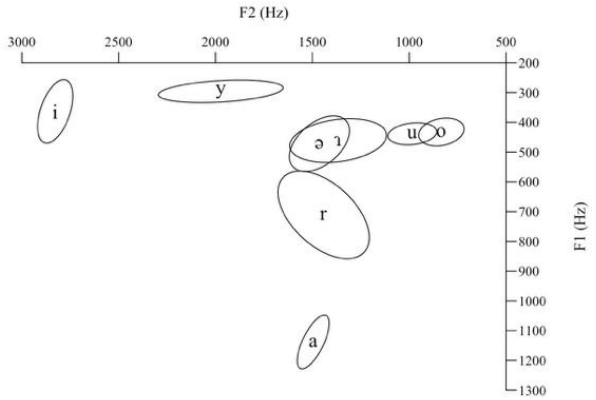


Figure 1. 1-sigma ellipses for Jingmen vowels [i, y, u, o, a, ɿ, ə] without suffixation, with rhotic [r] shown for reference

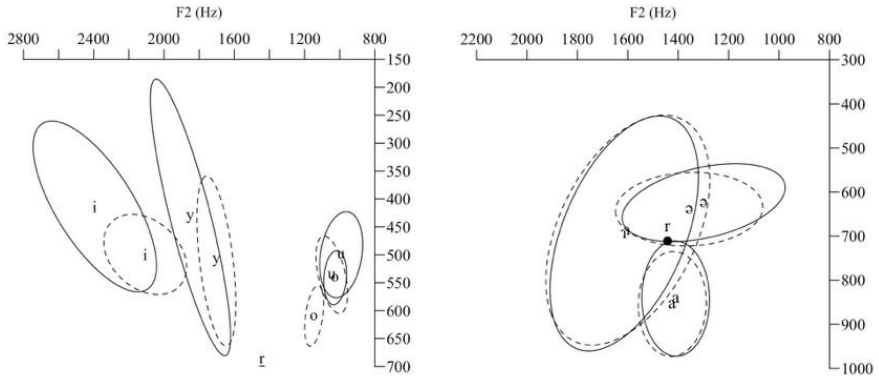


Figure 2. 1σ ellipses for Jingmen vowels with suffixation: front/back vowels [i, y, u, o] (right) and mid vowels [a, ɿ, ə] (left), in relation to rhotic [r]. Solid = 3/5 point; dashed = 4/5–5/5 average.

More specifically, for [y], [i], and [u], the trajectories of their final 25% overlap considerably with their midpoint values, suggesting that the influence of [r] extends beyond the vowel offset and already affects the vowel quality during its middle portion. In contrast, [o] displays a more distinct displacement between the midpoint and the final 25%, indicating a more consistent shift toward [r] at the offset.

The right panel of Figure 2 illustrates another pattern of rhotic influence. Unlike the gradual transitions seen on the left, these vowels [a], [ɿ], and [ə] exhibit little to no dynamic change between their midpoint and offset positions. Instead,

the entire vowel quality has shifted as a whole toward [r], forming a stable new phonetic realization. In addition to the diversity in whether resyllabification occurs, which is conditioned by the feature of frontness, there is also variation in how closely the vowels approach [r]. The front and back vowels move near [r], while the middle vowels largely overlap with it.

Discussion

The acoustic examination of plain and rhotic-suffixed vowels generally confirms the vowel centralization tendencies. However, unlike Beijing Mandarin, where resyllabification introduces an additional [ɚ] for most vowels except [u] (Lee 2005), the pattern in Jingmen Mandarin corresponds more closely to that observed in Southwestern Mandarin, where non-high vowels are entirely replaced by [ɚ], and high vowels exhibit gliding trajectories toward [ɚ] (Huang, Hsieh, & Chang 2020). Therefore, Jingmen Mandarin exhibits similar types of phonological patterns but with its own distinctive characteristics—namely, the rhotic suffix is realized as a low-mid [r] rather than a central [ɚ]; the vowels [o] and [u] are mid rather than high in height, yet they still exhibit gliding tendencies toward [r]. Overall, the pattern appears to be more sensitive to frontness. Besides, this suggests that although the suffix is morphologically realized as *-zi*, its phonetic and phonological effects on the vowel system are more comparable to those of the *-er* suffix.

References

Huang, J., Hsieh, F., Chang, Y. 2020. Er-Suffixation in Southwestern Mandarin: An EMA and Ultrasound Study, Interspeech.

Lee, W. 2005. A phonetic study of the "er-hua" rimes in Beijing Mandarin, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal.

Liu, H. 2017. Jingmen fangyan yanjiu [Jingmen dialect research]. Wuhan, China: Huazhong shifan daxue chubanshe [Central China Normal Univeristy Press].

Tonal preservation verse prosodic transfer in L3-Mandarin question intonation

Shujing Xu[1], Grace Wenling Cao[2]
[1]The Hong Kong Polytechnic University, Hong Kong
[2]University College Dublin, Ireland

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/026/000686

Abstract

This study examines prosodic transfer in L3-Mandarin among Cantonese-L1 speakers, investigating how pitch height and boundary timing signal interrogativity. Cantonese typically marks questions with a salient final-syllable F0 rise, whereas Mandarin employs overall pitch elevation while preserving final lexical tones. Twenty Cantonese-English-Mandarin trilinguals produced statements and questions in a reading conversation; their Mandarin proficiency was rated by native speakers. Results show asymmetric transfer: L1 boundary tone timing persisted in certain question types, yet the Mandarin-specific fall-rise tonal contour was maintained. Higher proficiency facilitated Mandarin-like pitch height modulation. The findings illustrate a duality in L3 acquisition, where prosodic transfer from the L1 coexists with target-like tonal production, highlighting how tonal typology constrains bilingual intonation.

Keywords: prosodic transfer, L3 acquisition, Mandarin, Cantonese, boundary tone

Introduction

Hong Kong's "Biliteracy and Trilingualism" policy fosters a population of Cantonese-English-Mandarin trilinguals. This provides a critical context for investigating cross-linguistic influence, particularly in prosody, between the two tonal languages of Cantonese and Mandarin. Crucially, they employ distinct strategies to signal interrogativity. Cantonese utilizes a boundary tone, which is realized as a salient final-syllable F0 rise that overrides lexical tones (Chen, 2020; Ge & Li, 2019). When questions end with T23 (low rising) which is contingent with boundary tone, T23 under boundary tone accelerates its F0 rise to approximate T25 (high rising). In contrast, Mandarin questions preserve lexical tonal shapes in the final sentence position (Chen, 2022) but raise overall pitch rather than imposing a final rise (Liu, 2009).

Despite this clear typological difference, the intonation patterns of L3-Mandarin produced by Cantonese-L1 speakers remain underexplored. This study investigates whether Cantonese speakers transfer their L1 boundary tone strategy into their Mandarin-L3, specifically in utterances ending with the rising tone (T214). The following research questions are addressed:

1. Do Cantonese-L1 speakers transfer the final-syllable timing of the Cantonese boundary tone to their L3 Mandarin questions?

2. Do Cantonese-L1 speakers' Mandarin proficiency affect the transfer? If yes, what are the patterns?

Methodology

The trilingual speech data were drawn from a forensic phonetic corpus (Cao & Mok, 2023). Participants were 20 young (M = 21.26, SD= 2.32), gender-balanced Cantonese-L1 speakers from Hong Kong. They completed an elicitation task in Cantonese and Mandarin, producing four sentence types: statement, yes-no question, intonation question, and wh-question. As showed in Table 1, all target sentences ended with the syllable "腦" (meaning "brain" in English; corresponding to tone T23 in Cantonese [nou23] and T214 in Mandarin [nau214]), and the bolded characters were used to visualize F0 contours. To access proficiency, 100 native Mandarin listeners rated the standardness of each speaker's Mandarin accent based on a 10-second recording, using a 10-point scale.

Target sentences were annotated in Praat. Using the ProsodyPro script (Xu, 2013), 10 F0 points were extracted per syllable for subsequent calculation of mean F0 and normalized for gender difference. Boundary tone divergence points were identified through visual inspection and validated via ANOVA. A linear mixed-effects model was implemented in R to assess the effects of proficiency and sentence types on final-syllable F0 mean value.

Table 1. The list of sentences for the elicitation task.

Sentence Type	Mandarin	Cantonese
Statement	你們拿了她的電腦。	你哋攞咗我部電腦。
	(English: You take her computer.)	
Yes-no Q	你們是不是也拿了她的電腦？	你哋係唔係都攞咗佢部電腦？
	(English: Did you take her computer as well?)	
Intonation Q	你們也拿了她的電腦？	你哋攞咗佢部電腦？
	(English: You take her computer?)	
Wh- Q	為什麼拿了她的電腦？	點解攞我部電腦？
	(English: Why did you take her computer?)	

Results

The F0 contours of the last three syllables (Figure 1) revealed systematic differences between L1-Cantonese and L3-Mandarin. In Cantonese, yes-no questions diverged from statements at the final syllable, while intonation

questions overlapped with statements until the final syllable's midpoint before sharply rising to mimic T25. In L3-Mandarin, however, intonation questions diverged earlier at the penultimate syllable, whereas yes-no and wh-questions retained final-syllable divergence. Crucially, all L3-Mandarin sentence types preserved T214's fall-rise shape.

Descriptively, the pitch of the final syllable was higher in questions than statements for both languages (see in Figure 1), though this difference was not statistically significant in ANOVA. Notably, in L3-Mandarin, the F0 contours for questions were generally higher than for statements, whereas in Cantonese they showed more overlap. This pattern in L3-Mandarin aligns with the native Mandarin strategy of using overall higher pitch to mark questions (Liu, 2009). To investigate this, a linear fixed model was fitted to examine the effect of Mandarin proficiency and sentence types on F0 mean value of the last syllable in Mandarin-L3. Participants demonstrated a fair level of Mandarin proficiency (M = 5.86, SD = 0.91). The analysis revealed a significant positive effect of Mandarin proficiency for F0 values (β = 11.41 Hz, t = 6.09, p<0.001), and a mild negative effect of statement (β = -9.51 Hz, t = -1.97, p=0.05).

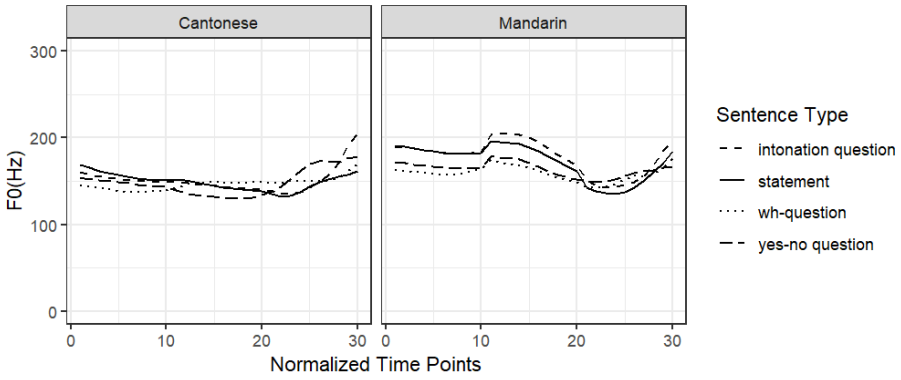


Figure 1. The last three syllables' F0 contours of four sentence types in Cantonese and L3-Mandarin

This study demonstrates asymmetric prosodic transfer in Cantonese-L1 speakers' L3-Mandarin. The timing of the L1 boundary tone partially persists, as seen in the final-syllable divergence of yes/no questions in L3-Mandarin. However, trilingual speakers successfully maintained the target Mandarin T214 hall-rise contour across all sentences, contrasting with Mandarin-L1 speakers' use of falling contours in statements and falling-rising in the yes-no, confirmation and particle questions (Liu, 2009).

Furthermore, higher Mandarin proficiency significantly predicted increased final-syllable F0 in questions, suggesting that the more advanced participants were, the higher pitch they used in the final syllable of all questions. It is

consistent with the strategy adopted by native Mandarin speakers (Liu, 2009). This duality reflects a tension in L3 prosody acquisition: the Cantonese-dominant prosodic strategy of boundary tone timing competes with the Mandarin requirement for lexical tonal preservation. The findings highlight how tonal typology constrains bilingual intonation.

Acknowledgements

The study is based on the first author's master's thesis completed at The Chinese University of Hong Kong. I wish to express my sincere gratitude to my thesis supervisors, Dr. Chunyu Ge for his expertise and patience, and Professor Peggy Mok for her mentorship and support throughout the process.

References

Cao, G.W., Mok, P. 2023. The Acoustics of cross-linguistic filled pauses in Cantonese-English-Mandarin trilingual speech. Proceedings of 20th International Congress of Phonetic Science (ICPhS), 3814-3818

Chen, Y. 2022. Tone and Intonation. In C.-R. Huang, Y.-H. Lin, & I.-H. Chen (Eds.), The Cambridge Handbook of Chinese Linguistics (1st ed., pp. 336–360). Cambridge University Press.

Ge, C., Li, A. 2019. Intonation of Cantonese interrogative sentences with and without sentence final particle. ICPhS 2019, Melbourne, Australia, 2455-2459.

Xu, Y. 2013. ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France. 7-10.

Liu, F. 2009. Intonation systems of Mandarin and English: A functional approach. The University of Chicago.

Phonetics 2025 Hong Kong: Proceedings International Conference on Phonetic Research and Applications

Index of names