

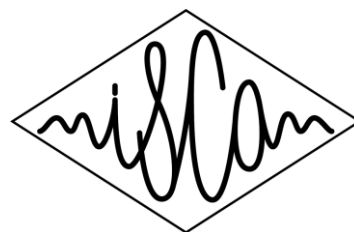


3rd International Conference on Tone and Intonation

TAI 2025

16. – 18. May 2025

in Herrsching near Munich, Germany



Variation in real-time lexical tone processing: A psycholinguistic perspective

Caicai Zhang (The Hong Kong Polytechnic University)

Variation is ubiquitous in human speech. A speech sound has different acoustic realizations when produced by different speakers or surrounded by different speech sounds, posing a challenge for fast and consistent speech perception in the listener's brain. Beyond acoustic variation, there are systematic phonological alternations, where the context-specific phonological form must be selected in a super-fast manner prior to articulation in the speaker's mind. However, how variation is accommodated in speech perception or planned in speech production is not well understood. In this talk, I will focus on variation on a short timescale, i.e., a few hundred milliseconds of real-time speech perception and production, probed via the lens of a highly variable phonological entity – lexical tone. I will present evidence showing that listeners deploy multiple cues, including contextual acoustic cues, population F0 knowledge, and their own vocal F0 cues, to achieve fast and accurate tone perception, especially when the stimuli are variable, ambiguous, or acoustically degraded. On the production side, our studies on Mandarin Third Tone sandhi show that speakers encode the abstract tonal category at an earlier phonological encoding stage and the surface, context-specific tonal form at a later phonetic encoding or motor preparation stage before articulation. These neural processes appear to be largely similar across lexicality (real vs. pseudowords) and word frequency (high vs. low frequency) conditions. These findings have implications for advancing speech perception and production theories. Psycholinguistic methods are instrumental in understanding basic mental processes in real-time lexical tone processing, highlighting that the human speech system is fast, flexible, and multi-pronged.

Structured variation in cross-modal coordination: insights from prosody production and perception

Petra Wagner (Bielefeld University)

It is a long established fact that the expression of prosodic events is not restricted to fundamental frequency, but encompasses a myriad of cues, including but not limited to intensity, voice quality, duration and articulatory precision. In the last decades, further evidence showed that the expression of prosody is not even restricted to the modality of speech, but can extend to facial expressions, head and body movements as well as gesticulations (e.g., gesture strokes). To this day, we do not have reached a consensus on how the interaction of these multiple cues is organized in production, and how listeners integrate these many -highly variable- cues in perception.

In my talk, I will address the degrees of freedom, variation and synchronization of prosody-related events in cross-modal production, and discuss how listeners may follow different strategies in cue integration. Lastly, I will suggest an interaction model, which relies on prosodic structure for cross-modal anchoring and coordination, thereby increasing the robustness of communication.

Musical pitch perception by speakers of African tone languages

Sharon Rose, University of California San Diego, USA

Speaking a tone language is reported to confer benefits in processing pitch in music (Pfordresher & Brown 2009; Wong et al. 2012; Hutka et al. 2015, among others). Yet, almost all research in this domain has tested speakers of East Asian languages, namely Mandarin, Cantonese, and Vietnamese, where the effect appears to be robust. African languages make up a large percentage of the tone languages spoken in the world, but are underrepresented in cognitive science research, particularly in this domain. Two previous studies address whether speakers of African tone languages also have a musical pitch processing advantage. Bradley (2016) tested fifteen speakers of Yorùbá and reported that both Mandarin and Yorùbá speakers are more sensitive to interval and contour musical pitch processing than English speakers. However, Creel, Obiri-Yeboah & Rose (2023) demonstrate that this ability is not universal among tone language speakers; 80 speakers of Akan, a language of Ghana, do not show the advantage with respect to interval processing. Akan has two level tones, and no contour tones, whereas Yorùbá has three level tones, and the East Asian languages tested have four or more tones, most of which are contours. In this talk, I will present follow-up research to the Creel et al (2023) study and explore i) the properties of tone systems that might confer better musical pitch processing, including number and type of tones, and ii) what musical cultural factors may impact results. Over forty speakers each of two Niger-Congo languages, Yorùbá (Nigeria), and Baatonum (Benin) participated in the same experiment as the Akan participants. Baatonum has four level tones, making it an ideal comparison to Mandarin, which also has four tones. We will report on whether the number and type of tones affects performance on the task. In addition, in prior studies, participants with musical training typically outperform those with no musical training (Bidelman et al 2013). However, musical training in a West African context must be assessed differently. In particular, speech surrogate systems, in which the tones of language are played on musical instruments such as the talking drum, are part of the cultures of all African groups tested, but its prevalence and cultural embedding varies. We will explore the possible impact of understanding talking drums on the results.

Joint research with Samuel Akinbo, Samuel Asitanga, Sarah Creel, Micheal Obiri-Yeboah and Yaya Yadoma.

References

- Bidelman, G. M., Hutka, S., and Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: evidence for bidirectionality between the domains of language and music. *PLoS One* 8:e60676. doi: 10.1371/journal.pone.0060676.
- Bradley, E. D. (2016). Phonetic dimensions of tone language effects on musical melody perception. *Psychomusicology*, 26(4), 337–345.
- Creel, S. C., Obiri-Yeboah, M., & Rose, S. (2023). Language-to-music transfer effects depend on the tone language: Akan vs. East Asian tone languages. *Memory & Cognition*, 51(7), 1624–1639. doi: 10.3758/s13421-023-01416-4.
- Hutka, S., G. M. Bidelman, S Moreno. (2015). Pitch expertise is not created equal: Cross-domain effects of musicianship and tone language experience on neural and behavioural discrimination of speech and music, *Neuropsychologia* 71, 52-63. doi: 10.1016/j.neuropsychologia.2015.03.019.
- Pfordresher, P. Q., and Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Atten. Percept. Psychophys.* 71, 1385–1398. doi: 10.3758/APP.71.6.1385.
- Wong, P. C. M., Ciocca, V., Chan, A. H. D., Ha, L. Y. Y., Tan, L. H., and Peretz, I. (2012). Effects of culture on musical pitch perception. *PLoS One* 7:e33424. doi: 10.1371/journal.pone.0033424

The dynamical approach to intonation: why and how?

Simon Roessig (University of York)

Dynamical systems have proven to be successful and powerful in describing a wide variety of natural phenomena and are attracting increasing attention in linguistics and phonetics. They also offer interesting perspectives for modelling intonation – primarily due to three properties: First, dynamical descriptions capture the evolution of states over time, highlighting the importance of temporal aspects in representations. Second, the interplay of continuity and discreteness is at the core of dynamics, offering promising solutions for the integration of phonetics and phonology. Third, in dynamical systems, a clear notion of noise and variability exists, making them potentially well suited for capturing the ubiquitous multiplicity of prosodic data. While dynamical approaches to intonation have been explored for some time, the development of more explicit computational models has accelerated in recent years. In this talk, I will outline fundamental properties of the dynamical approach, review significant advancements in the field, and discuss challenges and future directions.

Distributional Learning Across Contexts: Learning Cantonese Tones in Naturalistic Speech

Fengyue (Lisa) Zhao¹, Jennifer Kuo¹

¹Cornell University

Keywords: Distributional learning, Cantonese, Tones, Variability, Naturalistic speech

Motivation: Infants initially discriminate most sound contrasts but quickly attune to those of their native language. This raises the question: how do infants identify the relevant acoustic dimensions for learning phonetic categories? The **distributional learning** account proposes that infants track the distribution of sounds, and identify acoustic dimensions as contrastive if their distribution has two or more distinct peaks (i.e. multimodal distributions) [1]. However, while multimodalities appear in controlled experiments, they are rarely found in naturalistic, highly variable speech, suggesting that multimodality is not a reliable way to identify contrastive dimensions [2]. Recent work comparing languages with/without vowel length contrasts suggests that even without multimodality, contrastive dimensions show more **contextual variability**: when a dimension is contrastive, the shape of its distribution will vary more across contexts [3]. The **distributional learning across contexts** hypothesis proposes that infants utilize this contextual variability to distinguish phonetic categories. This study tests this hypothesis by examining Hong Kong Cantonese tones, exploring whether ease of acquiring different tonal contrasts is linked to their contextual variability in distribution shape. Cantonese serves as a valuable test case due to the overlapping acoustic distributions between its six tones: high-level (T1), high-rising (T2), mid-level (T3), low-falling (T4), low-rising (T5), and low-level (T6).

Methods: We analyzed the Multi-ethnic Hong Kong Cantonese Corpus (MeHKCC) [4], which consists naturalistic speech recordings from 25 native Cantonese female speakers. 59,385 monosyllabic and disyllabic content words were extracted. Pairwise F0 contour comparisons showed varying acoustic overlap among tones, except for the phonetically distinct T1 (e.g., distinct pair: T1T4, overlapping pairs: T3T5, T2T5; see Fig.1). Based on acoustic overlap and documented acquisition difficulty [5], tone pairs were categorized into: (1) *Easy* pairs, which are phonetically distinct and easy to learn (e.g., T1T4); (2) *Overlapping* pairs, which are acoustically overlapping but learnable (e.g., T3T5); and (3) *Merger* pairs, which are acoustically overlapping and challenging to learn (e.g., T2T5). We predict that contextual variability in distribution shapes aligns with developmental acquisition patterns, with *Easy* contrasts showing the most separation and variability, followed by *Overlapping* and *Merger* pairs. Although *Overlapping* and *Merger* pairs both show acoustic overlap, we predict that *Overlapping* pairs are more learnable due to the greater contextual variability in their distributional shapes.

To test this, nine F0 descriptors (mean, median, variance, max-min, onset, 25%, 75%, offset, duration) were extracted, and t-distributed stochastic neighbor embedding (t-SNE) was used to reduce these dimensions to a 2D space. Distributional differences were quantified using Earth Mover's Distance (EMD) [6] for pairwise tone comparisons across contexts. Contexts were defined as combinations of (1) neighboring sounds (e.g., stops, fricatives, nasals), (2) syllable position in a word (i.e., first or second syllable in a word), and (3) prosodic position (i.e., utterance-initial, -medial or -final).

Results: Analyses were conducted for all tone pairs, with T1T4 (*Easy*), T3T5 (*Overlapping*), and T2T5 (*Merger*) selected for illustration. Fig. 2 shows the frequency distribution of the *Overlapping* tone pair T3T5 after dimensionality reduction. While tone pairs show unimodal distribution when pooled across contexts (Panel A), they show different distribution shapes across specific contexts (Panel B shows two illustrative contexts). Figure 3 presents a boxplot of EMD for the three tone pairs, where each data point represents the pairwise EMD of two tones within a single context. Higher mean EMD values indicate greater distributional separation in general, while higher variance across contexts reflects greater contextual variability. Across four EMD metrics—mean, median, variance, and maximum—*Easy* pairs consistently show the highest values, followed by *Overlapping* pairs, and then *Merger* pairs (values provided in Fig. 3). This hierarchy aligns with developmental acquisition patterns: tones with greater separation and contextual variability are learned more readily than tones with lower values. Analyses of all 15 tone pairs reveal similar trends, with more nuanced interactions between distributional learning across contexts and acoustic realizations.

Discussions: This study explored the learning of multiple tone contrasts, a relatively unexplored area in distributional learning. Findings suggest that infants may rely on distributional shapes across contexts to learn contrasts, offering a plausible mechanism for learning in the absence of invariance in speech signals. Future direction will expand to additional corpora with additional contexts, and develop computational learning models to quantitatively capture the learning trajectories of all tone pairs.

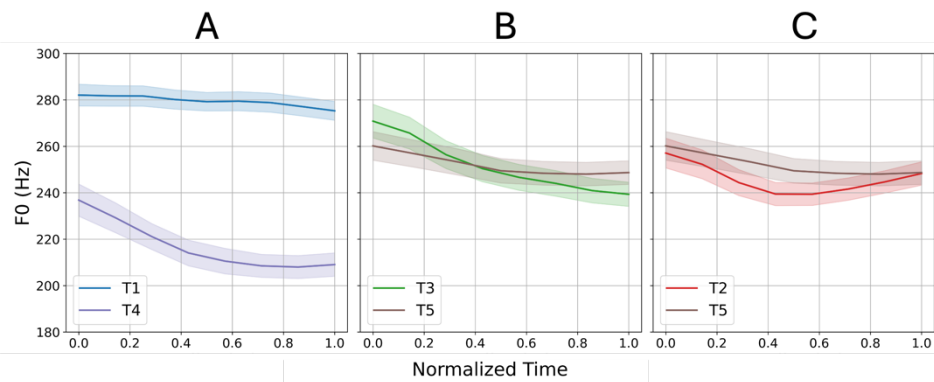


Figure 1. Pairwise F0 contours for three example tone pairs from a female speaker, with mean and 95% confidence intervals, time-normalized. (A) T1T4 shows clear phonetic distinction, while (B) T3T5 and (C) T2T5 exhibit varying degrees of overlap.

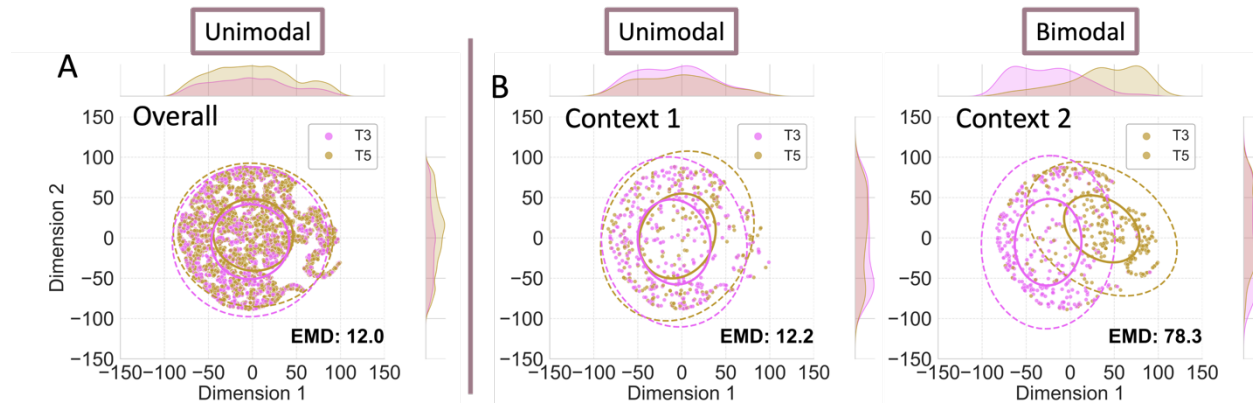


Figure 2. (A) The overall frequency distribution of the *Overlapping* tone pair T3T5 along a reduced 2D space shows a unimodal distribution, despite T3 and T5 being contrastive tones. (B) Frequency distributions for the same tone pair across two specific contexts reveal different distributional shapes. Context 2 (e.g., nasal onsets without codas, second syllable, utterance-medial) exhibits greater separability (EMD = 78.3) compared to Context 1 (EMD = 12.2). Even though the overall frequency distribution is unimodal, we can see differently shaped distributions across context.

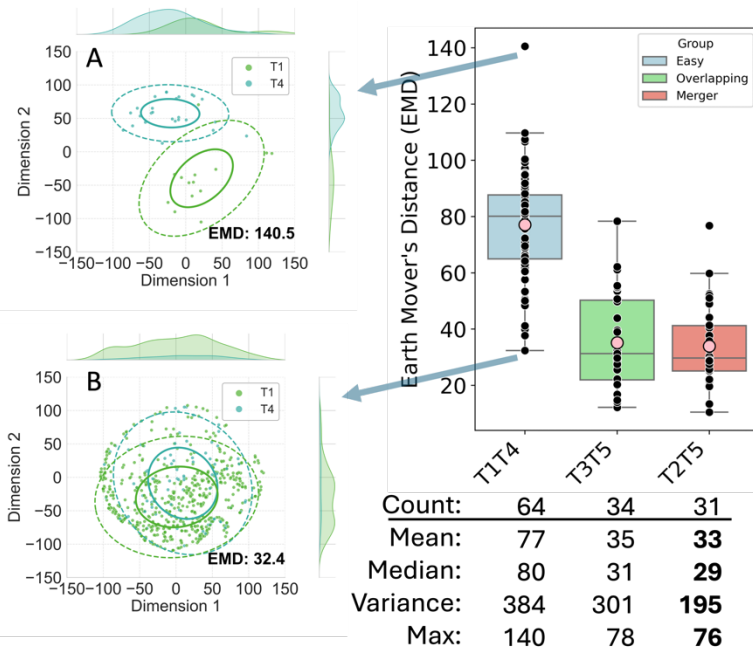


Figure 3. Earth Mover's Distance (EMD) distributions for three tone pair categories: *Easy* (T1T4), *Overlapping* (T3T5), and *Merger* (T2T5). Each data point represents the pairwise EMD between two tones within a specific context. Panel A illustrates a context with the greatest separability, while Panel B shows a context with the lowest separability. Lowest values for four EMD metrics (mean, median, variance, and maximum) were bolded.

- References:** [1] Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), B101–B111. [https://doi.org/10.1016/s0010-0277\(01\)00157-3](https://doi.org/10.1016/s0010-0277(01)00157-3)
- [2] Bion, R. A. H., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning Phonemic Vowel Length from Naturalistic Recordings of Japanese Infant-Directed Speech. *PLoS ONE* 8(2): e51594. <https://doi.org/10.1371/journal.pone.0051594>
- [3] Hitczenko, K., & Feldman, N. (2022). Naturalistic speech supports distributional learning across contexts, *Proc. Natl. Acad. Sci. U.S.A.* 119 (38) e2123230119, <https://doi.org/10.1073/pnas.2123230119>
- [4] Yu, A., Delisle, N., Martin, N., Zhang, V., Yao, Y., & To, C. (2024). The Multi-ethnic Hong Kong Cantonese Corpus. *CorpusPhon satellite workshop at LabPhon19*. <https://ccds.edu.hku.hk>
- [5] Mok, P. P. K., Fung, H. S. H., & Li, V. G. (2019). Assessing the Link Between Perception and Production in Cantonese Tone Acquisition. *Journal of Speech, Language, and Hearing Research*, 62(5), 1243–1257.
- [6] “Wasserstein metric”, https://en.wikipedia.org/wiki/Wasserstein_metric

How do children process complex tone sandhi? The case of Xiamen Southern Min

Huangyang Xie, Weijun Zhang, Peggy Pik Ki Mok

The Chinese University of Hong Kong

huangyangxie@cuhk.edu.hk, weijunzhang@cuhk.edu.hk, peggymok@cuhk.edu.hk

The mental representation of tone sandhi has received much research interest. While studies on child language acquisition can provide valuable insights, only a few studies have investigated this topic. Research shows that Mandarin-speaking children could achieve an adult-like accuracy in tone sandhi production for real words by age 3. However, their ability to apply tone sandhi to novel materials, such as those with accidental gaps, does not reach an adult-like level until around 7 years of age [1, 2]. In a study involving only three children and using both real words and novel words made up of real syllables, Hsieh [3] argued that the tone sandhi in Taiwanese Southern Min (a cyclic pattern with tone substitution) is neither productive nor learnable. He proposed that the sandhi tones were stored as allomorphs without generalization. In contrast, Li and Mok [4] observed a clear developmental trajectory for tone sandhi acquisition in children speaking Xiamen Southern Min. The present study revisits this debate by collecting more children data from Xiamen Min speakers using both real and novel words.

In this study, there were 12 old (60-87 years old), 19 middle-aged (35-56 years old) speakers, 15 teenagers (13-19 years old) and 49 children (4-12 years old) who did both the Peabody Picture Vocabulary Test (PPVT-4) and a production experiment. Before the experiment, the participants (or their parents) were asked to complete a questionnaire on their language background. The participants were then asked to complete an adapted version of a subset of the PPVT-4 in both Mandarin Chinese and Xiamen Southern Min. Different tasks of the production experiment were then completed in a fixed order which was a picture-naming task with 5 sections: 1) monosyllabic real words; 2) disyllabic real words (with one picture, DO); 3) disyllabic real words (combining two pictures appearing in section 1, DT); 4) pseudo words (combining two pictures resulting in a nonword, AO); 5) combining an accidental gap syllable on the sandhi position and a real monosyllable (AG).

The overall accuracy of the four age groups (*Fig. 1*) is consistent with Ge & Mok [5]: the accuracy increases from children to old speakers. However, participants of all age groups performed similarly in the AG condition: only around 30% of the tokens were produced with correct sandhi patterns. Also, it is attested that all age groups had much higher correct rates (about 60%) for two sandhi patterns: 44>22 and 24>22, whereas the other sandhi patterns exhibited an accuracy lower than 20% (*Fig. 2*).

A linear mixed-effects model was applied to the data of all non-adults which confirms that age is not significantly associated with accuracy ($p = 0.805$), whereas PPVT scores are significantly correlated with accuracy ($\beta = 3.425$, $p < 0.001$). For analysis, non-adults were grouped into 4 PPVT quartiles in descending order (level 1 the highest). As shown in *Fig. 3*, this reveals a distinct contrast in accuracy trends between AG and non-AG words: the accuracy of non-AG words decreases significantly with lower PPVT scores, while AG words are at about 30% correct across all quartile groups. The accuracy patterns of different sandhi rules within AG words are consistent with those observed in all age groups. The model shows that AO and AG words have significantly lower accuracy than DO words, and within AG words, 44>22 and 24>22 achieve significantly higher accuracy than the other sandhi rules (*Fig. 5*).

The findings suggest that speakers, regardless of proficiency, employ different strategies to process complex tone sandhi on phonologically invalid words (AG) compared to real or novel words abiding by phonotactic constraints. The PPVT results also show that language proficiency affects tone sandhi application (except for AG). The Xiamen tone sandhi patterns may differ in markedness, which may account for the substantial gap between the accuracies of different sandhi patterns. More analysis of the error and acoustic patterns is underway.

Keywords: complex tone sandhi, child acquisition, AG words, Xiamen Southern Min

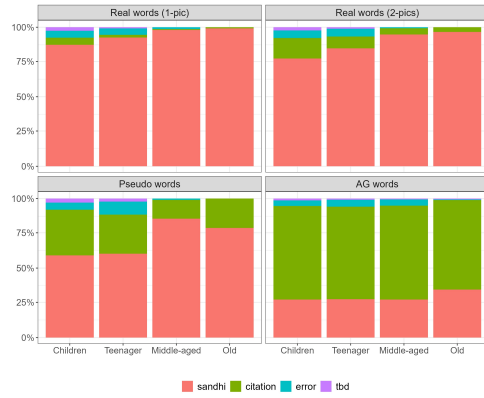


Figure 1: Accuracy of different word types by age groups



Figure 2: Accuracy of AG words by sandhi rules and age groups

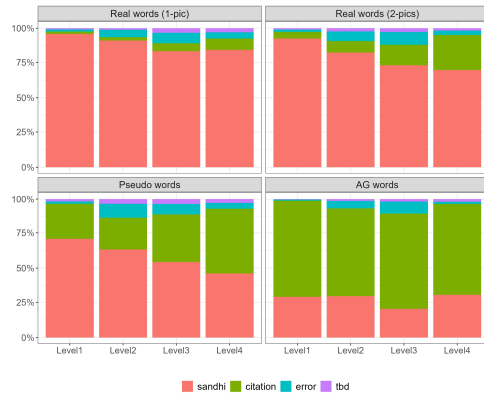


Figure 3: Accuracy of different word types of non-adults

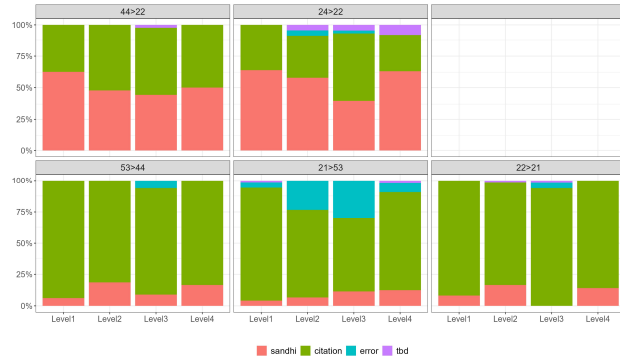


Figure 4: Accuracy of AG words by sandhi rules of non-adults

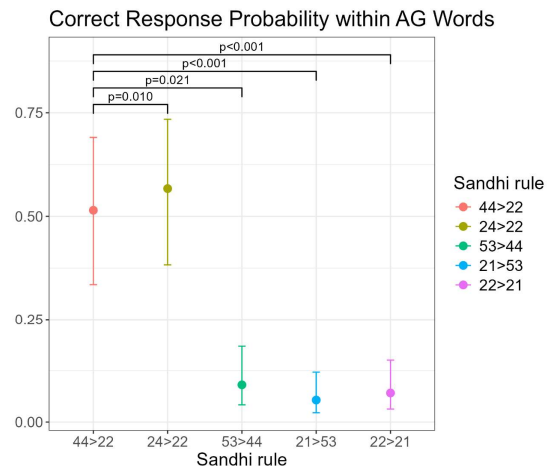
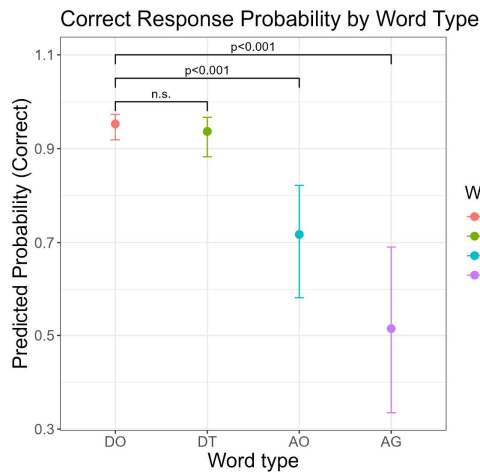


Figure 5: Predicted accuracy of non-adult participants grouped by (a) word types and (b) sandhi rules within AG words

Selected References:

[1] Huang, X., Zhang, G., & Zhang, C. (2018). A Preliminary Study on the Productivity of Mandarin T3 Sandhi in Mandarin-speaking Children. In *Proc. TAL 2018* (pp. 88-92). [2] Huang, X., Zuo, Y., & Zhang, C. (2019). Seven-year-olds reach an adult-like productivity in the application of Mandarin tone sandhi. In *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 3125-3129). [3] Hsieh, H.-I. (1975). How generative is phonology. In Koerner, E. F. K. (eds.), *The Transformational-Generative Paradigm and Modern Linguistic Theory*. John Benjamins, 109–144. [4] Li, X. & Mok, P. (2020) The acquisition of tone sandhi of the Xiamen dialect. In *Proceedings of the 10th International Conference on Speech Prosody 2020*, 479-483. Tokyo, Japan. [5] Ge, C. & Mok, P. (2024) The effect of phonotactic constraints on tone sandhi application: A cross-sectional study of Xiamen Min. Proc. In *Proceedings of Speech Prosody 2024*, 403-407. Leiden.

Learning second language lexical tone discrimination: Under what conditions is the perceptual boost from incremental cue training retained?

Yanyu Li, Laurence White & Ghada Khattab
Speech and Language Sciences, Newcastle University

Key words: incremental cue training, tone discrimination

First language (L1) experience can affect perceptual weighting of acoustic cues relevant to meaningful second language (L2) contrasts [1]. For challenging segmental contrasts (e.g., /i/-/ɪ/ for L1 Spanish listeners), exaggerating underweighted cues to increase perceptual attention has been effective [2]. Here we test cue exaggeration in the tonal domain, specifically for tonal languages where *F0* direction is a primary cue (e.g., Mandarin). Speakers of non-tonal languages (e.g., English) have been shown to be less perceptually sensitive than tonal speakers to *F0* direction cues (e.g., in tone identification tasks) [3]. We test the effect of incremental exaggeration of *F0* direction in tone discrimination training, to explore the conditions where a possible perceptual boost from cue exaggeration can persist when exaggeration is removed.

Over three studies, we trained and tested 246 L1 English listeners on four pseudo-tones synthesised based on Mandarin, and primarily contrasting on *F0* direction (Fig.1). Study 1 (N=86) compared two Training Types (**Incremental** vs **Fixed**) with participants randomly assigned to each of the groups. The **Incremental** group was trained on tonal stimuli (*high-rising 45* vs *high-falling 41*) where *F0* movement relevant to *rising/falling* was exaggerated at training outset and reduced blockwise (Fig.1) to baseline (Training 3). The **Fixed** group were trained throughout with the baseline version of the same tone pair. In Pretest and Posttest, both groups were tested on an identical untrained tonal contrast *level 44* vs *falling-rising 425*. Each testing/training trial consisted of an ABX task (*Is Sound X the same as Sound A or Sound B?*), with visual feedback ("correct"/"incorrect") in training blocks only. Study 1 used Training 1 60Hz and Training 2 30Hz exaggeration. Study 2 reduced exaggeration to Training 1 30Hz and Training 2 15Hz. Study 3 used Study 1 *F0* exaggeration, but pseudorandomly varied *F0* height, to test if cue exaggeration is more effective in high variability training contexts: thus, onset *F0* values were pseudorandomly varied within and between trials, whilst tonal slopes remained unchanged.

Response accuracy was analysed through mixed-effect logistic regression models, separately for test and training data, with fixed effects of Training Type (**Fixed** & **Incremental**), Exaggeration Magnitude (Study 1 – **30Hz/block** vs Study 2 – **15Hz/block**) or *F0* Height (Study 1 **Consistent** vs Study 3 **Varied**), and an ordered factor of Block (Training 1 vs 2 vs 3; Pretest vs Posttest). There were by-participant and by-syllable random intercepts, and a block-by-participant random slope. Effects were tested via likelihood ratios in backwards model comparison. For Studies 1 vs 2 (Fig.2 left), the **Incremental** group was more accurate than the **Fixed** group only with stimuli exaggeration at 60Hz and 30Hz, not with exaggeration at 15Hz. A Training Type x Block interaction, $\chi^2(2) = 38.99, p < .001$, indicated that accuracy decreased over training for the **Incremental** group, but increased for the **Fixed** group. This interaction was also found in Study 1 vs 3 (Fig.2 right), $\chi^2(2) = 34.25, p < .001$, where the **Incremental** group outperformed the **Fixed** group on exaggerated stimuli at 60Hz and 30Hz, $\chi^2(1) = 21.77, p < .001$. For all three studies, accuracy improved from Pretest to Posttest. Additionally, the Study 3 participants, with varied *F0*, had lower accuracy in both test ($\chi^2(1) = 23.54, p < .001$) and training blocks ($\chi^2(1) = 14.15, p < .001$).

Both training methods (**Incremental** & **Fixed**) improved discrimination, but the perceptual boost in Incremental training was not sustained when exaggeration was at 15Hz or was removed. Moreover, there was no sustained boost to Incremental training with greater *F0* height variability in Study 3. Ongoing work is examining whether the immediate perceptual boost from cue exaggeration can be reinforced and sustained given multiple separate training sessions.

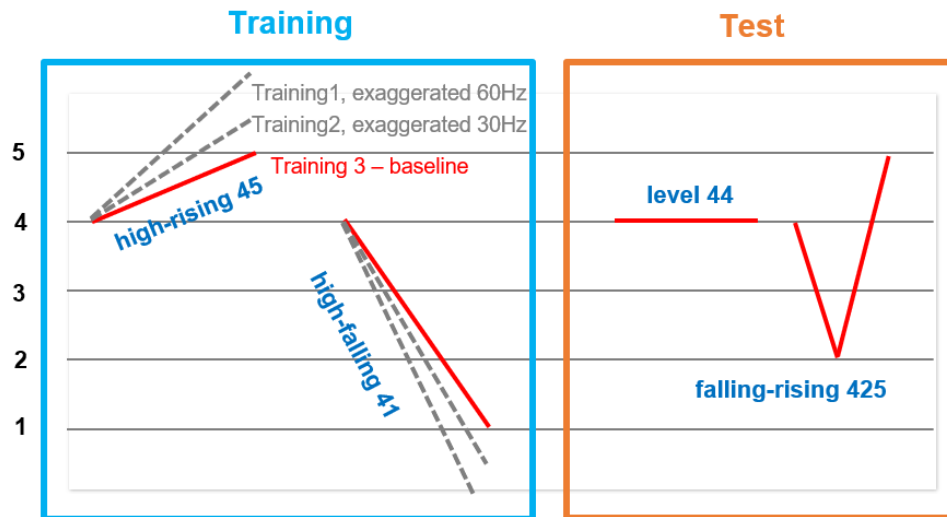


Figure 1. The four pseudo-tones represented in Chao digits [6]. High-rising 45 and high-falling 41 (shown with levels of exaggeration for the Incremental group) were used in training blocks; Level 44 and falling-rising 425 in test blocks.

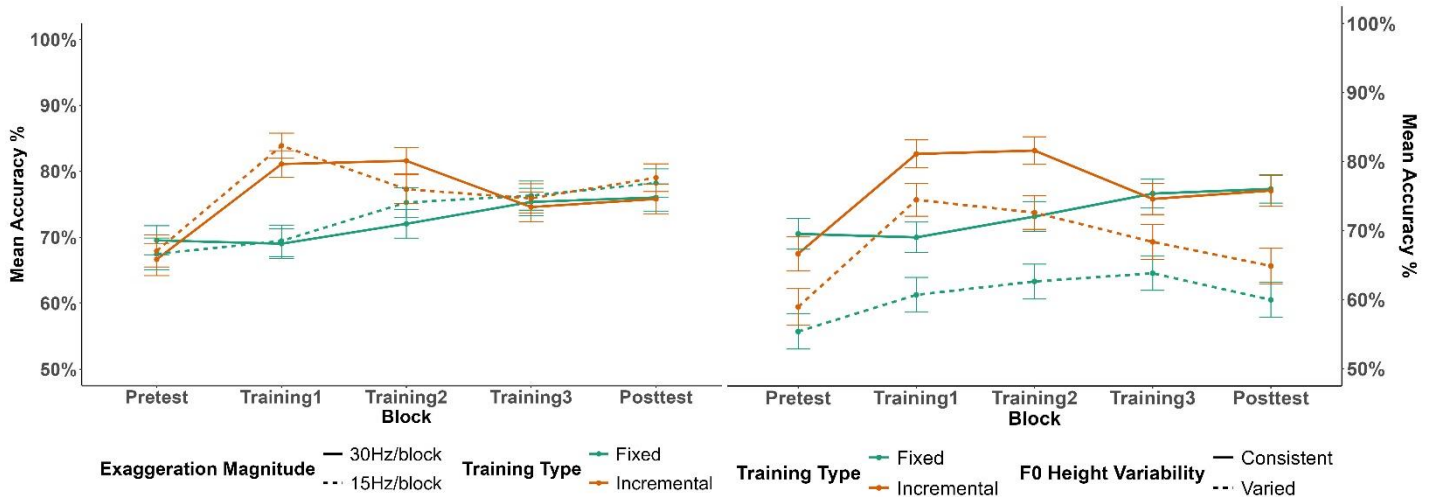


Figure 2. Average accuracy (%) for Study 1&2 (left: contrasting F0 exaggeration magnitude) and Study 1&3 (right: contrasting F0 height variability) by Training Type (Incremental & Fixed) in training blocks (Training 1, 2, 3) and test blocks (Pretest & Posttest). Study 1 is always in solid lines. Error bars indicate 95% confidence interval for accuracy means.

- [1] P. K. Kuhl and P. Iverson, 'Linguistic experience and the "perceptual magnet effect"', *Speech Percept. Linguist. Exp. Issues Cross-Lang. Res.*, pp. 121–154, 1995.
- [2] M. V. Kondaurova and A. L. Francis, 'The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods', *J. Phon.*, vol. 38, no. 4, pp. 569–587, Oct. 2010, doi: 10.1016/j.wocn.2010.08.003.
- [3] A. L. Francis, V. Ciocca, L. Ma, and K. Fenn, 'Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers', *J. Phon.*, vol. 36, no. 2, pp. 268–294, 2008.
- [4] E. M. Ingvalson, L. L. Holt, and J. L. McCLELLAND, 'Can native Japanese listeners learn to differentiate /r-l/ on the basis of F3 onset frequency?', *Biling. Lang. Cogn.*, vol. 15, no. 2, pp. 255–274, Apr. 2012, doi: 10.1017/S1366728911000447.
- [5] H. Pashler and M. C. Mozer, 'When does fading enhance perceptual category learning?', *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 39, no. 4, pp. 1162–1173, 2013, doi: 10.1037/a0031679.
- [6] Y.-R. Chao, 'A system of tone letters', *Maître Phon.*, 1930.

Indexing the attention orienting function of edge tones – the role of individual variability using pupillometry

Maria Lialiou¹, Martine Grice¹, Jesse A. Harris², Petra B. Schumacher¹

¹University of Cologne, ²UCLA

Intonational events are phonologically anchored to specific positions in the prosodic structure, forming either pitch accents (anchored to stressed syllables), or edge tones (anchored to edges of constituents). In the autosegmental-metrical (AM) theory of intonational phonology (e.g., [1]), edge tones have been strictly associated with a phrasing function, while pitch accents have been associated with a prominence-cueing function. However, two serial recall studies (e.g., [2-3]) have shown that rising edge tones lend prominence to the whole domain they delimit, calling this dichotomy into question. Building on these serial recall studies and on the recent finding that rises associated with the edges of constituents induce an attention orienting response in a similar way to accentual rises ([4]), the current study explores the role of individual variability in edge tones as an anchor for attention orienting, and thus prominence.

Using pupil dilation response (PDR) as a proxy for attention orienting, 60 native German listeners (54f; 6m) were presented with sequences of serially ascending numbers (*standards*; e.g., 21 22 23 24 25 26...) with occasional out-of-the-sequence numbers (*deviants*; e.g., 25 in 21 22 23 25 26...). Standard numbers featured shallow falling intonation (henceforth, neutral intonation), a pattern that is also found on non-final items of a sequence or list in German, while deviant numbers were produced with one of three intonational patterns: neutral intonation, domain-final rises, or domain-final falls. Domain-final rises and falls are functionally distinct: rises indicate continuity; falls denote finality (e.g., [5-8]). However, both types of edge tone can mark the end of units in a sequence. They thus fulfil a similar function. Given the expectancy violation basis of attention orienting (e.g., [9]), we expect that the presentation of a deviant will disrupt the anticipated pattern, capturing attention and thus inducing a PDR. Based on the reported enhanced orienting function of rising pitch (e.g., [4, 10-11]), we predict that a final rise in deviants will result in a more robust PDR compared to deviants produced with final falls or with neutral intonation. Individuals may vary in the use or control of attentional resources towards incoming information. Participants' individual variability was measured on the basis of three cognitive skills: processing speed, inhibitory ability, and working memory capacity (WMC), by implementing a version of the odd-man-out ([12]), flanker ([13]), and digit span ([14]) tasks, respectively.

The cognitive measurements yielded weak correlations, and thus PDR (normalised) was modelled only as a function of the ordered factor prosodic condition (treatment coded; rise/fall/neutral) and its interaction with the continuous flanker scores (reflecting participants' inhibitory ability), from 0 to 3sec after deviant onset in 10ms bins by fitting a generalized additive mixed model (GAMM). Fig 1 illustrates results of PDR over time as a function of prosodic condition (colour-coded) and flanker scores (from left [*lower inhibition*] to right [*higher inhibition*] panels). Fig 2 illustrates difference smooths (calculated as smooth A – smooth B) with 95% CIs for the comparisons among prosodic conditions as a function of flanker scores (x-axis). Individuals' inhibitory skills modulated PDR to deviants differently across prosodic conditions. Individuals with higher flanker scores, and thus better inhibitory skills, showed sustained PDRs only to rising deviants, as opposed to neutral and falling ones, with no difference between the latter two ([rise] > [fall = neutral]). In contrast, individuals with lower flanker scores, and hence weaker inhibitory skills exhibited sustained PDRs to both rising and falling PDRs, but not to the baseline neutral ([rise = fall > neutral]). The findings highlight individual differences in the allocation of attentional resources and suggest that edge tones effectively capture attention across cognitive profiles. This supports the view that tones at the edges of constituents can serve as attention-orienting devices, with implications for understanding for prominence and auditory attention.

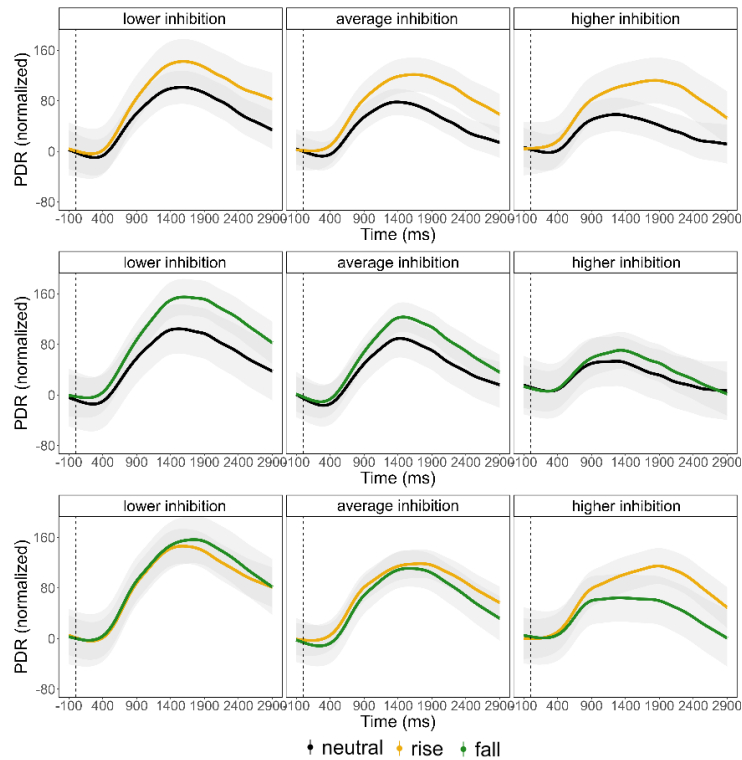


Fig1: GAMM smooths of PDR (normalised) with 95% CIs, across prosodic conditions and flanker scores. PDRs are time-locked to the onset of the deviant (zero ms), as indicated by the vertical dashed line.

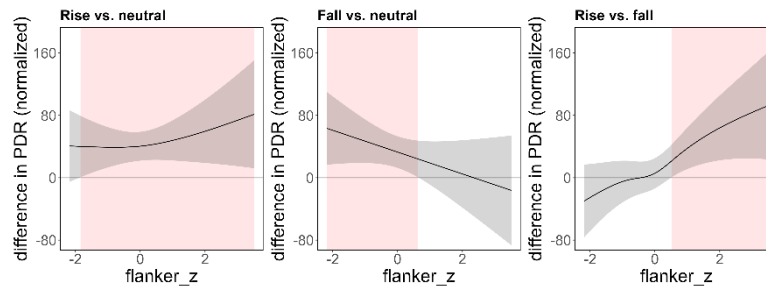


Fig 2: Difference smooths of PDR (normalised) with 95% CIs for the comparisons among prosodic conditions as a function of flanker scores. The red shaded areas depict where the 95% CI does not include zero, indicating the windows of significant differences.

References

- [1] Ladd, D. (2008). *Intonational Phonology*. 2nd edn. (Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.
- [2] Savino, M., Winter, B., Bosco, A., & Grice, M. (2020). Intonation does aid serial recall after all. *Psychonomic Bulletin & Review* 27(2). 366–372.
- [3] Grice, M., Savino, M., Schumacher, P.B., Röhr, C.T., & Ellison T.M. (to appear). Rises on Pitch Accents and Edge Tones Affect Serial Recall Performance at Item and Domain levels. *Laboratory Phonology*.
- [4] Lialiou, M., Grice, M., Röhr, C.T., & Schumacher, P.B. (2024). Auditory Processing of Intonational Rises and Falls in German: Rises Are Special in Attention Orienting. *Journal of Cognitive Neuroscience*. 1–24.
- [5] Grabe, E. (1998). *Comparative Intonational Phonology: English and German*. MPI Series in Psycholinguistics 7. Wageningen, Ponsen en Looien. (PhD Thesis).
- [6] Baumann, S., & Trouvain, J. (2001). On the prosody of German telephone numbers. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 557–560. ISCA.
- [7] Chen, A. (2003). Language Dependence in Continuation Intonation. In *Proceedings of the 15th International Congress of Phonetic Sciences*, 1069–1072.
- [8] Peters, J. (2018). Phonological and semantic aspects of German intonation. *Linguistik Online* 88(1).
- [9] Vachon, F., Hughes, R.W., & Jones D.M. (2012). Broken expectations: Violation of expectancies, not novelty, captures auditory attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38(1). 164–177.
- [10] Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, 42(4), 313–329.
- [11] Rinne, T., Degerman, A., & Alho, K. (2005). Superior temporal and inferior frontal cortices are activated by infrequent sound duration decrements: An fMRI study. *NeuroImage*, 26(1), 6672.
- [12] Frearson, W., & Eysenck, H.J. (1986). Intelligence, reaction time (RT) and a new "odd-man-out" RT paradigm. *Personality and Individual Differences* 7(6). Place: The Netherlands Publisher: Elsevier Science, 807–817.
- [13] Eriksen, B.A., & Eriksen, C.W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics* 16(1). Place: US Publisher: Psychonomic Society, 143–149.
- [14] Wechsler, D. (1987). *WMS-R: Wechsler Memory Scale-Revised : manual*. San Antonio: Psychological Corp.: Harcourt Brace Jovanovich.

The Effect of Cognitive Abilities on the Phonetic Accuracy of Intonation Imitation

Wenwei Xu, Yiya Chen

Leiden University Centre for Linguistics, Leiden University
w.xu@hum.leidenuniv.nl; yiya.chen@hum.leidenuniv.nl

There has been increasing attention to individual extralinguistic factors in the acquisition, perception, and production of speech [e.g., 1, 2]. Regarding cognitive abilities such as musical aptitude and working memory, mixed effects have been reported on speech imitation. For example, while musical aptitude or experience has been shown to enhance tone and intonation imitation [3, 4, 5], no such effect was reported in [6]. Similarly, working memory has been argued to facilitate tone and intonation imitation [3, 7], but its effect appears minimal in [4]. Notably, in these studies, imitation results are typically assessed using rating or labeling tasks [4-6] with only [3, 7] performing acoustic analyses. This reliance on metalinguistic perceptual assessments may obscure subtle phonetic variations, which can nevertheless be important for our understanding of the role of individual working memory capacity and musical aptitude in speech processing. This study aimed to address this issue further using an intonation imitation paradigm and detailed acoustic analysis of speech data.

33 Standard Chinese speakers (13F, all undergraduate students) participated in the study. In two blocks, they imitated pseudo sentences with synthesized intonation contours. Each contour features an utterance-final rise/fall or utterance-medial peak/trough, for which the pitch movement spans over either one or three syllables (**Figure 1**), akin to typical intonation events in West Germanic languages. In the first block, participants were asked to respond immediately, while in the second, they could only respond upon hearing a 500 ms' sine tone issued 2000 ms after the stimulus offset. After this task, participants completed an auditory N-back task and the Musical Ear Test (melody subtest) [8], which measure working memory and musical aptitude, respectively. Working memory is predicted to show a more robust effect (if any) in the immediate block, given that acoustic information usually decays in the memory after a 2 s' delay [9, 10]. Since all participants were L2 English learners, and their experience with a West Germanic language could potentially influence imitation performance, they were also asked to complete the LexTALE Test [11], which provided a measure of their English proficiency as a control factor.

Two sets of measures (global vs. local) were used to quantify imitation accuracy after time-normalized F0 was converted to ERB and standardized within speakers. Between each pair of targets and responses, the global accuracy was measured by the Euclidean distance of the standardized ERB (dERB) or its first derivative (dD1), as well as the Pearson correlation (PC) of the standardized ERB. The local accuracy was defined as deviations in the imitated height (Hd) and timing (Td) of a key pitch event in each imitated contour (e.g., for the final rise, the terminal pitch value and the start of the rise, respectively). For each of the five measures, linear mixed effects models were fitted with the individual factors as predictors, where interactions between the individual factors, intonation contours, and imitation blocks were included when they improved the model fit. The coefficients were interpreted after correcting *p*-values for multiple testing using Hommel's method (**Table 1**).

Regarding global accuracy, higher musical aptitude was associated with higher accuracy for the imitation of the long rise in the delayed block, shown by both dERB and PC. No significant coefficient was found for working memory or English proficiency for any of the measures. As for local accuracy, all the individual factors showed a significant effect on pitch height and/or timing in various conditions. More specifically, better memory was associated with higher accuracy in terms of pitch height within the immediate block, as well as local timing across contours and blocks. Furthermore, higher musical aptitude was associated with higher accuracy in local timing for the long rise in the immediate block. Finally, English proficiency showed a negative relationship with accuracy in local timing for the short fall in the delayed block.

The results suggest that working memory is useful for predicting imitation accuracy in terms of more subtle phonetic details such as height and timing of local pitch events, which impressionistic approaches might fail to capture. Moreover, the facilitative effect of working memory on imitating pitch height in immediate but not delayed imitation accords well with our prediction that better working memory boosts the imitation of acoustic information. Musical aptitude, on the other hand, seems more likely to modulate the global accuracy of imitated contours, which may explain why musical experience is relevant for perceived imitation accuracy in [4, 5]. Collectively, there is clear evidence that the relationship between imitation accuracy and the individual aptitude factors examined in this study also depends on what aspects of intonation patterns are imitated, suggesting a three-way interaction between target patterns, listeners' language experience, and their cognitive abilities.

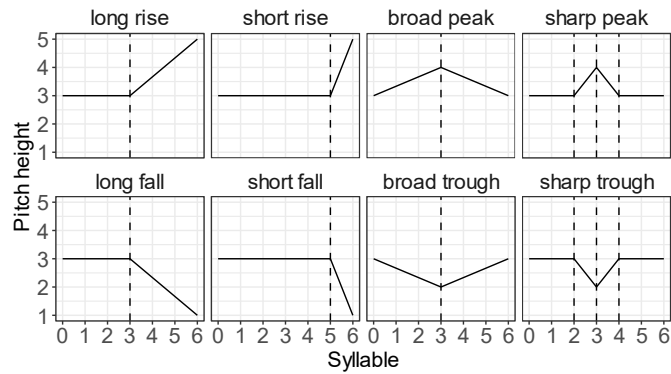


Figure 1. Intonation contours (schematic).

Table 1. Effects of individual factors on imitation accuracy.

Factor	Contour	Block	Measure	<i>p</i> -value	Effect
Working memory	(across)	Immediate	Height	.036	Facilitation
	(across)	(across)	Timing	.008	
Musical aptitude	Long rise	Delayed	dERB	.022	
			PC	.009	
		Immediate	Timing	< .001	
English proficiency	Short fall	Delayed	Timing	.008	Inhibition

References

- [1] Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, 130(1), 461–472.
- [2] Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, 11(2), 261–271.
- [3] Laméris, T. J., Li, K. K., & Post, B. (2023). Phonetic and phono-lexical accuracy of non-native tone production by English-L1 and Mandarin-L1 speakers. *Language and Speech*, 66(4), 974-1006.
- [4] Li, P., Zhang, Y., Baills, F., & Prieto, P. (2024). Musical perception skills predict speech imitation skills: differences between speakers of tone and intonation languages. *Language and Cognition*, 16(3), 647–665.
- [5] Pei, Z., Wu, Y., Xiang, X., & Qian, H. (2016). The effects of musical aptitude and musical training on phonological production in foreign languages. *English Language Teaching* 9, 19.
- [6] Toh, X. R., Lau, F., & Wong, F. C. K. (2022). Individual differences in nonnative lexical tone perception: Effects of tone language repertoire and musical experience. *Frontiers in Psychology* 13, 940363.
- [7] Petrone, C., d'Alessandro, D., & Falk, S. (2021). Working memory differences in prosodic imitation. *Journal of Phonetics*, 89, 101100.
- [8] Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and Individual Differences*, 20(3), 188–196.
- [9] Baddeley A. D. (2003). Working memory and language: an overview. *Journal of communication disorders*, 36(3), 189–208.
- [10] Zahner-Ritter, K., Einfeldt, M., Wochner, D., James, A., Dehé, N., & Braun, B. (2022). Three kinds of rising-falling contours in German wh-questions: Evidence from form and function. *Frontiers and Communication*, 7.
- [11] Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, 44(2), 325–343.

Intonational marking of focus types in Seoul Korean

Jiyoung Jang¹, Argyro Katsika²

¹*Hanyang Institute for Phonetics and Cognitive Sciences of Language;*

²*University of California, Santa Barbara*

Key words: intonation, focus structure, information structure, edge-prominence, Seoul Korean

Intonation is widely recognized as a key factor in signaling prominence at the phrase level, primarily through pitch movements that are associated to linguistic elements that are considered rhythmically or conceptually important. Languages differ in how they use these pitch movements to mark prominence: The prominence-marking pitch movements align with the heads of phrases in head-prominence languages, and are called pitch accents, while they align with the edges of prosodic units in edge-prominence languages, and are thus considered edge tones [1]. Previous research suggests that when these prominence markers denote focus, they are not simply binary, corresponding to a mere distinction between focused and unfocused elements, but instead reflect a hierarchical organization that corresponds to focus structure (e.g., [2-6]). Specifically, pitch accents have been found to vary in terms of F0 and other correlates of prominence, such as articulatory strengthening, depending on the type of focus being expressed—ranging from broad to narrow to contrastive focus—indicating that prominence is a multi-dimensional and multi-level system. The current study extends this investigation to Seoul Korean, an edge-prominence language. Unlike languages with lexical stress, where phrasal prominence is marked by pitch accents on heads, i.e., stressed syllables [1], Seoul Korean marks prominence by the means of Accentual Phrases (APs), the underlying tonal pattern of which has been modelled as [LH...LH]¹. The focused word is assumed to start an AP or a higher phrase, with subsequent AP boundaries up to the end of the Intonational Phrase (IP) often undergoing attenuation or elimination, referred to as dephrasing [7]. Given that the focused word is at the beginning of a (possibly longer than common) AP, we aim to determine whether focus types are phonetically distinguished in their F0 contours of the AP-initial LH rise. In addition, we examine if a hierarchy of prominence emerges, where F0 systematically increases from broad, to narrow, and to contrastive focus, similar to that observed in head-prominence languages.

To investigate this question, pairs of prompt and test sentences were designed to elicit five focus types in IP-medial target words: contrastive focus (CF), narrow focus (NF), broad focus (BF), unfocused post-focally with unfocused post-focally with narrow focus on a preceding word (UN), and contrastive focus on a preceding word (UC) (Table 1). Two target prosodic words consisting of three syllables (/mapu-lul/, /napi-lul/) were used. Data from 5 native speakers of Seoul Korean (1F, 4M) were analyzed. In total, 383 tokens were analyzed. F0 was extracted using the Straight algorithm in VoiceSauce [8] and z-scored by speaker. Data were analyzed in R [9] using Generalized Additive Mixed Models [10] with a by-level factor smooth for *Focus Type* and random reference and difference smooths for *Speaker*, *Place of Articulation* (labial vs. alveolar word-initial consonant), and *Repetition*.

The results revealed a significant effect of *Focus Type* on F0, with focus types showing distinct pitch patterns across conditions. F0 variation did not simply distinguish in-focus from out-of-focus elements, but rather differentiated between multiple levels of focus. Three distinct levels of F0 prominence were observed: (1) CF and NF, (2) BF, (3) UN and UC. As shown in Figure 1, CF and NF exhibited a significantly greater expansion of the rise, with higher F0 values for H compared to BF (CF–BF: $edf = 6.646$, $F = 3.757$, $p < 0.001$; NF–BF: $edf = 6.710$, $F = 2.928$, $p < 0.01$), which itself displayed higher F0 values than UN and UC (BF–UN: $edf = 5.496$, $F = 6.083$, $p < 0.001$; BF–UC: $edf = 5.317$, $F = 7.086$, $p < 0.001$), which are the conditions assumed to undergo dephrasing.

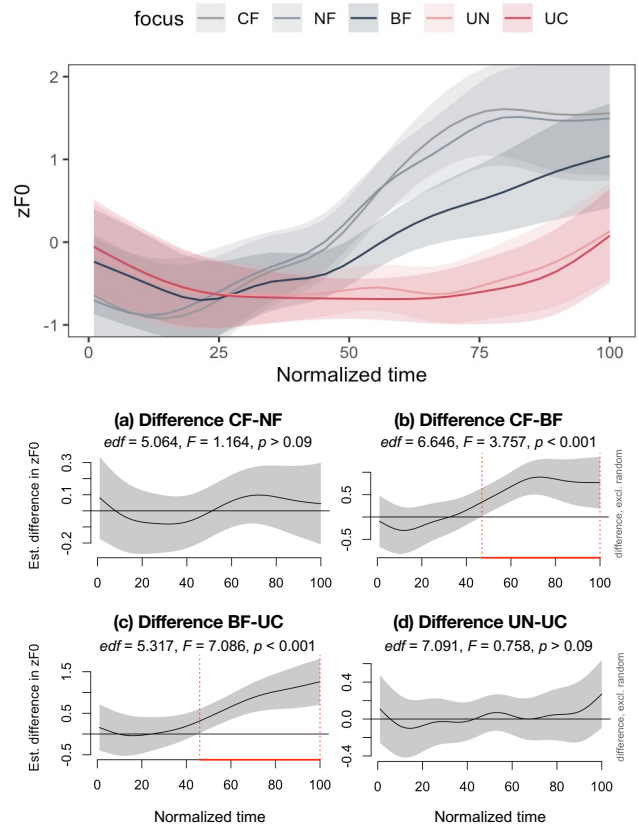
These results suggest that focus type plays a key role in modulating pitch, with the most prominent focus types (CF, NF) associated with the greatest F0 excursions of the H tone, followed by BF, and the least prominent types (UN, UC) exhibiting the least increase in F0 contours. These findings align with findings on articulatory correlates of Seoul Korean [11], and together indicate that Korean encodes focus structure prosodically. The hierarchy of focus types that emerge phonetically is similar to that found in head-prominence languages (e.g., [2-6]), suggesting a multi-level and multi-dimension system of prominence and an interaction between prosodic structure and information structure that may be a cross-linguistic feature, shared across different types of the prosodic typology.

¹ The realization of the AP-initial tone depends on the laryngeal configuration of the initial segment: H tone for aspirated, fortis, and fricatives; otherwise, L tone (see [1,7]).

Table 1. Sample stimuli per *Focus Type* for target word /mapu-lul/. Target words are underlined and focused words are in bold.

Focus	Example sentences
CF	A: ‘Did Minam visit the farmer?’ B: mi.nam.i.ka <u>ma.pu</u> .lul paŋ.mun.he.s* ^Λ Minam-NOM horseman -ACC visit-PAST ‘Minam visited the horseman .’
NF	A: ‘Who did Minam visit?’ B: mi.nam.i.ka <u>ma.pu</u> .lul paŋ.mun.he.s* ^Λ Minam-NOM horseman -ACC visit-PAST ‘Minam visited the horseman .’
BF	A: ‘What happened?’ B: mi.nam.i.ka <u>ma.pu</u> .lul paŋ.mun.he.s* ^Λ Minam-NOM horseman-ACC visit-PAST ‘Minam visited the horseman .’
UN	A: ‘Who visited the farmer?’ B: mi.nam .i.ka <u>ma.pu</u> .lul paŋ.mun.he.s* ^Λ Minam -NOM horseman-ACC visit-PAST ‘ Minam visited the horseman.’
UC	A: ‘Did Junseok visit the farmer?’ B: mi.nam .i.ka <u>ma.pu</u> .lul paŋ.mun.he.s* ^Λ Minam -NOM horseman-ACC visit-PAST ‘ Minam visited the horseman.’

Figure 1. (Top) GAMM smooths by *Focus Type*. (a-d) Estimated differences. (Note: difference plots for NF–BF, BF–UN are omitted due to page limits.)



References

- [1] Jun, S. A. (2005). *Prosodic typology: The phonology of intonation and phrasing*. OUP Oxford.
- [2] Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, 25(7-9), 1044-1098.
- [3] Hermes, A., Becker, J., Mücke, D., Baumann, S., & Grice, M. (2008). Articulatory gestures and focus marking in German. In *Proceedings of Speech Prosody 2008*, 457-460.
- [4] Katsika, A., Jang, J., Krivokapić, J., Goldstein, L., Saltzman, E. (2020). The role of focus in accental lengthening in American English: Kinematic analyses. *Proceedings of Speech Prosody 2020*, Tokyo, Japan.
- [5] Katsika, A., Jang, J., Krivokapić, J., Goldstein, L., & Saltzman, E. (2023). A hierarchy of prominence: The production and perception of focus in American English. In *Proceedings of the International Congress of Phonetic Sciences* (pp. 1696-1700).
- [6] Roessig, S., & Mücke, D. (2019). Modeling dimensions of prosodic prominence. *Frontiers in Communication*, 4, 44.
- [7] Jun, S. A. (1993). *The Phonetics and Phonology of Korean*. PhD dissertation, The Ohio State University.
- [8] Shue, Y.-L., Keating, P. A., Vicens, C., & Yu, K. (2011). VoiceSauce: A Program for Voice Analysis. *Proceedings of International Congress of Phonetic Sciences*, 1846-1849.
- [9] R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- [10] Wood, S. (2019). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation version 1.8-38 from CRAN (1.8-38) [R]. <https://rdrr.io/cran/mgcv/>.
- [11] Katsika, A. & Jang, J. (2024). Prosodic encoding of focus and edge-prominence: an articulatory study of Seoul Korean. *Proceedings of the 19th Conference on Laboratory Phonology* (pp. 55-56), Seoul, South Korea.

Perception compared to clustered intonation variation in German *wh*-questions

Heiko Seeliger, Constantijn Kaland
University of Cologne

Recent intonation research shows that cluster analysis on F0 contours is a valuable method, e.g. [1-4]. However, the results of cluster analysis have only rarely or indirectly been investigated in perception studies [5-8]. The initial cluster analysis underlying the current study investigated variation in the boundary tones of nuclear contours of German *wh*-exclamatives and *wh*-questions ([2], [9]). This was done in two separate cluster analyses: one clustering of the whole F0 contour and one clustering of only the contour of the final two syllables. This approach was able to distinguish phonologically meaningful contour classes. That is, we were able to separate high-level plateaus from late falls (henceforth: Level and Fall). The Falls are of particular interest as they are not in the GToBI [10] core inventory of commonly occurring nuclear contours. However, two clusters appeared to consist of both Level and Fall contours, casting doubt on how well these can be separated. These mixed clusters were singled out for further perception analysis [5], eliciting native listeners' judgments of contour similarity. While contours were indeed judged to be more similar for within-cluster pairs than for between-cluster pairs, the results also showed an influence from differences in utterance duration and in F0 register, which had strong negative correlations with perceived similarity. The current study aims to control for these factors in order to enable a more systematic comparison of contour pairs and their potentially meaningful differences in perception.

We made the following adjustments to the stimuli and methodology of the present study: (i) selection of 10 contours (*wh*-questions only), 5 from each cluster (down from 10 from each cluster), (ii) systematic presentation of every possible combination, including of identical stimuli (instead of mostly between-cluster combinations), (iii) duration normalization of every contour to the mean duration of all 10 contours, (iv) F0 normalization, by shifting the median F0 of every contour to the median of all F0 values of the 10 contours. F0 range was not normalized, since differences in F0 range were also visible to the original cluster analysis reported in [2] (duration and F0 median also being normalized there). The stimuli were overall balanced for speaker gender, with a slight imbalance within clusters (Fall: 2 male, 3 female; Level: 3 male, 2 female). The experimental task was otherwise identical to the one reported in [5]: Hummed speech contours were played in pairs. Participants (native speakers of German; $n = 27$) rated the perceived similarity of each stimulus pair on a five-point scale, whose end points were labeled 'identisch' (identical) and 'unähnlich' (dissimilar). Every listener contributed 55 similarity ratings (45 non-identical + 10 identical contour pairs).

In order to test whether the contours in the stimuli ($n = 10$) match the Level and Fall distinctions found in [2], we performed a cluster analysis on this subset (Fig. 1). Three late-falling contours (originally all in the Fall cluster) formed a cluster that did not merge or split with/into other clusters when three clusters were used in the analysis. This suggests that the late falls constitute a distinct class of their own. The perception ratings are shown in Fig. 2. The results are split by between/within cluster comparisons for the subset analysis. Comparisons of identical stimuli are plotted separately, as these were judged to be extremely similar across the board. Original cluster membership had no significant impact on perceived similarity, but membership in the subset clusters did: Contours from the same subset cluster were judged to be more perceptually similar than contours from different subset clusters (LMM, $p < 0.05$). In other words, the late-falling contours also form a distinct class in perception, not just in the subset cluster analysis. A tendency was shown in that the late-falling contours received the highest mean rating for non-identical comparisons.

Overall, we take the following main findings from this study: (i) The original clusters, Fall and Level, were indeed 'mixed', also in terms of perception. Since the original analysis clustered over 1100 contours, it might well be expected that not every cluster is completely 'pure' (in terms of, e.g., nuclear contour). (ii) Performed on this subset of stimuli, cluster analysis is successful in separating perceptually distinct F0 contours. (iii) As already indicated by the findings of [5], naive listeners are able to perform the fairly abstract task of assigning similarity ratings to pairs of hummed contours. The ratings correlate with acoustic differences in F0 range and delta-F0 (range: $F0_{max} - F0_{min}$; delta: mean of absolute differences between successive F0 points) in the latter two quarters of the contour. This suggests a sensitivity to the nuclear part of the contours, despite the delexicalized nature of the stimuli (see Fig. 3). Taken together, the results add a fine-grained analysis that confirms the outcomes of our earlier work, suggesting a nuclear intonation contour for German that has not been accounted for by the GToBI core inventory.

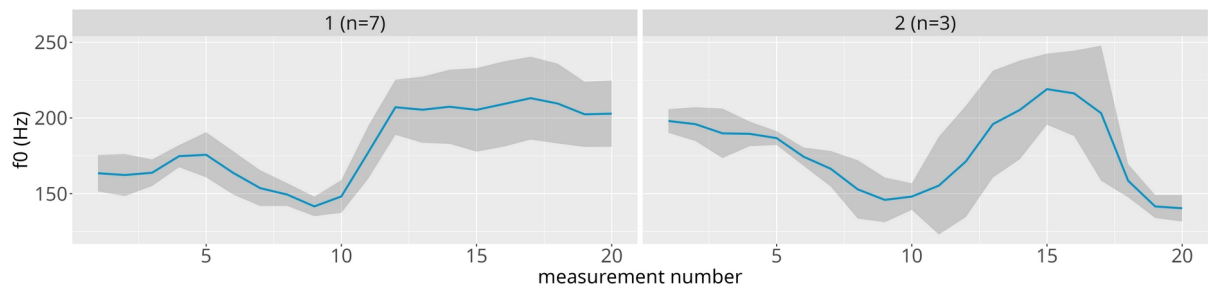
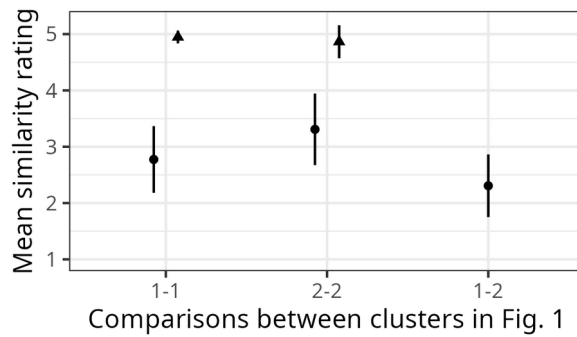
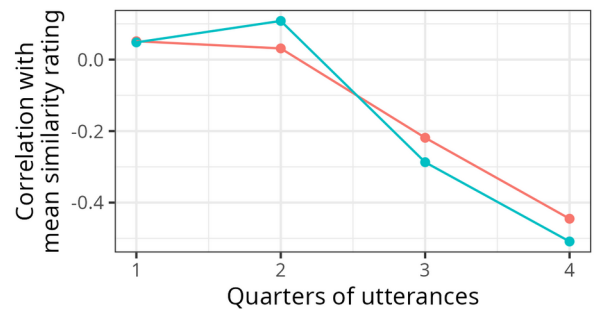


Figure 1: Clusters identified in the stimuli subset. Late-falling contours are shown on the right.



Identical stimuli • no ▲ yes



Abs. difference between contour pairs in:

— Delta F0 — F0 range

Figure 2: Mean ratings and standard deviations for contour comparisons split by cluster membership

Figure 3: Correlations between (i) two measures of acoustic differences between contours and (ii) mean similarity ratings, split up by quarters of utterances

References

- [1] Kaland, Constantijn. 2023. Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *Journal of the International Phonetic Association* 53(1), 159–188.
- [2] Seeliger, Heiko & Kaland, Constantijn. 2022. Boundary tones in German *wh*-questions and *wh*-exclamatives – a cluster-based approach. In *Proc. Speech Prosody 2022* (27–31).
- [3] Kaland, Constantijn; Steffman, Jeremy & Cole, Jennifer. 2024. K-means and hierarchical clustering of f0 contours. *Proc. Interspeech 2024* (1520–1524).
- [4] Jeon, Hae-Sung; Kaland, Constantijn & Grice, Martine. 2024. Cluster analysis of Korean IP-final intonation. In *Proc. Speech Prosody 2024* (1025–1029).
- [5] Seeliger, Heiko; Lützel, Anne & Kaland, Constantijn. 2023. The perception of German *wh*-phrase-final intonation: a contour clustering evaluation. In *Proc. 2nd International Conference on Tone and Intonation (TAI 2023)* (10–14).
- [6] Cole, Jennifer; Steffman, Jeremy; Shattuck-Hufnagel, Stefanie & Tilsen, Sam. 2023. Hierarchical distinctions in the production and perception of nuclear tunes in American English. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 14(1), 1–51.
- [7] Kaland, Constantijn. 2023. Intonation contour similarity: f0 representations and distance measures compared to human perception in two languages. *Journal of the Acoustical Society of America*, 154(1), 95–107.
- [8] Collier, René. 1975. Perceptual and Linguistic Tolerance in Intonation. *International Review of Applied Linguistics in Language Teaching*, 13(1–4), 293–308.
- [9] Repp, Sophie & Seeliger, Heiko. Under review. Contrast + givenness, local + non-local. The influence of complex information-structural settings on the prenuclear, nuclear and post-nuclear regions in exclamatives and questions. *Language and Speech*. (Pre-print: 10.31124/advance.171232666.64045580/v1)
- [10] Grice, Martine; Baumann, Stefan & Benz Müller, Ralf. 2005. German intonation in autosegmental-metrical phonology. In Jun, Sun-Ah (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*, Oxford University Press, 55–83.

F0 correlates of perceived speaker surprise in American English: Accents vs. Edge Tones

Rebekah Stanhope, Thomas Sostarics, and Jennifer Cole | Northwestern University

rstanhope@u.northwestern.edu, TSostarics@u.northwestern.edu,

jennifer.cole1@northwestern.edu

Background. Previous work has proposed that the L + H* pitch accent is the primary encoding of speaker surprise in English, with variation in the scaling of this accent potentially contributing additional ‘emphasis’ or other expressive effects [1]. However, other studies have linked perceived speaker surprise to expanded pitch range [2] and high pitch level across the entire utterance [3, 4], suggesting that listeners’ perceptions of speaker surprise may also be influenced by continuous variation in the scaling of other pitch accents, such as L* and H*, as well as intonational features at other parts of the contour, e.g. edge tones. In this study, we investigate how perceived surprise varies according to the phrase-final (nuclear) intonation of the utterance, from the rightmost pitch accent to the end of the intonational phrase, comparing the relative contributions of pitch accent and edge tone scaling. Additionally, we attempt to replicate previous findings that pitch range and level impact perceived surprise. **Methods.** Participants (n = 55) listened to declarative utterances such as *Megan’s a grandma*, resynthesized to vary F0 in the nuclear region, and rated how surprised the speaker sounded on a scale of 1 (not surprised) to 6 (very surprised). Participants rated 25 contours from a 5x5 phonetic continuum crossing the target F0 of the pitch accent (*accentual pitch*, ranging from 70-110 Hz, i.e. L* to H*) and the final F0 of the edge tones (*ending pitch*, 61-149 Hz, spanning L-L% to H-L% to H-H%). Each contour was repeated five times, with each repetition instantiated by one of five different sentences. These materials are the same as those used in [5], except for the alignment of the pitch accent targets (see p.2). We modeled the likelihood of higher surprise ratings using a Bayesian mixed-effect ordinal regression model with accentual pitch, ending pitch, and their interaction as predictors (see additional detail in Table 1). We explored the influences of pitch range (i.e. size of the pitch excursion across the nuclear region) and pitch level (in terms of mean and max F0) on perceived surprise through separate models—see Table 1 for additional details. **Results.** The empirical data (Fig.1) show that ratings were highest when accentual pitch was low (Acc. Step 1) and ending pitch was high (End. Step 5), and lowest when both accentual and ending pitch were low. There is wide variation in ratings along the accentual pitch continuum spanning from L* to H*, indicating that perceptions of surprise are sensitive to variation in accents other than L + H*. We find positive credible main effects of accentual pitch ($\hat{\beta} = 0.11$, CI [0.08, 0.14]) and ending pitch ($\hat{\beta} = 0.28$, CI [0.22, 0.34]), reflecting that, overall, surprise ratings increase as these F0 targets increase, with the magnitude of this effect being greater for ending pitch. We also find a credible interaction ($\hat{\beta} = -0.03$, CI [-0.04, -0.02]) indicating that at higher ending pitch steps, increases in accentual pitch lead to decreases in mean rating (Fig. 2). These results show that in declarative sentences, the pitch accent is not the sole locus of the intonational encoding of surprise; the credible effect of ending pitch and credible interaction between ending pitch and accentual pitch both provide evidence that perceived surprise also varies with edge tone scaling. Our models exploring pitch range and level show credible positive main effects of pitch excursion size ($\hat{\beta} = 0.09$, CI [0.07, 0.11]), max F0 ($\hat{\beta} = 0.38$, CI [0.30, 0.48]), and mean F0 ($\hat{\beta} = 0.35$, CI [0.28, 0.44]), indicating that surprise ratings increase with these cues, in line with the findings of [2, 3, 4]. This result provides additional evidence to suggest that listeners’ perceptions of speaker surprise should be sensitive to multiple intonational features, potentially spanning the entire utterance, pointing to multiple directions for further inquiry. We suggest that future work investigate the contributions of prenuclear accents, and directly compare the effects of bitonal accents, including L + H*, to monotonal ones. **Keywords:** *intonational meaning, speech perception.*

Fig. 1 (right): Empirical mean surprise rating for each continuum step shown in color with numeric values in the top left corner. Stacked bars show the proportions of 1-6 ratings in each cell.

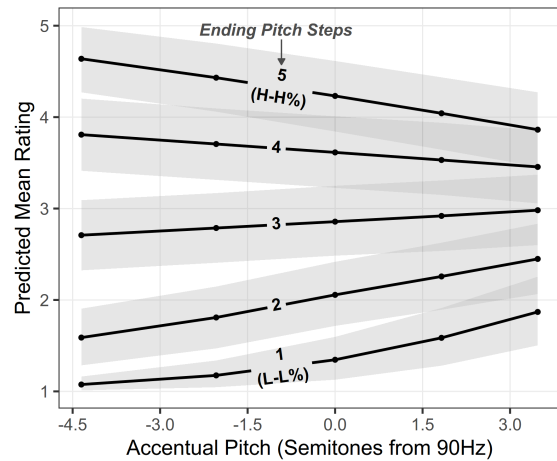


Fig. 2 (above): Model-predicted mean ratings for each ending pitch step as accentual pitch increases.

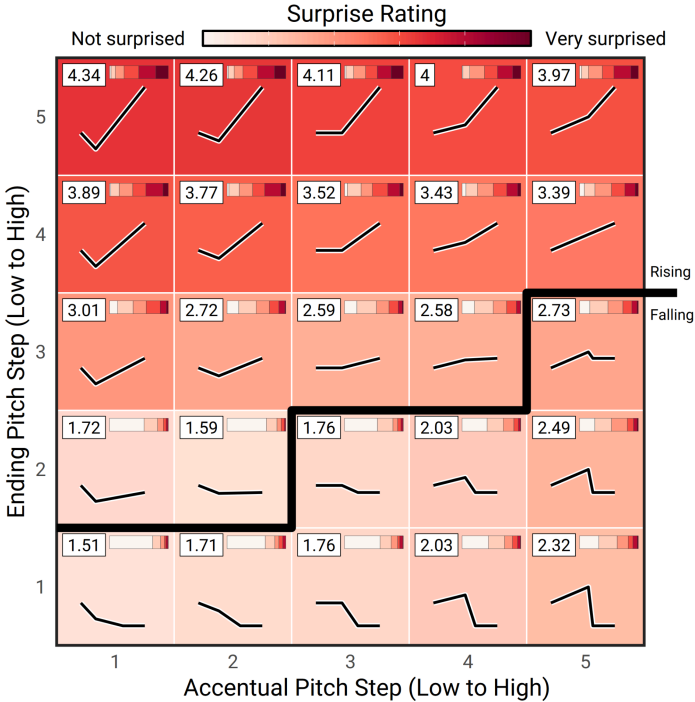
Pitch accent alignment: Alignment of the accentual pitch target varies in equidistant steps from 30% of the stressed syllable for the lowest target value (L*) to 100% of the stressed syllable (i.e. the end) for the highest target value (H*).

Table 1: Model descriptions, *contour shape* refers to whether a contour is *falling* or *rising* (see Fig.1).

Model name	Predictors and random effects <i>*Predictors were transformed to semitones from the midpoint of the accentual pitch continuum (90 Hz)</i>	Scale Parameter
Accentual and ending pitch	rating ~ accentual pitch * ending pitch + (1 + accentual pitch * ending pitch participant) + (1 utterance)	Contour shape (falling, rising) as a predictor; random slopes of shape by participant, random intercepts by participant and utterance
Pitch excursion	rating ~ excursion size * contour shape+ (1 + excursion size * contour shape participant) + (1 utterance)	
Mean F0	rating ~ mean F0 (in nuclear region) * contour shape + (1 + mean F0 * contour shape participant) + (1 utterance)	
Max F0	rating ~ max F0 (in nuclear region) * contour shape + (1 + max f0 * contour shape participant) + (1 utterance)	

References

[1] Rett, Jessica & Sturman, Beth. (2020). Prosodically marked mirativity. *Proceedings of WCCFL 38*. [2] Gussenhoven, C., & Rietveld, T. (2000). The behavior of H and L under variations in pitch range in Dutch rising contours. *Language and Speech*, 43(2), 183–203. [3] Ladd, D.R., Silverman, K., Tolkmitt, F., Bergmann, G., Scherer, K.R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect. *Journal of the Acoustical Society of America* 78, 435–444. [4] Liu, X. Xu, Y. Zhang, W. & Tian, X. (2021). Multiple prosodic meanings are conveyed through separate pitch ranges: Evidence from perception of focus and surprise in Mandarin Chinese. *Cognitive, Affective, & Behavioral Neuroscience*. 21. 1-12. [5] Sostarics, T., & Cole, J. (2023). Pitch Accent Variation and the Interpretation of Rising and Falling Intonation in American English. In Proc. INTERSPEECH 2023 (pp. 97–101).



Exploring the foundations of intonational variation in (Multicultural) London English

Elisa Passoni¹, Sam Hellmuth² and Joe Pearce² (QMUL¹/University of York²)

Introduction: Recent work on London English has documented a highly systematic new phonological system in inner city areas. This new variety, Multicultural London English (MLE [1, 2]), is argued to have arisen out of intensive, multiethnic social contact, initiating a wave of structural innovation 30 years ago [3]. Segmental properties of MLE have been extensively researched [1, 4], but, despite anecdotal reports of an ‘MLE intonation’ [5], this possibility has not yet been systematically explored. The only available description of inner city London intonation is a reported inventory of nuclear contours in read speech data from the *English Intonation in the British Isles* (IViE) project [6]; the IViE inner city London speakers shared a core inventory of nuclear contours with e.g. Cambridge speakers (fall H*L_%, fall-rise H*L_H%, rise-plateau L*H_H%), but two contours were near exclusively used in London (rise L*_H%, high-plateau H*_H%), in yes/no questions and declarative questions [7]. Grabe notes that whether “the different LH options for questions add nothing beyond degree of interrogativity to communicative impact [...] is an empirical question” (p. 21). The current paper: RQ1) revisits these generalisations using fPCA analysis of F0 applied to the same data and RQ2) explores for the first time patterns observed in unscripted data from the same speakers.

Methods: The IViE London data were collected at the end of the 1990s from 12 teenaged monolingual English speakers of Caribbean descent (6F/6M) with five tasks eliciting scripted (read sentences, read story) and unscripted speech (story retelling, map task and free conversation). Here we focus on read sentences (*dec/whq/ynq/dqn*), map task (*map*) and conversation (*con*). Following procedures devised for the *Generations of London English* project [8], unscripted data was transcribed and diarized in WhisperX [9] and checked for errors by the first author; for scripted data we used published task scripts. All data were force aligned in Praat TextGrid format using Montreal Forced Aligner (MFA) [10] at word and phone level with a GLE customized pretrained British English acoustic model/pronunciation dictionary. For RQ1 we drew on the original prosodic annotation labels published on the IViE website for half of speakers; for London we assume these are the 6 speakers analysed in [7]. IViE tone labels were manually accessed via the IViE website for the four sentence types (*dec/whq/ynq/dqn*) which all have a stress-initial disyllable as last lexical item, in which the preaccentual, accented and postaccentual syllables were manually annotated for these 4 sentence types x 12 speakers (N=202); 5 tokens were excluded due to a pause before the last lexical item. We extracted F0 in Hz at 10ms intervals through the three labelled syllables using a Praat script, plus timepoint of four landmarks: 1) preaccentual syllable start; 2) accented syllable start; 3) accented syllable end; 4) postaccentual syllable end. F0 was normalised by converting to semitones relative to each speaker’s minimum. We used mixed effects linear regression on principle components, derived from landmark registered fPCA with time-warping [11] on the time series data, to test: first, for the 6 speakers where we have IViE prosodic annotation labels (N=100), whether all five assigned *contour* labels are significantly differentiated; second, for all 12 speakers in the full dataset (N=197), whether cluster labels obtained from hierarchical clustering and/or k-means clustering on fPCA output [12] support analysis in terms of five distinct contours. We also identified *whq/ynq/dqn* in unscripted *con/map* using non-prosodic criteria (presence of wh-word/syntactic inversion and/or next-turn-proof interlocutor response), yielding 35 *ynqs*; 42 *whq*; 26 *dqn* (N=103); their nuclear contours were auditorily labelled by first and second authors.

Results: Figure 1 shows predicted fPCA curves for three PCs by assigned contour label in the IViE-labelled data subset; for all three PCs, there is no significant difference between fall H*L_% vs. fall-rise H*L_H%, nor between rise L*_H% vs. rise-plateau L*H_H%), though the high-plateau H*_H% is distinct. Figure 2 shows predicted fPCA curves for the full read speech dataset (N=197) by cluster label derived from k-means clustering (k=5); for all PCs, predicted curves for five clusters display overlap of some confidence intervals. Figures 3-4 show reconstructed F0 contours in registered time by sentence type and gender, for the five IViE labels in the labelled subset and for the k-means cluster label in the full dataset, respectively. The mapping of identified clusters in the full dataset roughly matches that reported by [7] for the labelled subset, but the number of distinct contours in inner city London English may be less than envisaged in that analysis. Our initial exploration of unscripted questions (Table 1) reveals i) similar parallel distribution of contours across *ynqs* and *whqs* as in read speech, though with many more falls, and ii) more varied realisation of *dqns*, but with frequent use of a rise-fall contour that was not noted in the analysis of the London read speech in [7].

Conclusion: Reanalysis of the IViE London data via fPCA offers an interim answer to the empirical question raised by [7] in showing an apparent distinction between only two types of rise in inner city London English question (which we characterise as a rise vs. high-plateau). In future work we will extend the analysis to include direct comparison of fPCA modelling of IViE London and Cambridge scripted and unscripted data.

Figure 1: Predicted fPCA curves by assigned contour label in IViE labelled data subset (N=100).

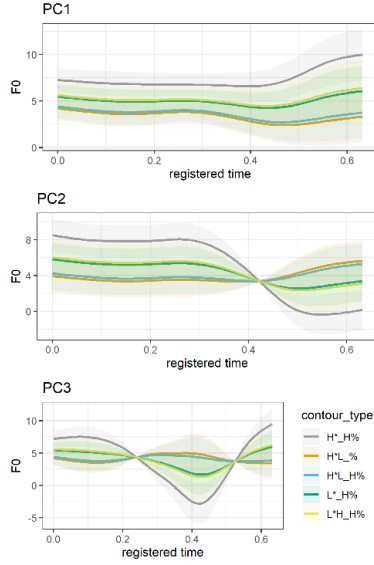


Figure 2: Predicted fPCA curves by k-means cluster label (k=5) in unlabelled full data subset (N=197).

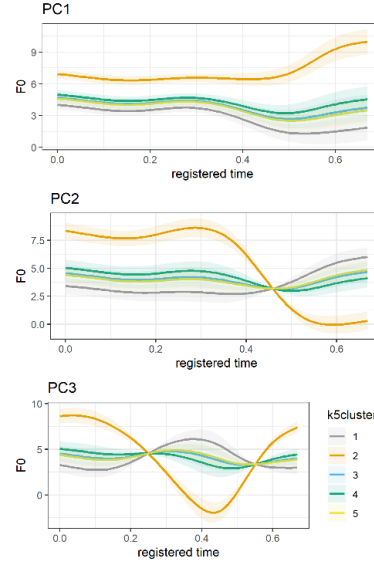


Figure 3: Reconstructed F0 contours by sentence type, published IViE contour label and speaker sex (N=100).

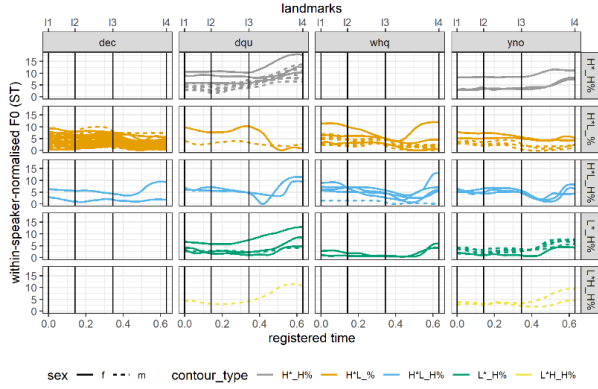


Figure 4: Reconstructed F0 contours by sentence type, k-means cluster label (k=5) and speaker sex (N=197).

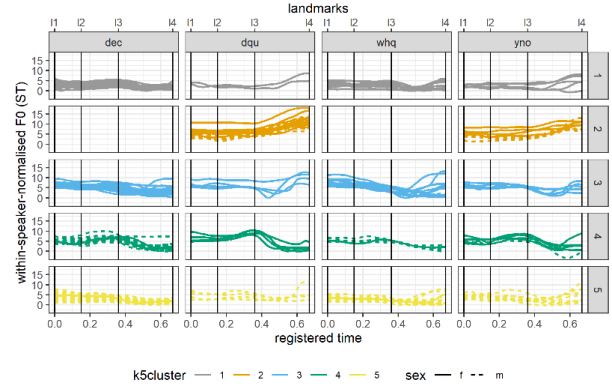


Table 1: Contours by sentence type in unscripted speech (N=103).

ynq	whq	dqn	contour:
66%	64%	27%	fall
3%	2%	4%	fall-rise
0%	5%	4%	rise-plateau
14%	7%	12%	rise
6%	2%	0%	high-plateau
3%	14%	54%	rise-fall
9%	5%	0%	no consensus

References

- [1] P. Kerswill, E.N. Torgersen, and S. Fox, "Reversing "drift": Innovation and diffusion in the London diphthong system," *Language variation and change*, vol. 20, no. 3, pp. 451-491. 2008.
- [2] J. Cheshire, P. Kerswill, S. Fox, and E. Torgersen, "Contact, the feature pool and the speech community: The emergence of Multicultural London English," *Journal of Sociolinguistics*, vol. 15, no. 2, pp. 151-196. 2011.
- [3] P. Kerswill and E. Torgersen, "Tracing the origins of an urban youth vernacular: Founder effects, frequency, and culture in the emergence of Multicultural London English," in *Advancing Socio-grammatical Variation and Change*. Routledge, K.V. Beaman, I. Buchstaller, S. Fox, J.A. Walker, Editors, 2020. pp. 249-276.
- [4] P. Kerswill, A. Gibson, D. Sharma, K. McCarthy, E. Passoni, S. Hellmuth, "Individual profiles amidst a multiethnolect: acoustic heterogeneity in London English", *Journal of the Acoustic Society of America*, vol 154 (4) , A334, 2023
- [5] Pronunciation Studio (2008-2024) Multicultural London English – the Urban English Accent [Online]. Available: <https://pronunciationstudio.com/mle-multicultural-london-english-accent/>
- [6] F. Nolan and B. Post, "The IViE Corpus", *The Oxford Handbook of Corpus Phonology*, J. Durand, U. Gut, and G. Kristoffersen, Editors. OUP: Oxford, 2014, pp 475 - 485
- [7] E. Grabe, "Intonational variation in urban dialects of English spoken in the British Isles," in *Regional Variation in Intonation*, P. Gilles and J. Peters, Editors. Tuebingen: Niemeyer, 2004. pp. 9-31
- [8] D. Sharma, P. Kerswill, K. McCarthy, and S. Hellmuth, *Generations of London English: Language and Social Change in Real Time*, UKRI ESRC. 2023-2026.
- [9] M. Bain, J Huh, T. Han, A. Zisserman, WhisperX: Time-Accurate Speech Transcription of Long-Form Audio, in *Interspeech 2023*.
- [10] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. "Montreal forced aligner: Trainable text-speech alignment using kald," in *Interspeech*, 2017.
- [11] M. Gubian, F. Torreira, L. Boves, Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts, *Journal of Phonetics*, vol. 49, pp. 16-40, 2015.
- [12] C. Kaland, J. Steffman, and J. Cole. "K-means and hierarchical clustering of f0 contours," in *Proc. Interspeech 1520-1524*, 2024.

Prosody without intonation

Christian DiCanio (University at Buffalo) and Richard Hatcher (Hanyang University)

Keywords: intonation, tone, fieldwork, production, prosody

What does intonational or prosodic structure look like in complex lexical tone languages? In such languages there is typically less freedom for F_0 to be manipulated for indicating pragmatic meaning or phrasal constituency (Bennett and Elfner 2019, Connell 2017). In certain tone languages intonational edge tones may be found, e.g. Cantonese (Wong et al. 2005), while in others, pitch range expansion or register shift may be used prosodically to indicate narrow or contrastive focus, e.g. in Mandarin (Xu 1999), and/or finality, e.g. Yoloxóchitl Mixtec (DiCanio et al. 2018, 2021). Yet, might it be possible for a complex tone language to lack intonation altogether?

In this talk, we present phonological data and the results from two production studies investigating the realization of tones and syllables in Itunyoso Triqui (IT), an Otomanguean language spoken in Oaxaca, Mexico (DiCanio 2008, 2010). IT possesses nine lexical tones (/5, 4, 3, 2, 1, 43, 32, 31, 13/) and tone has a heavy morphological load within the language. The phonological data shows evidence for some, lower levels of a prosodic hierarchy. In particular, there is evidence of stress. Most phonological contrasts occur in the stem-final syllable, which is also lengthened, whereas pre-tonic syllables have fewer contrasts. Many phonological and lexical patterns also target a word-final iambic foot; and additional morphophonological patterns target a higher prosodic/morphological word. Despite evidence for low-level prosodic structure, none of the observed patterns involve continuous changes in F_0 .

To examine whether F_0 is used to encode phrasal constituency, our first field experiment investigated the production of IT tones in utterance non-final and final contexts following a repetition task. F_0 and duration values from 11 speakers productions of disyllabic words were examined. Words produced in utterance-final position were lengthened relative to those produced in non-final position. For level tones, no change in F_0 accompanied the change in utterance position. For falling tones, a greater F_0 fall was observed in utterance-final position than in non-final position. When normalized relative to the durational changes though, these differences preserved similar slope values on utterance-final syllables (see Figure 1). In sum, these findings argue for utterance-final prosodic lengthening with no associated boundary tone.

In our second experiment, we examined whether a more global pattern of F_0 declination occurs in the language. We constructed sentences consisting entirely of high tone /5/, mid tone /3/, and low tone /1/. F_0 and intensity values were extracted at 10 equidistant time points across each vowel for the same speakers who participated in the first experiment. All F_0 and intensity values for each utterance were examined along a single curve. Generalized additive mixed models reveal no evidence for F_0 declination across these tone-controlled utterances, though some evidence for a final pre-pausal drop in intensity was observed.

The results from this work demonstrates that (a) IT lacks intonational tones and (b) the grammatical system does not permit final F_0 lowering. If the presence of an intonational system is demonstrated primarily by the presence of pitch accents and boundary tones, IT seems to lack both of these. As such, it appears to be an exception to the universal *frequency code* arguing for declination as a cross-linguistic universal (Gussenhoven 2004). An obvious question in a language like this is just how pragmatic information is encoded. Where IT is lacking in intonational structure, it makes up for in its rather large set of 39 final pragmatic particles. Moreover, information structure in the language is largely handled within the morphosyntax (c.f. Kalinowski 2015, Torreira et al. 2014). What seems to distinguish tonal languages with boundary tones from those lacking them is the degree to which prosodic and pragmatic distinctions have been grammaticalized elsewhere. Itunyoso Triqui is a strong example of this. Yet despite evidence for some prosodic structure, there is no evidence for intonation in the language.

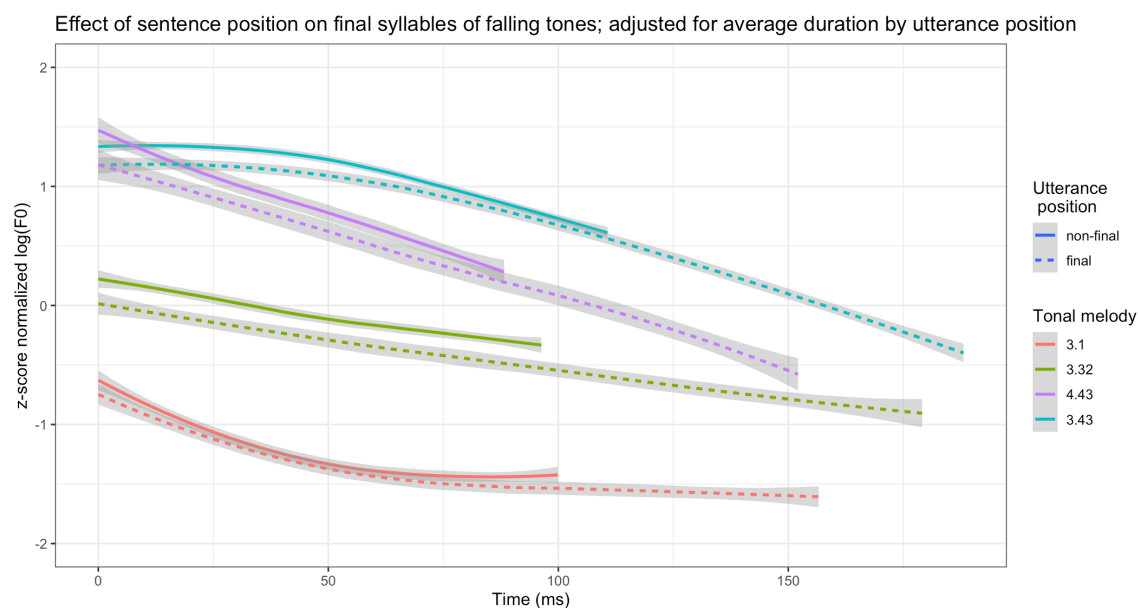


Figure 1: Itunyoso Triqui final falling tones in utterance-medial and utterance-final contexts. Note that the tonal contours here are the final syllables of the given melodies of disyllabic words, e.g. the [32] in /3.32/.

References

- Bennett, R. and Elfner, E. (2019). The syntax-prosody interface. *Annual Review of Linguistics*, 5:151–171.
- Clopper, C. G. and Tonhauser, J. (2013). The prosody of focus in Paraguayan Guaraní. *International Journal of American Linguistics*, 79(2):219–251.
- Connell, B. (2017). Tone and Intonation in Mambila. In Downing, L. J. and Rialland, A., editors, *Intonation in African Tone Languages*, pages 132–166. Berlin/Boston: De Gruyter.
- DiCanio, C. T. (2008). *The Phonetics and Phonology of San Martín Itunyoso Trique*. PhD thesis, University of California, Berkeley.
- DiCanio, C. T. (2010). Illustrations of the IPA: San Martín Itunyoso Trique. *Journal of the International Phonetic Association*, 40(2):227–238.
- DiCanio, C., Benn, J., and Castillo García, R. (2018). The phonetics of information structure in Yoloxóchitl Mixtec. *Journal of Phonetics*, 68:50–68.
- _____ (2021). Disentangling the effects of position and utterance-level declination on the production of complex tones in Yoloxóchitl Mixtec. *Language and Speech*, 64(3):515–557.
- Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Research Surveys in Linguistics. Cambridge University Press.
- Kalinowski, C. (2015). *A Typology of Morphosyntactic Encoding of Focus in African Languages*. PhD thesis, University at Buffalo.
- Torreira, F., Roberts, S. G., and Hammarström, H. (2014). Functional trade-off between lexical tone and intonation: typological evidence from polar-question marking. In *Proceedings of the 4th Tonal Aspects of Language Symposium*, pages 100–103. Nijmegen, the Netherlands.
- Wong, W. Y. P., Chan, M. K. M., and Beckman, M. E. (2005). An Autosegmental-Metrical analysis and prosodic annotation conventions for Cantonese. In Jun, S.-A., editor, *Prosodic typology*, chapter 10, pages 271–300. Oxford University Press.
- Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, 27:55–105.

A typological assessment of tonal languages in contact

Ricardo Napoleão de Souza

University of Helsinki, Svenska Handelshögskolan

The existence of so-called ‘tonal hotspots’ in Southeast Asia, Sub-Saharan Africa and Meso-America, where tone languages from different families converge (e.g. Yip 2002, Maddieson 2013), has contributed to the view that lexical tones diffuse easily in contact situations (e.g. Matisoff 2001, Ratliff 2007, among others). However, more rigorous assessments of tone systems often reveal a less unified picture (c.f. Brunelle & Kirby 2015, Kirby & Brunelle 2020). Similarly, developments in studies of tonogenesis (e.g. Thurgood 2002, Hyslop 2022) give support to language-internal explanations for how tone develops. To complicate matters further, the creole literature reports conflicting outcomes in cases when lexifier and substrate languages differ in the phonological status of lexical tones (e.g. Bordal Steien & Yakpo 2020). These issues are exacerbated by a lack of understanding of the actual mechanisms behind contact-induced change in word prosody systems (Napoleão de Souza & Sinnemäki 2022).

This study addresses these gaps by investigating how a series of phonetic and phonological variables related to lexical tones behave in contact situations. Using a sample specifically designed to assess contact effects (Di Garbo & Napoleão de Souza 2023), it evaluates how tonal properties change and/or diffuse, while controlling for inheritance factors in a principled way (Sinnemäki et al. 2024, see also Guy 2011, Witzlack-Makarevich et al. 2022).

Method. A database consisting of 50 pairs of languages in contact was scanned for instances in which tonal languages, including languages with restricted tone (i.e. ‘pitch accent’, Hyman 2006), were involved. In total, 12 contact situations encompassing at least one tonal language were identified: five pairs in Africa, three in the Americas, two in Asia, one each in Europe and Oceania. In the database, a language characterized as the recipient (e.g. Muak Doi [ukk, Austroasiatic]) is grouped with an unrelated donor language (e.g. Lü [khb, Tai-Kadai]), and with its closest relative not involved in the contact scenario (e.g. Pnar [pbv, Austroasiatic]) as a form of control.

The tone data were then collected from descriptive materials using a coding design based on typological work (e.g. Maddieson 2013, Jun 2014, Gordon 2016, see Sinnemäki et al. 2024). Variables analyzed include overall number of tones, number and shape of contour tones, laryngeal setting, lexical vs. grammatical function for tones, presence of stress, among others. Similarities between languages are computed by assigning binary values to individual properties. When the feature values in a contact pair are the same while differing from the control language (e.g. X/ X/ Y), we assign a one (1). When recipient and control languages have the same feature value (e.g. X/ Y/ X), we assign a zero (0). This scheme treats ‘ones’ as contact-induced change, whereas ‘zeroes’ represent either no change or independent developments. Values are then summed across features and divided by the total number of possibilities, yielding a ‘tone adaptation score’ (0-1) for every scenario. These scores are then turned into a continuous probability distribution (i.e. Beta distribution) to assess uncertainty.

Results & Discussion. Of the 36 languages analyzed (24 languages in contact, plus 12 control languages), 31% were tonal, 36% had restricted tone, and another 31% had stress. A preliminary analysis of adaptation scores yielded the following results. Surprisingly, half of the sample showed no contact-induced change in any tone properties where it was possible (i.e. 18 languages). This may suggest limited rates of diffusion for tonal properties. Indeed, tone adaptation scores obtained for the rest of the sample were overall rather low, with the grand mean being 0.3 for the remaining scenarios. Qualitatively, two changes were observed in more than one case: development of restricted tone in a language that previously had stress (2 instances), and the addition of a new tone contour in an erstwhile recipient tonal language. While these findings support the hypothesis that tones are added in contact varieties, the low rates of change seem to align with previous research that suggested that

similarities between tone systems in an area may be superficial. The full analyses with the Beta distributions will help clarify the latter point.

SAMPLE

Set ID	Language Name	ISO, Family	Role in Database
2	Bade	bde; AFA	<i>Recipient</i>
2	Manga Kanuri	kby; SAH	<i>Donor</i>
2	Lele	lln; AFA	<i>Control</i>
3	Mursi	muz; SUR	<i>Recipient</i>
3	Hamer-Banna	amf; SMC	<i>Donor</i>
3	Tennet	tex; SUR	<i>Control</i>
4	Kambaata	ktb; AFA	<i>Recipient</i>
4	Wolaytta	wal; TNO	<i>Donor</i>
4	Xamtanga	xan; AFA	<i>Control</i>
6	Nobiin	fia; NUB	<i>Recipient</i>
6	Beja	bej; AFA	<i>Donor</i>
6	Karko	kko; NUB	<i>Control</i>
8	Langi/Rangi	lag; ACG	<i>Recipient</i>
8	Alagwa	wbj; AFA	<i>Donor</i>
8	Zulu	zul; ACG	<i>Control</i>
13	South Saami	sma; URL	<i>Recipient</i>
13	Swedish	swe; IEU	<i>Donor</i>
13	Skolt Saami	sms; URL	<i>Control</i>
15	Yurok	yur; ALG	<i>Recipient</i>
15	Karok	kyh; KYH	<i>Donor</i>
15	Naskapi	nsk; ALG	<i>Control</i>
33	Muak Doi	ukk; ATC	<i>Recipient</i>
33	Lü/Tai Lue	khh; TAI	<i>Donor</i>
33	Pnar	pbn; ATC	<i>Control</i>
34	Burmese	mya; STB	<i>Recipient</i>
34	Mon	mnw; ATC	<i>Donor</i>
34	Kurtokha/Kurtöp	xkz; STB	<i>Control</i>
38	Hopi	hop; UAZ	<i>Recipient</i>
38	Zuni	zun; ZUN	<i>Donor</i>
38	Ute-Southern Paiute	ute; UAZ	<i>Control</i>
44	Kala Lagaw Ya	mwp; PMN	<i>Recipient</i>
44	Meryam Mir	ulk; ETF	<i>Donor</i>
44	Umpila	ump; PMN	<i>Control</i>
48	Yuhup	yab; NHP	<i>Recipient</i>
48	Macuna	myy; TCA	<i>Donor</i>
48	Nadëp	mbj; NHP	<i>Control</i>

SELECTED REFERENCES

- Bordal Steien, G., & Yakpo, K. (2020). Romancing with tone. *Language*, 96(1), 1-41.
- Brunelle, M., & Kirby, J. (2015). Re-assessing tonal diversity and geographical convergence in Mainland Southeast Asia. In *Languages of Mainland Southeast Asia: The state of the art*, pp. 82-110.
- Di Garbo, R. & Napoleão de Souza, R. (2023). A sampling technique for worldwide comparisons of language contact scenarios. *Linguistic Typology*, 3, 553-589.
- Hyslop, G. (2022). Toward a typology of tonogenesis. *Australian Journal of Linguistics*, 42(3-4), 275-299.
- Jun, S. (2014) Prosodic typology: by prominence type, word prosody, and macro-rhythm. In *Prosodic Typology II: The Phonology of Intonation and Phrasing*. Oxford.
- Matisoff, J. A. (2001). Genetic versus contact relationship: prosodic diffusibility in South-East Asian languages. In *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, 291-327.
- Napoleão de Souza, R., & K. Sinnemäki (2022). Beyond segment inventories: Phonological complexity and suprasegmental variables in contact situations. *Journal of Language Contact*, 15, 439-480.
- Sinnemäki, K., Di Garbo, F., Napoleão de Souza, R., & Ellis, M. (2024) A typological approach to language change in contact situations. *Diachronica*, 3, 379-413.
- Witzlack-Makarevich, A., Nichols, J., (...) Bickel, B. (2022). Managing AUTOTYP Data: Design Principles and Implementation. In *The Open Handbook of Linguistic Data Management*. Cambridge: MIT Press, pp. 632-642.

Culminativity and the Neutralization of Vowel Length in Four Neo-Štokavian Dialects

Elizabeth Zsiga (Georgetown University) and Draga Zec (Cornell University)

The Neo-Štokavian dialects (part of the Bosnian-Croatian-Serbian dialect continuum) are well known for their system of "pitch accents," in which words are traditionally described as having one of four prosodic shapes: long falling, short falling, long rising, and short rising (**Figure 1**). As early as Lehiste and Ivić (1986), the inventory of these accent types has been analyzed as the result of the interaction of three separate systems: stress, tone, and vowel length. While many studies have analyzed the interaction of stress and tone in these dialects (e.g., Inkelas & Zec 1988, Zec 1999, Smiljanić 2002), in this paper we present new data on the distribution of vowel length, the third parameter in the prosodic system. Crucially, we demonstrate that vowel length is lost when not coupled with either stress or H tone, an unusual interaction of different types of prominence.

Following Zec and Zsiga (2022), we assume that the place of stress depends on the place of H tone, at most one per word, which is lexically assigned. If the H tone is assigned to the initial syllable, stress and tone will coincide on that syllable, and the accent is termed "falling." If the H is assigned to a non-initial syllable, the preceding syllable will be stressed, and the accent is termed "rising." Possible patterns on tri-syllabic words are shown in **Figure 2**. This distribution of H tone and stress can be captured by positing a syllabic trochaic foot, which is the locus of culminativity: while stress falls on its head, H tone can be associated either with its head or its non-head. Lexically, vowel length is subject to the same principles of culminativity as are stress and tone. There can be at most one long vowel per word, which must occur in the head foot, either on its head or its non-head. Outside of the head foot, syllables cannot bear tone and cannot be long. Words demonstrating all possible combinations of stress, tone, and vowel length in di- and tri-syllables are given in **Figure 3**.

In our study, we investigate the realization of tone, stress, and vowel length in four Neo-Štokavian dialects: Novi Sad, Belgrade, Valjevo and Čačak. Twenty-eight adult female speakers took part in the study, seven from each dialect region. Materials consisted of the fifteen words in Figure 3. Each target noun was paired with a verb (in most cases, *imamo* 'we have') to create a two-word phrase. Then, for each phrase, a context sentence was created to prompt a response where the target word would be in focus. Responses were elicited with the target word in both phrase-initial and phrase-final position. For each target word, we measured the location of the pitch peak and the duration of all syllables. In earlier work, we have reported on the realization of H tones, showing that, in words with falling accent, the peak is consistently realized on the tonic syllable, while in words with rising accents, the location of the pitch peak varies by position and by dialect. H is always realized on the post-tonic syllable in phrase-initial position, and may retract to the tonic in phrase-final position, though the exact pattern of retraction differs by dialect.

Here we present the results for vowel length, shown in **Figure 4**. The stressed syllables, those corresponding to the foot head, are longer than post-stress, non-head syllables. Even with the increase in duration due to stress, the long/short distinction is maintained on all head syllables. On non-head syllables, the length distinction is maintained *only* if the syllable bears a lexical H. In non-head syllables that do not bear a H tone, the length distinction is neutralized. That is, the presence of vowel length crucially depends on the presence of either stress or H tone. Importantly, this interaction is only found at the lexical stratum, affecting lexical vowel length and its interactions with the other two prominences. In the postlexical stratum, however, where H tone retracts from a final syllable when the target word occurs in phrase-final position, vowel length is not affected. For example, even when the H tone retracts in /'vola:n^H/, the [a:] retains its length (mean 112 ms., vs. 77 ms for the [a] in /'jovan^H/).

We conclude that Neo-Štokavian pitch accent presents an unusual case of culminativity, with vowel length dependent not only on stress, but on tone. Further, the data provides evidence for the separation of the lexical and post-lexical strata of the phonology.

Figure 1: Four pitch accent types





	Falling		Rising	
a. Short		‘mother.acc.sg’		‘proper name.nom.sg’
b. Long		‘sea.nom.sg’		‘proper name.acc.sg’

Figure 2. Possible configurations of stress and tone

a. Falling	b. Rising	c. Rising
(‘o o) o	(‘o o) o	o (‘o o)
H	H	H

Figure 3. Words in the study

<u>description</u>	<u>traditional</u> <u>orthography</u>	<u>foot</u> <u>structure</u>	<u>representation</u>	<u>gloss</u>
<i>Falling accent: H on initial syllable</i>				
no long V	māmu	(‘cvHcv)	‘ma _H mu	mother
	növine	(‘cvHcv)cv	‘no _H vine	newspaper
long tonic	môre	(‘cv:Hcv)	‘mo: _H re	sea
	rêvije	(‘cv:Hcv)cv	‘re: _H vije	fashion shows
long post-tonic	nëmān	(‘cvHcv:)	‘ne _H ma:n	dragon
	nëmāni	(‘cvHcv:)cv	‘ne _H ma:ni	dragons
<i>Rising accent: H on final syllable</i>				
no long V	jòvan	(‘cvcv _H)	‘jova _H n	(proper name)
	ramèna	cv(‘cvcv _H)	ra‘mena _H	shoulders
long tonic	nénu	(‘cvcv _H)	‘ne:nu _H	(proper name)
	romāni	cv(‘cv:cv _H)	ro‘ma:ni _H	novels
long post-tonic	vòlān	(‘cvcv: _H)	‘vola: _H n	steering wheel
	juvèlir	cv(‘cvcv: _H)	ju‘veli: _H r	goldsmith
<i>Rising accent: H on medial syllable</i>				
no long V	màrame	(‘cvcv _H)cv	‘mara _H me	scarves
long tonic	mólove	(‘cv:cv _H)cv	‘mo:lo _H ve	shopping malls
long post-tonic	àrlje	(‘cvcv: _H)cv	‘ari: _H l’je	place name

Figure 4. Results

	Head H	Head no H	NonHead H	NonHead no H	outside of head foot
long	208 ms	195 ms	110 ms	77 ms	Ø
short	129 ms	135 ms	85 ms	76 ms	82 ms

Keywords: Tone, Pitch Accent, Vowel Length, Culminativity, Neo-Štokavian dialects**References:**

- Inkelas, Sharon & Draga Zec (1988) Serbo-Croatian pitch accent. Interactions of tone, stress, and intonation. *Language* 64: 227-248.
- Lehiste, Ilse & Pavle Ivić (1986) *Word and Sentence Prosody in Serbocroatian*. MIT Press.
- Smiljanić, Rajka (2002) *Lexical, Pragmatic and Positional Effects on Prosody in Two Dialects of Croatian and Serbian*. Ph.D. Dissertation. University of Illinois.
- Zec, Draga (1999) Footed tones and tonal feet: Rhythmic constituency in a pitch-accent language. *Phonology* 16: 225-264.
- Zec, Draga & Elizabeth Zsiga. (2022) Tone and stress as agents of cross-dialectal variation: the case of Serbian. In H. Kubuzono, J. Ito, and A. Mester (Eds.), *Prosody and Prosodic Interfaces*. Oxford University Press.

An evolutionary approach to grammatical tone and its relation to lexical tone

Sandra Auderset (University of Bern)

Keywords: grammatical tone, tone change, evolutionary phonology, Mixtec

In this talk, I take an evolutionary approach (Blevins 2004) to ‘grammatical’ tone and explore whether it can be distinguished from ‘lexical’ tone by diachronic processes. Typological and comparative approaches to tone often distinguish between lexical and grammatical tone based on functional criteria (Hyman 2016, Palancar 2016, Rolle 2018, Lionnet et al. 2022). Lexical tone is usually defined as contrastive pitch that distinguishes (at least some) morphemes. Grammatical tone, however, is not defined simply as the tone linked to grammatical morphemes, but rather as a ‘tonal operation that is not general across the phonological grammar’ (Rolle 2018: 1) or ‘tone which is assigned by the grammar’ (Hyman 2016: 6). Grammatical tone exhibits greater diversity in its functions and is less well-studied than lexical tone. Consequently, we do not yet have a good understanding of cross-linguistic distributions and general diachronic pathways (Caballero & German 2021). Drawing primarily on data from Mixtec languages (Mixtecan, Otomanguan; Mexico), I show that regular tone correspondences across lexical items and verbal paradigms can provide important clues about the emergence of ‘grammatical’ tone while at the same providing a diachronic criterion for distinguishing between tone patterns that are primarily used to make meaning distinctions and such that are associated with a morphosyntactic functions.

Mixtec languages provide an ideal starting point for such investigations because they are all tonal and tone can be reconstructed to the proto-language and most likely all the way back to Proto-Otomanguan (Dürr 1987). Mixtec varieties rely on tone for lexical distinctions and tone also plays an important role in verbal inflectional classes marking aspect, mood, and in some cases polarity (Palancar 2016, Palancar et al. 2016). Recent years have seen an uptake in descriptions of tonal AM-marking and its interaction with segmental markers and stem alternations in Mixtec languages (e.g. Costello 2014, Palancar et al. 2016, Swanton & Mendoza Ruiz 2021). The evolution of lexical tones from Proto-Mixtec to the modern varieties is relatively well understood (Dürr 1987, Swanton & Mendoza Ruiz 2021, Auderset 2024). I compare the tone correspondences and changes found in ‘lexical’ morphemes to those that arise in aspect-mood marking based on a data set that covers verb paradigms from 8 Mixtec varieties compiled from published sources as well as ongoing documentation projects. The data analyzed so far suggests that the imperfective-marking high tone has no segmental origin, as opposed to the perfective-marking low tone and the tone patterns appearing on the irrealis forms. However, the tone correspondences of the imperfective high tone are different from those of other high tones found on ‘lexical’ items, but seem to recur with the derivational high tone used in the formation of adjectives.

I compare these findings with what is known about the emergence of grammatical tone in other languages and language families (cf. Vydrin 2012 and Konoshenko 2023 on Mande, and Lovick & Tuttle 2024 on Dene). Preliminary results indicate that regular correspondences and diachronic pathways provide a way of understanding tones and their

function that does not rely on making reference to vaguely defined concepts like ‘general’ phonology or grammar.

References

- Auderset, Sandra. 2024. Rates of change and phylogenetic signal in Mixtec tone. *Language Dynamics and Change* 14(1)
- Caballero, Gabriela & Austin German. 2021. Grammatical Tone Patterns in Choguita Rarámuri (Tarahumara). *International Journal of American Linguistics* 87(2), 149-178.
- Chao, Yuen-Ren. 1930. ə sistim əv “toun-letəz” [a system of tone letters]. *Le maître phonétique* 8(30): 24–27.
- Costello, Rachael A. 2014. *Aspect and mood in Jicaltepec Mixtec*. MA thesis. Dallas: Graduate Institute of Applied Linguistics.
- Dürr, Michael. 1987. A preliminary reconstruction of the Proto-Mixtec tonal system. *Indiana* 11: 19–61.
- Konoshenko, Maria. 2017. Tone in grammar: What we already know and what we still don’t? *Voprosy Jazykoznanija* (4), 101–114.
- Lionnet, Florian, Laura McPherson & Nicholas, Rolle. 2022. Theoretical approaches to grammatical tone. *Phonology* 39(3), 385–398.
- Lovick, Olga, and Siri G. Tuttle. 2024. Pitch Patterns in Standard Negation in Alaskan Dene and the Development of Grammatical Tone. *International Journal of American Linguistics* 90(4), 397-444.
- Palancar, Enrique L. 2016. A typology of tone and Inflection: A view from the Oto-Manguean languages of Mexico. In Palancar & Léonard (eds.). *Tone and Inflection: New Facts and New Perspectives*. Amsterdam: De Gruyter Mouton
- Palancar, Enrique, Jonathan D. Amith & Rey Castillo García. 2016. Verbal inflection in Yoloxóchitl Mixtec. In: Palancar, Enrique L. and Jean Léo Léonard (eds.). *Tone and Inflection: New facts and perspectives*. Amsterdam: DeGruyter. 295-336.
- Rolle, Nicholas (2018). *Grammatical tone: typology and theory*. PhD dissertation, University of California, Berkeley.
- Swanton, Michael W. & Juana Mendoza Ruiz. 2021. Observaciones sobre la diacronía del tono en el tu’un Savi (mixteco) de Alcozauca de Guerrero Estudios Lingüísticos y Filológicos en Lenguas Indígenas Mexicanas. In: Arellanes, Francisco y Lilián Guerrero (eds.). *Estudios lingüísticos y filológicos en lenguas indígenas mexicanas: Celebración de los 30 años del Seminario de Lenguas Indígenas*, 309–345. México: Universidad Autónoma de México.
- Vydrin, Valentin. 2012. Aspektual’nye sistemy južnyx mande v diaxroničeskoj perspektive [Aspectual systems of Southern Mande in diachronic perspective]. *Acta Linguistica Petropolitana* 8(2). 566–647.

How merged are Cantonese tones in spontaneous speech?

Roger Yu-Hsiang Lo and Molly Babel

Department of Linguistics, University of British Columbia

Introduction: Recent literature describes the tonal system in Cantonese as undergoing several mergers (e.g., [10, 1, 6, 2, 13, 7]). For example, Fung and Lee [2] show that Tone 2 (T2) and T5 are fully merged with respect to fundamental frequency (F0) while T3-T6 and T4-T6 are described as partially-merged and nearly-merged, respectively. Importantly, however, these observations come from lab-elicited speech and therefore may not reflect the degree of merger observed in more natural communication. Nagy, Tse, and Stanford [7] represents the first systematic endeavour to investigate the F0 merger patterns in spontaneous Cantonese speech. Based on mean F0 values over the syllable, F0 values at a fixed reference point, and F0 slopes, they conclude that these three tone pairs are more merged in spontaneous speech than in lab speech. With their study as inspiration, this project also set out to explore the Cantonese tone mergers in spontaneous speech. Crucially, we innovate in the following directions: (1) instead of relying on “point” measurements as a proxy for the tonal trajectory, we incorporate the entire dynamic F0 trajectory in our analysis; (2) we consider syllable duration associated with the merging pairs, which is typically “normalized away”; and (3) we foreground individual differences in addition to population-level merger patterns, given that substantial variation at the individual level is expected for a language system undergoing changes.

Data: Data come from conversational interviews with 34 early Cantonese-English bilinguals (17 female, 17 male) in the SpiCE corpus [3]. Individual syllable boundaries were delimited using the Montreal Forced Aligner [5], and the STRAIGHT algorithm [4] implemented in VoiceSauce [9] was used to extract F0 in Hertz at 11 equal-distance points over the voiced portion (as estimated by Google REAPER [11]) of the syllable. We excluded tokens where either no voicing was detected or the ratio of maximum F0 to minimum F0 was greater than 2 to mitigate measurement errors due to pitch halving/doubling, which left us with a total of 99,416 tokens. The F0 values were converted into semitones, using individual talkers’ mean F0 in T1 from read speech from the SpiCE corpus. For duration, we used forced-aligned boundaries to compute it at the syllable level for the same tokens used in the above F0 analysis.

Analysis: We employ a generalized additive model (as implemented in the *mgcv* package; [12]) to model F0 trajectories, represented by the 11 equal-distance points, as a function of the linguistic and social variables considered by Nagy, Tse, and Stanford [7]. Specifically, we allowed F0 trajectories to vary in overall height in response to a set of fixed variables—TONE (T1-6), ONSET (sonorant, obstruent), SYLLABLE TYPE (checked, non-checked), PRECEDING TONE (T1-6, none), FOLLOWING TONE (T1-6, none), POSITION IN UTTERANCE (numeric in $[0, 1]$)—and random variables—CHARACTER, POSITION IN WORD, GENDER, and TALKER. We also allowed the shape and wigglyness to vary for each tone and talker. For duration, we fitted an approximate Bayesian mixed-effects model (using the *INLA* package; [8]) that predicts the logarithmic syllable duration based on TONE, SYLLABLE TYPE, PRECEDING TONE, FOLLOWING TONE, POSITION IN UTTERANCE, CHARACTER, POSITION IN WORD, and TALKER.

Results: The predicted F0 trajectories at the population level and for four example talkers are shown in Figure 1. At the population level, although the relative height between the tones follows the patterns of the citation forms, the overlapping confidence intervals for T2/T5 and T3/T4/T6 indicate that the tonal space for these tones might be similar. Indeed, predicted contours at the individual level reveal that talkers exemplify different (non-)merger patterns, reflecting the heterogeneity of the speech community. Figure 2 depicts the predicted syllable duration at the population level and for the same four talkers. Results at both levels suggest that merging pairs (i.e., T2-T5, T3-T6, T4-T6) still retain duration differences. Overall, the combined results of F0 and duration paint a picture where the neutralization between reported merging tone pair appears to be incomplete, at least in natural conversation.

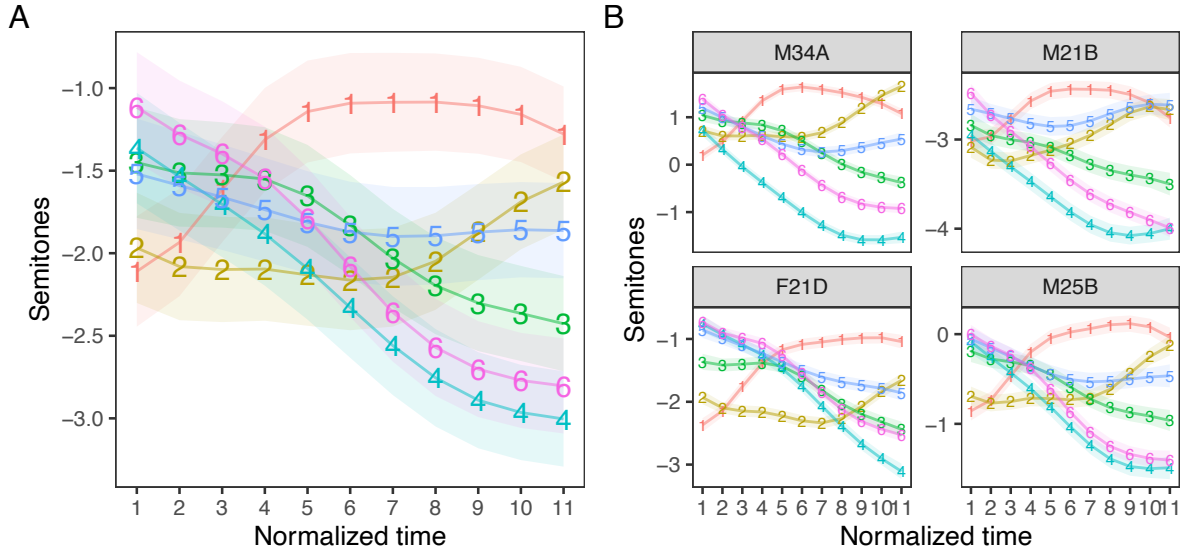


Figure 1: Predicted mean F0 contours for the six Cantonese lexical tones at (A) the population level and (B) the individual level for four talkers that typify no merger (M34A), the T2-T5 merger (M21B), the T3-T6 merger (F21D), and the T4-T6 merger (M25B). The shaded areas represent the 95% confidence interval.

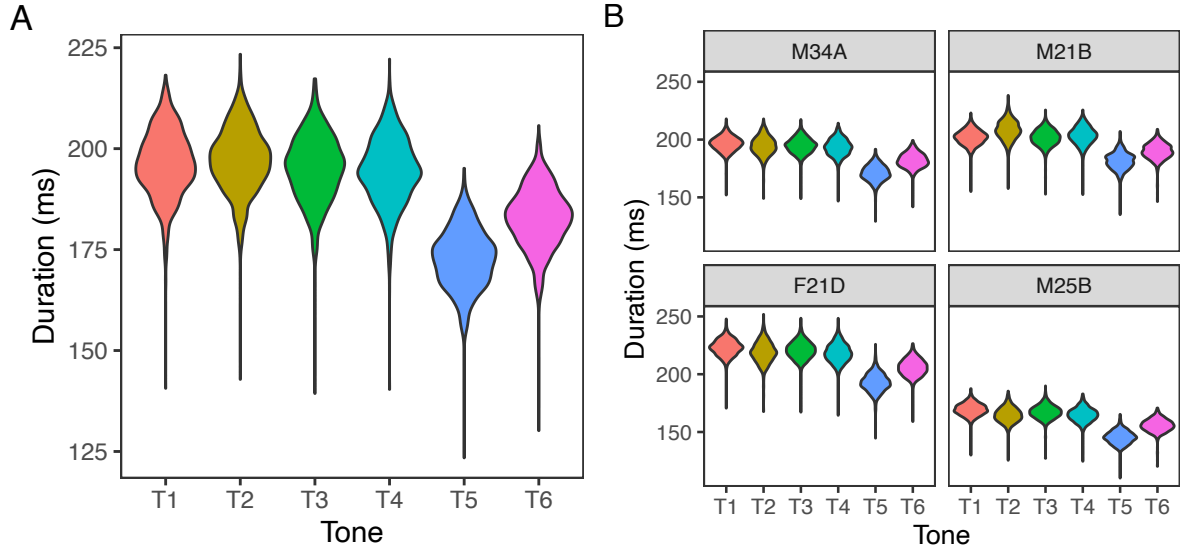


Figure 2: Predicted mean syllable durations for the six Cantonese lexical tones at (A) the population level and (B) the individual level for the same four talkers as in Figure 1.

References: [1] Robert S. Bauer, Kwan-hin Cheung, and Pak-man Cheung. “Variation and merger of the rising tones in Hong Kong Cantonese”. In: *LVC* 15 (2004). [2] Roxana S. Y. Fung and Chris K. C. Lee. “Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception”. In: *JASA* 146 (2019). [3] Khia A. Johnson. *SpiCE: Speech in Cantonese and English*. 2021. [4] Hideki Kawahara, Alain de Cheveigné, and Roy D. Patterson. “An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: Revised TEMPO in the STRAIGHT suite”. In: *The proceedings of the 5th ICSLP 98*. 1998. [5] Michael McAuliffe et al. “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi”. In: *Proceedings of INTERSPEECH 2017*. 2017. [6] Peggy P. K. Mok, Donghui Zuo, and Peggy W. Y. Wong. “Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese”. In: *LVC* 25.3 (2013). [7] Naomi Nagy, Holman Tse, and James N. Stanford. “Have Cantonese tones merged in spontaneous speech?” In: *The phonetics and phonology of heritage languages*. Ed. by Rajiv Rao. Cambridge: Cambridge University Press, 2024. [8] Håvard Rue and Sara Martino. “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations”. In: *J. R. Stat. Soc. Ser. B Methodol.* 72 (2009). [9] Yen-Liang Shue et al. “VoiceSauce: A program for voice analysis”. In: *The proceedings of the ICPHS XVII 2011*. 2011. [10] Lydia K.H. So. “Tonal changes in Hong Kong Cantonese”. In: *Current Issues In Language and Society* 3.2 (1996). [11] David Talkin. *REAPER: Robust Epoch And Pitch Estimator*. [12] Simon N. Wood. “Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models”. In: *J. R. Stat. Soc. Ser. B Methodol.* 73 (2011). [13] Jingwei Zhang. “Tone mergers in Cantonese: Evidence from Hong Kong, Macau, and Zhuhai”. In: *Asia-Pacific Language Variation* 5.1 (2019).

Tone merger in Datong Jin Chinese: Evidence from production and perception

Zhenyi Liao & Lei Liang (Nankai University)

Datong Jin Chinese is one of the representative areas of the Jin dialect in northern Shanxi province, known for its checked syllables with a glottal coda [ʔ]. Datong's tonal system consists of five tone categories: Yin Ping (T1), Yang Ping (T2), Shang Sheng (T3), Qu Sheng (T4), and Ru Sheng (T5). Prior impressionistic studies reported that the checked tone (T5) is experiencing the loss of final coda and tends to merge with non-checked tones due to similar F0 contours and tone values (Bai, 2001).

In our acoustic study, 24 native speakers (17F) were recruited from the urban area of Datong city and were divided into three groups: younger group (N = 6, 4F, Mean = 23.17), middle-aged group (N = 8, 7F, Mean = 44.88), and older group (N = 10, 6F, Mean = 66.7). Figure 1 shows the F0 trajectories of the lexical tones in Datong. Inconsistent with previous predictions, we found that the register of T5 has raised from 32 to 53, which has led to a similarity in pitch contour between T5 and T3, making them closely positioned in the tonal space. Growth Curve Analyses and Linear Mixed-Effects Models were conducted in R to investigate F0 and duration differences between this checked-unchecked tone pair across different age groups. As shown in Figure 2, the two tones differed in F0 height, slope, curvature and duration ($ps < .01$) for both older and younger groups. However, for the middle-aged group, they were only significantly different in F0 height and curvature ($ps < .01$), but not in F0 slope and duration ($ps > .05$), indicating a potential merger.

Based on the production results, a 2AFC identification task (140 trials) was conducted to investigate the T3-T5 contrast in perception (or in other words, the primary cues in differentiating the two tones). Two isolated syllables [ta] (T3: “打”, T5: “答”) were produced three times by a 67-year-old male speaker. Four T3-T5 continua (7 steps \times 2 cues \times 2 syllables) were created by manipulating the F0 and duration using PSOLA in Praat. 32 native speakers (20F) were recruited locally to participate in this experiment, including younger group (N = 11, 7F, Mean = 27.27), middle-aged group (N = 9, 6F, Mean = 42.67), and older group (N = 12, 7F, Mean = 68.75). As shown in Figure 3, regardless of whether the F0 continuum was based on T3 or T5, the two identification curves ran in parallel, indicating that F0 changes did not affect participants' identification. However, in the duration continuum based on T5, participants tended to identify the stimuli as T5 when the duration fell below 227ms (step 5), yet as the duration surpassed this threshold, they tended to perceive them as T3. These results suggest that duration is the primary cue to differentiate between T3 and T5, with F0 playing a secondary role. Additionally, the differences between the T3 and T5 continua also indicate the role of glottal stop in enhancing the perception of T5. This underscores glottalization as a necessary cue for T5 identification, given its retention in the T5-based continua. A more specific comparison of the identification results across the three age groups is shown in Figure 4, which is consistent with the production results. In the T5 duration continuum, the identification boundary for T5 is as follows: older < younger < middle-aged, indicating that middle-aged participants also show an extension of duration in perception.

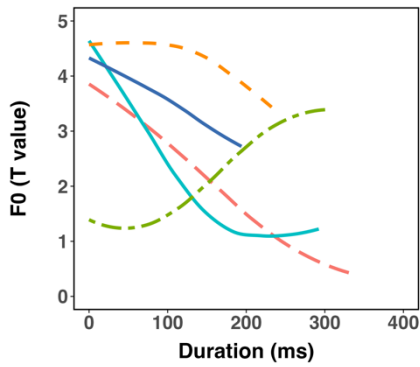


Figure 1. F0 trajectories in Datong Jin Chinese
(T value across all speakers)

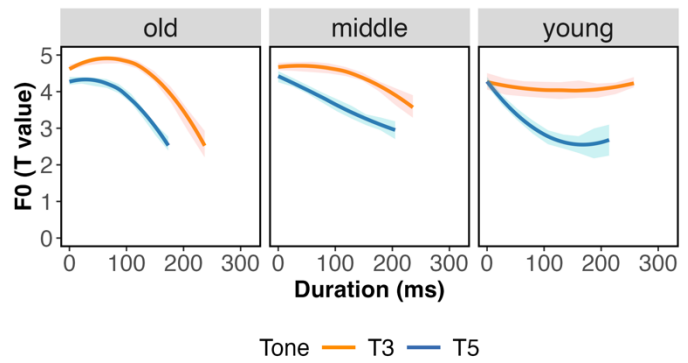


Figure 2. F0 trajectories of T3-T5
(across three age groups)

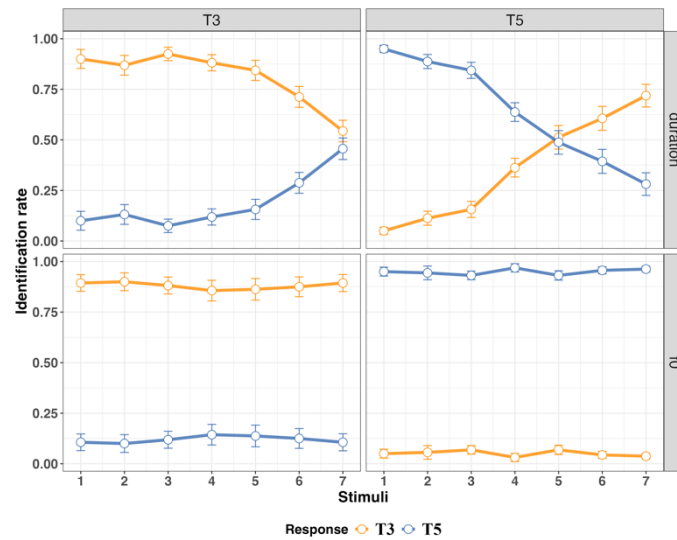


Figure 3. Identification curves for the four continua across all participants. Error bars indicate ± 1 standard error.

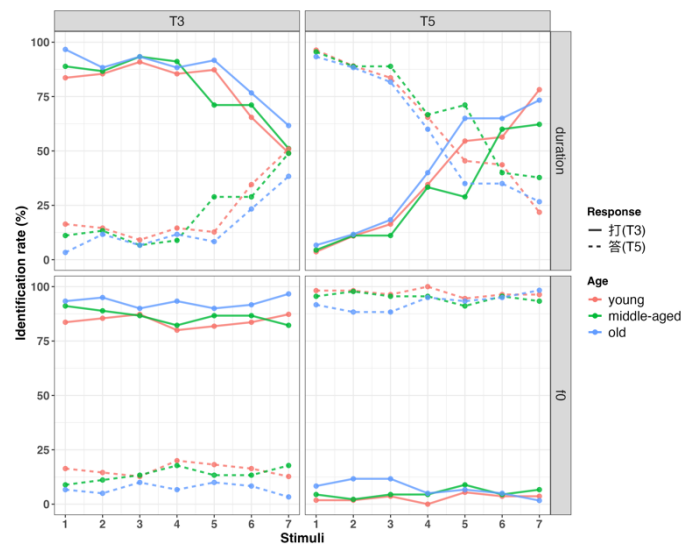


Figure 4. Identification curves for the four continua across three age groups participants.

Intonation-modulated tonal features in Standard Mandarin

Katrina Kechun Li & Yiya Chen

Leiden University

`k.li@hum.leidenuniv.nl`, `yiya.chen@hum.leidenuniv.nl`

Intonation plays a crucial role in shaping tonal realisation in tone languages, with its influence varying in both the temporal scope of its interaction with tones and the extent of its impact on tonal f_0 patterns (see, e.g., Chen 2022 and references therein). In Standard Mandarin, question intonation is generally characterised by a globally raised f_0 trend with an accelerating utterance-final f_0 rise (e.g. Yuan et al. 2006), without affecting tonal identity (e.g. Liu et al. 2016). However, the critical acoustic features of tonal changes remain unclear, particularly given the variability of question intonation across speakers and speaker groups (as shown for English by Xie et al. 2021). This study focuses on Standard Mandarin and aims to delineate its tonal space engaged in naturalistic communication, by identifying stable tonal targets, and pinpointing the key f_0 changes associated with intonation modulation. Furthermore, we compare tonal realisation in large conversational spontaneous speech corpus with that from lab speech produced under carefully controlled intonation conditions.

The spontaneous speech was sourced from Magic-RAMC speech dataset (Yang et al. 2022). We focused on T1, T2 and T4 at sentence-final syllables, which yielded over 30,000 tokens from 663 speakers. The lab speech featured a carrier sentence ‘what he just said is X(?)’, paired with a prompt phrase of equal length that biased the speakers to produce either a statement (‘He told me’ or ‘I’m sure’) or a question (‘What did you say’ or ‘Are you sure’). The monosyllabic target word (X) covered T1, T2 and T4, with controlled and balanced onsets. The lab speech dataset contained 2,490 tokens from 8 speakers. Both datasets underwent consistent processing procedures. From the forced-aligned rhyme portion of each syllable, 20 equidistant f_0 points were extracted using a triangular smoothing window and subsequently normalised with z-scores for each speaker, with strict f_0 filtering protocols applied.

The results showed that the average f_0 patterns in the spontaneous speech corpus resembled statement contours in the lab speech (Figure 1). Figure 2 displays the tonal density contours and kernel density estimate (KDE) across all intonation conditions. Brighter colours (top) and higher KDE peak (bottom) indicate higher density, and thus reflect more stable tonal targets. Although the spontaneous speech involves richer (unlabelled) intonation conditions, both datasets exhibit similar tonal space. T2 clearly demonstrates an early, stable pitch target, while T4 appears to have a late target that is less concentrated. T1, on the other hand, lacks a clearly defined stable tonal target, suggesting that its level contour spans a broader tonal space. The results from Generalised Additive Mixed Model (GAMM) on the lab speech further revealed significant modulations for all three tones under question intonation, though the mechanisms vary (Figure 3). T1 and T4 show register raising from early on. In contrast, T2, despite known as the counterpart of T4 (as rising vs. falling), exhibits primarily late raising effects, consistent with its stable early tonal target observed in the spontaneous speech (Figure 2).

By combining a bottom-up approach (tonal density contours) with a top-down analysis (GAMM), this study identified key f_0 features of tonal changes under question intonation, particularly the asymmetrical effects on seemingly symmetrical tonal patterns (rising T2 vs. falling T4). The findings provided further evidence that intonational effects often illuminate underlying tonal targets (Chen & Gussenhoven 2008; Li 2024). In addition, this study offered direct evidence supporting the compatibility of analysing tone-intonation interactions between spontaneous speech and lab-controlled speech. Future research will explore the perceptual implications of these tone-intonation interactions and the influence of contextual factors on such dynamics.

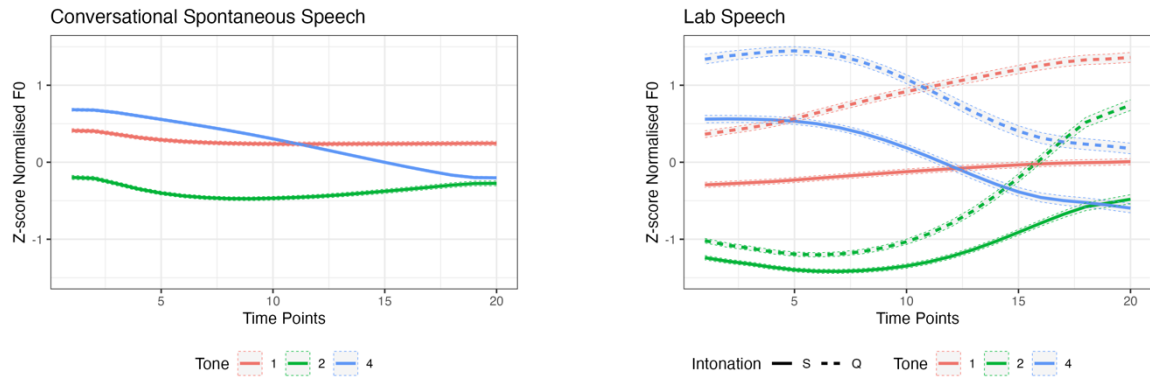


Figure 1. Average tonal contours with standard errors for T1, T2 and T4 in the two datasets.

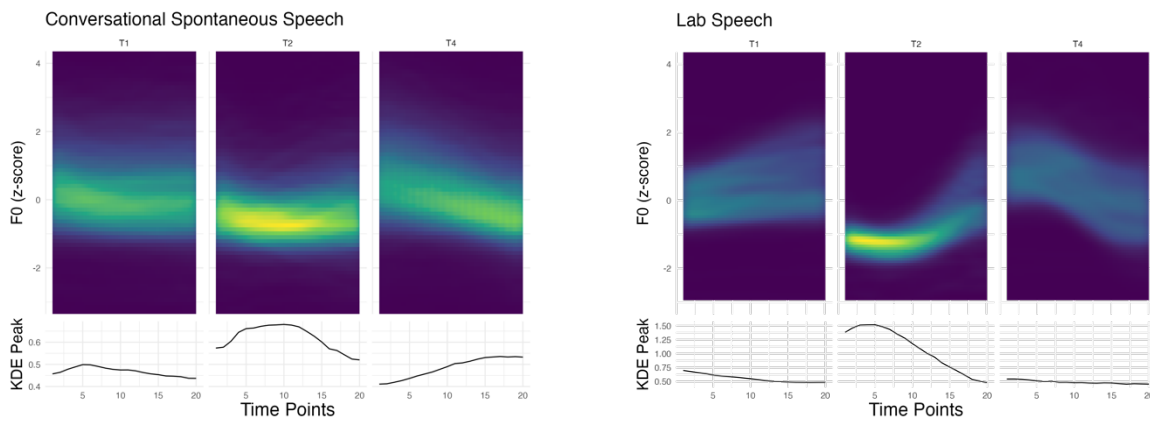


Figure 2. Tonal density contours (top) with KDE peak heights (bottom) for T1, T2 and T4 in the two datasets.

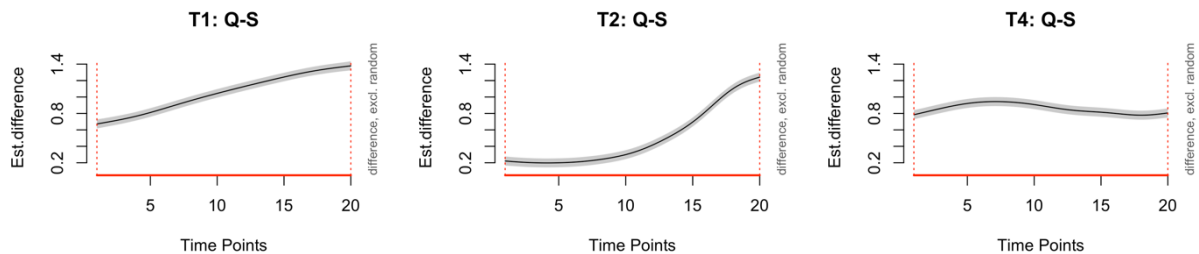


Figure 3. Difference curves between question and statement intonations for T1, T2 and T4 in lab speech.

References

- Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, 36(4), 724–746.
- Chen, Y. (2022). Tone and Intonation. In C.-R. Huang, I.-H. Chen, Y.-H. Lin, & Y.-Y. Hsu (Eds.), *The Cambridge Handbook of Chinese Linguistics* (pp. 336–360). Cambridge University Press.
- Li, K. (2024). *Focal prominence in tone languages: A case study of three Chinese languages* [PhD Thesis, University of Cambridge].
- Liu, M., Chen, Y., & Schiller, N. O. (2016). Online processing of tone and intonation in Mandarin: Evidence from ERPs. *Neuropsychologia*, 91, 307–317.
- Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, 211, 104619.
- Yang, Zehui, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, et al. ‘Open Source MagicData-RAMC: A Rich Annotated Mandarin Conversational (RAMC) Speech Dataset’. *arXiv Preprint arXiv:2203.16844*, 2022.
- Yuan, J. (2006). Mechanisms of Question Intonation in Mandarin. In Q. Huo, B. Ma, E.-S. Chng, & H. Li (Eds.), *Chinese Spoken Language Processing* (pp. 19–30). Springer.

Individual variability in the boundary-induced modulation of stops and tones in Thai

Alif Silpachai^a & Jeremy Steffman^b

^a*Radboud University*, ^b*University of Edinburgh*
alif.silpachai@ru.nl, jeremy.steffman@ed.ac.uk

The production of stops and tones can be modulated by a prosodic boundary. For example, /b/ is produced with more voicing, and the maximum f_0 of tones is lower at the beginning of a sentence than the beginning of a word, that is, in a domain-initial position, in Thai [1]. However, it is unclear whether such domain-initial strengthening (DIS) effects [2] differ between speakers. It is possible that DIS effects are achieved differently between speakers. Some people may strengthen /b/ and /p^h/, while others strengthen only one. Furthermore, DIS in one category may predict modulation in another. For example, a speaker who lowers the f_0 of one tone may also lower the f_0 of another tone to the same extent (cf. individual variability in the production of the mean VOTs of voiceless stops in American English [3]). To address this unclear picture, this study examined how individual variation modulates DIS effects on stops and tones in Thai.

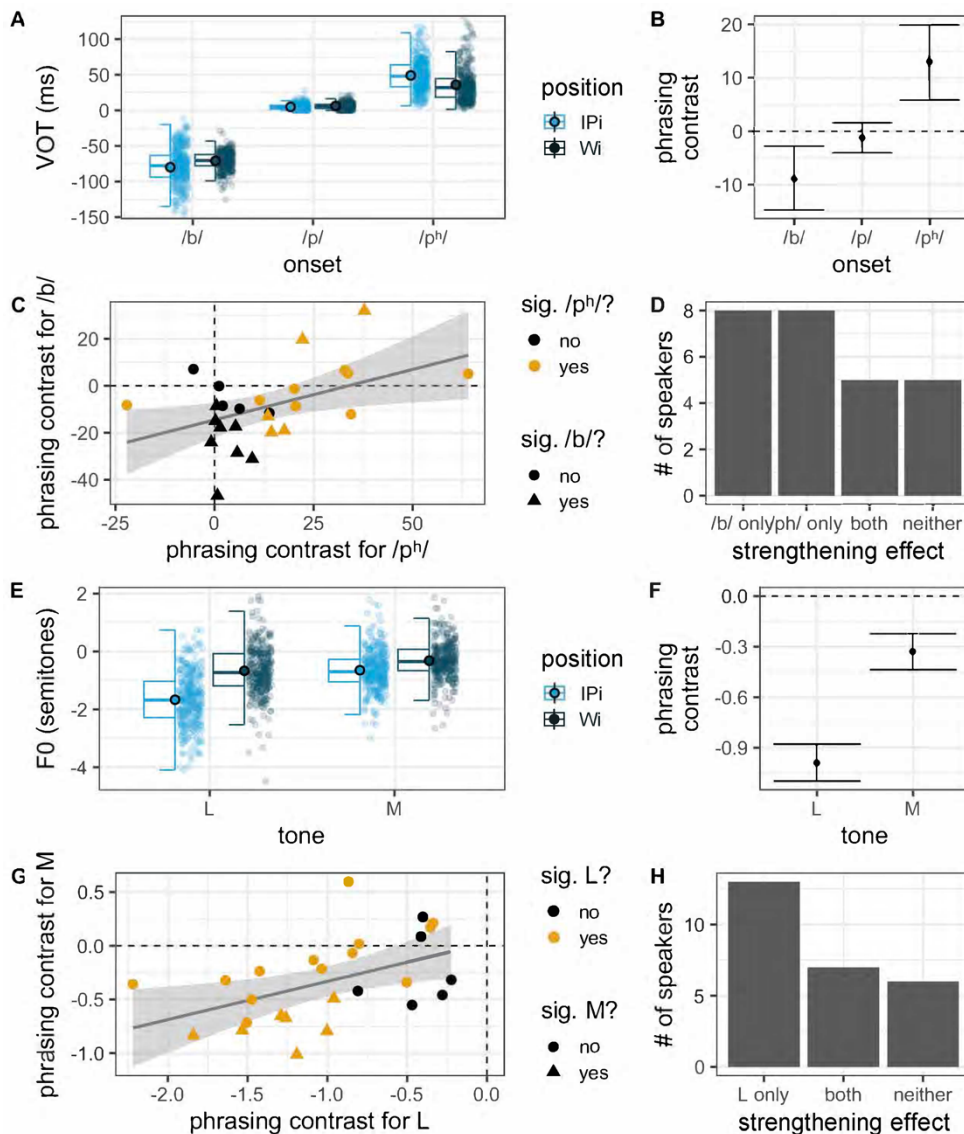
Twenty-six native speakers of Thai (15 females and 11 males, $M = 40.15$, $SD 13.70$, years old) produced monosyllabic words with /b/, /p/, or /p^h/ as the onset consonant having the mid (32) or low (21) tone [4] in two carrier sentences used to elicit the productions of the consonants and tones in IP- and word-initial (IPi and Wi) positions. The stimuli and carrier sentences were presented to the speakers in the Thai script. The sentence for IPi can be translated as “Bun Mi, (my/your/our) uncle/aunt’s _____ is over here”, whereas the sentence for Wi as “Bun has had (this) _____ for a while now.” Both sentences had six syllables. The VOTs of the consonants and the maximum f_0 of the tones, transformed to semitones relative to each speaker’s mean f_0 , were used as correlates of the stop and tone contrasts, respectively.

DIS effects were observed at the group level, as assessed with Bayesian mixed-effects models with speaker as a random effect. VOTs of /b/ and /p^h/ were more negative and more positive, respectively, while /p/’s VOT was not modified by phrasing (see Figures 1A and B). The maximum f_0 of both tones was lower in the IPi position, although the difference between the prosodic positions was more prominent for the low tone, which could be understood as contrast enhancement domain-initially (larger IPi differences in f_0 ; Figures 1E and F). For individual-level results, we fit models for each speaker and then estimated speaker-specific contrasts for the phrasing effect within each phonological category using emmeans [5]. These analyses showed that DIS was realized differently at the individual level. Estimating the difference between IPi and Wi positions showed that eight speakers modulated both /b/ and /p^h/, whereas eight modulated /b/ only, eight modulated /p^h/ only, and five did neither (Figure 1D). For tone, 13 speakers lowered the low tone, seven lowered both tones, and six did neither (Figure 1H); interestingly, no speaker lowered the mid tone only. A positive correlation between the positional effect for /b/ and /p^h/ showed that speakers who modified /p^h/ more tended to modify /b/ less (Figure 1C). For tone, the pattern was the opposite: Larger modulation (mostly lowering) for one tone correlated with larger modulation for the other (Figure 1G; n.b. the difference with Figure 1C is that tones underwent the same lowering, where VOT decreased for /b/ and increased for /p^h/).

Findings suggest that DIS, measured with VOT and maximum f_0 , is realized differently between speakers. It is possible that different factors underlie this individual variation (e.g., age and gender).

Keywords: Domain-initial strengthening, Prosodic enhancement, Lexical tone, Laryngeal contrast, Talker variability

Figure 1: Panel A: VOT for both positions and all three stops. Panel B: marginal group-level contrasts showing a credible (non-zero) strengthening effect for /b/ (more negative when IP-initial), and /p^h/ (more positive when IP initial). /p/ does not show a credible difference. Panel C: speaker-specific contrasts for VOT, showing the estimated size of the phrasing effect. Coloration indicates if this effect was significant for /p^h/ on the x axis, and shape indicates if this was significant for /b/, on the y axis. Panel D: counts of speakers with specific effects. Panel E: Max f₀ for the two tones in both positions. Panel F: the phrasing contrast showing credible lowering for both tones. Panel G: speaker-specific contrasts, mirroring Panel C. Panel H: counts of speakers with specific effects.



References [1] Silpachai, A. (2024). The boundary-induced modulation of obstruents and tones in Thai. *Journal of Phonetics*, 102, 101291. doi:10.1016/j.wocn.2023.101291 [2] Cho, T. (2016). Prosodic boundary strengthening in the phonetics–prosody interface. *Language and Linguistics Compass*, 10(3), 120–141. doi:10.1111/lnc3.12178 [3] Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61, 30–47. doi:10.1016/j.wocn.2017.01.001 [4] Gandour, J. (1979). Tonal rules for English loanwords in Thai. In N. D. Liêm (Ed.), *Southeast Asian Linguistic Studies* (Vol. 4, pp. 131–144). Department of Linguistics, Research School of Pacific Studies, Australian National University. [=Pacific Linguistics (Series C), No. 49]. [5] Lenth, R. (2023). emmeans: Estimated marginal means, aka least-squares means.

Production Evidence for the Perceptual Asymmetry of Mandarin Question and Statement Intonations

Tong Li & Yao Yao

The Hong Kong Polytechnic University, HKSAR, China
tongtanya.li@connect.polyu.hk, y.yao@polyu.edu.hk

The invariant mapping between prosodic cues and category of intended meaning (i.e., the form-function mapping) frequently occurs in the perception of prosody [1-4]. In Mandarin, previous studies have identified challenges in the perception of question versus statement intonations: identifying questions is more difficult when the final tone is rising (T2) than when it is falling (T4) [5, 6]. Although this perceptual asymmetry has been observed cross-linguistically (e.g., in Cantonese [7]), the underlying mechanism remains underexplored. Recent probabilistic models of speech perception grounded in Bayesian principles have indicated the critical role of cue distribution in the perception of prosodic variability in non-tonal languages such as English [3, 4] and German [2]. The probabilistic inference model [2] suggests that listeners learn and store statistical distributions specific to the speaker to overcome variability in prosody.

In this study, we aim to investigate the underlying mechanism of perceptual asymmetry of Mandarin question intonations by examining **the distributional structures of prosodic cues**. Twenty-four monosyllabic minimal word pairs with identical segmental compositions but differing in tone (T2 or T4) were embedded in the final position of carrier sentence “ta1 shuo1 X” (English: “she said X”). Statements and yes/no questions were produced by 6 female native speakers of Mandarin, resulting in a total of 576 critical sentences (24 word pairs \times 2 tone types \times 2 sentence types \times 6 participants). Acoustic measurements, including mean F0, F0 range, and vowel duration, were extracted for each syllable token. Distribution of both raw and speaker-normalized prosodic cues of critical syllables were analyzed and compared across sentence types and tone types.

Results show that consistent with previous studies [5, 6], question intonation consistently exhibits a higher F0 curve and shorter syllable durations compared to statement intonation, except for the final syllable, which is longer in questions. Additionally, the F0 range of the final syllable in questions varies with tone type, expanding significantly with a rising tone, while remaining unchanged with a falling tone. Focusing on the final syllables in greater detail, most speakers clearly distinguished questions and statements ending in T2 by mean F0 and F0 range, but questions and statements ending in T4 was more distinguished by duration and mean F0. The location and spread of the distributions of question and statement prosodic cues differ across speakers, leading to the substantial overlap in the overall distribution of raw prosodic features between questions and statements (Figure 1), especially in mean F0 and F0 range for T2 and in duration and mean F0 for T4. Normalizing prosodic cues by speaker increases the distinction in mean F0 (Figure 3), but substantial overlap remains in the F0 range of T2 and duration of T4, providing evidence for the inherent distributional variability of F0 range and duration characteristic to speakers that normalization cannot fully resolve.

In conclusion, the current study provides preliminary evidence for the role of distributional structures in the perceptual asymmetry between T2 and T4 question intonations in Mandarin. It highlights **differences in the systemic structure of cross-speaker variability between T2 and T4 conditions**: for T2-final intonations, F0 range and mean F0 demonstrate critical speaker-specific variability, whereas T4-final intonations reveal prominent variability of mean F0 and duration across speakers. However, whether this difference contributes to the greater difficulty in identifying question with final T2 compared to final T4 needs further verification by perceptual experiment. Our ongoing work is to examine whether listeners are equally sensitive to cue-distribution variability of mean F0, F0 range, and duration across speakers and whether their perceptual adaptation ability to categorize question intonation is related to the speaker-specific distribution of prosodic cues.

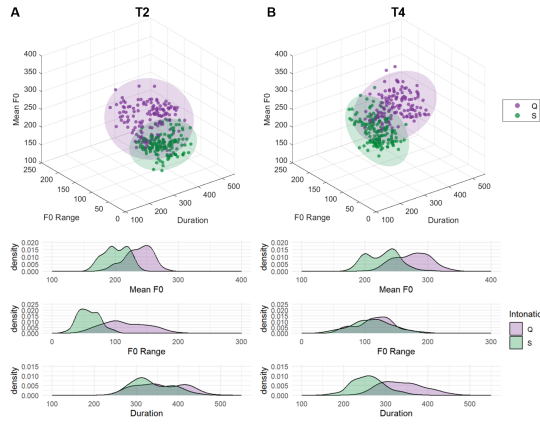


Figure 1: *Non-normalized cue distribution in sentence-final syllable across all speakers.*

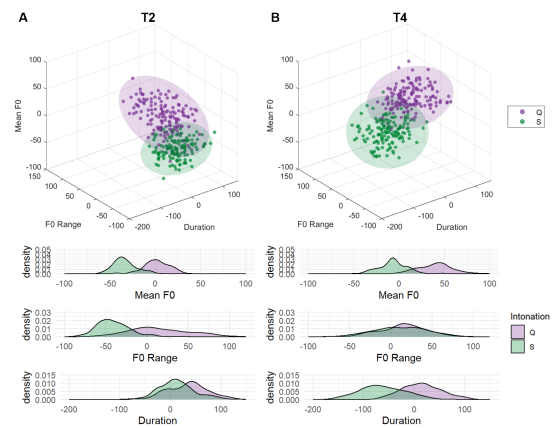


Figure 3: *Speaker-normalized cue distribution in sentence-final syllable across all speakers.*

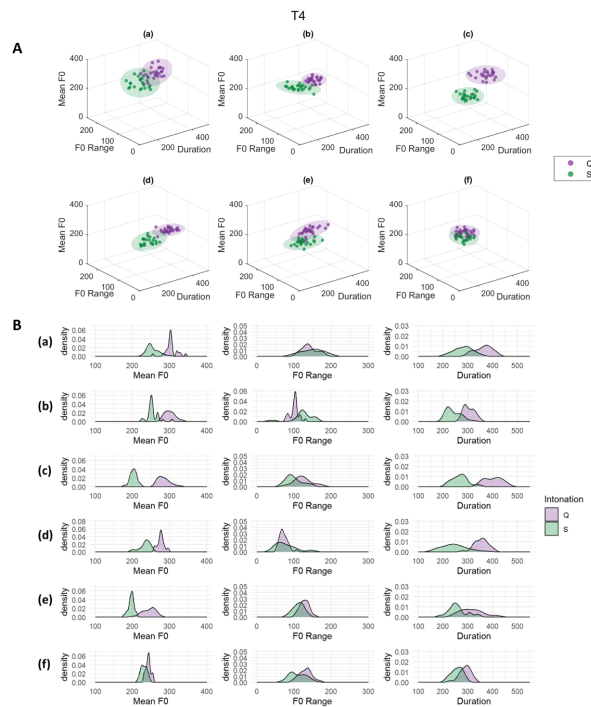
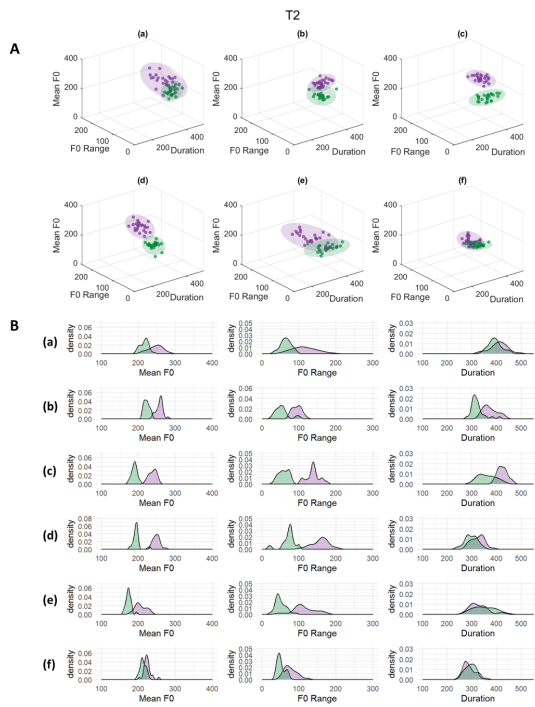


Figure 2: *Speaker-specific cue distribution in sentence-final syllable when sentences end with T2 (left panels) and T4 (right panels).*

References

- [1] Kurumada, C., & Roettger, T. B. (2022). Thinking probabilistically in the study of intonational speech prosody. *WIREs Cognitive Science*, 13(1), e1579.
- [2] Kurumada, C., Brown, M., & Tanenhaus, M. K. (2018). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin & Review*, 25(3), 1153–1160.
- [3] Kurumada, C., Rivera, R., Allen, P., & Bennetto, L. (2024). Perception and adaptation of receptive prosody in autistic adolescents. *Scientific Reports*, 14(1), 16409.
- [4] Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, 211, 104619.
- [5] Yuan, J. (2006). Mechanisms of Question Intonation in Mandarin. In Q. Huo, B. Ma, E.-S. Chng, & H. Li (Eds.), *Chinese Spoken Language Processing* (Vol. 4274, pp. 19–30). Springer Berlin Heidelberg.
- [6] Liu, M., Chen, Y., & Schiller, N. O. (2016). Online processing of tone and intonation in Mandarin: Evidence from ERPs. *Neuropsychologia*, 91, 307–317.
- [7] Xu, B. R., & Mok, P. (2012). Cross-linguistic perception of intonation by Mandarin and Cantonese listeners. *Proceedings of Speech Prosody 2012*, 99–102.

Variation and change in Bangkok Thai tonal contours: A comparison of data from the 1990's and 2020's

Cathryn Yang¹, Pittayawat Pittayaporn², James Kirby³

¹*SIL International and Payap University, Thailand*

²*Department of Linguistics & Center of Excellence in Southeast Asian Linguistics, Chulalongkorn University, Bangkok, Thailand*

³*Institute of Phonetics and Speech Processing, Ludwig Maximilian University of Munich, Germany*

Keywords: Bangkok Thai, tonal contours, re-study, growth curve analysis

This study examines variation and change in Bangkok Thai tone contours by comparing speech samples of comparable materials recorded 30 years apart. Our aim is to explore potential generational shifts in the tonal contours of the High tone, which has been reported as undergoing change-in-progress [1]. We record stimuli similar to those used in two tonal acoustic studies published in the 1990's: [2], which analyzed tone citation forms from monosyllabic words, comparing two generations (5 male and 5 female speakers born in the 1930s vs. 10 male speakers born in the 1960s); and [3], which examined tonal coarticulation in connected speech using 25 disyllabic sequences embedded in a carrier sentence, recording the same 10 young male speakers who participated in [2].

Method. This re-study is part of the Chulalongkorn Corpus of Spoken Thai (C-COST) project. The stimuli consist of a wordlist of 206 monosyllabic words and the 25-sentence set from [3]. The wordlist included the stimuli used in [2], plus additional words with a variety of onset-tone combinations. Each participant read the wordlist twice and the sentence set three times. We analyzed data of 23 speakers (13F, 10M) from the Bangkok Metropolitan Area (age range 18-29 yrs, n=14; age range 64-81 yrs, n=9).

Recordings were segmented and labeled using MAUS [4] with manual corrections, and f0 values of syllables ending in a vowel or sonorant consonant were extracted every 5 msec using Praat's pitch estimator [5] and processed in EMU [6]. To normalize f0 across speakers, values were converted to semitones relative to each speaker's mean f0 of the Thai Mid tone, which is typically realized as a mid-level tone with a slight final fall, to reflect a tone's relative distance to Mid within a speaker's tonal space. After filtering out short tokens (<50 ms) and tokens with less than 10 f0 values, 2,270 tokens were retained for analysis.

The f0 trajectory was analyzed using growth curve analysis (GCA) in R, with a linear mixed-effects model incorporating a third-order polynomial to capture the variation in pitch contour, as in [7]. Predictor variables included generation (old/young), sex (female/male), style (citation/sentence), duration (centered by style). Interactions between predictors and polynomial terms allowed assessment of how social and stylistic factors influence not only overall pitch height (intercept) but also slope, curvature, and cubic parameters of the contour. Preceding and following tone were included as additive fixed effects (with Mid tone as the reference level). Random intercepts for speaker and syllable accounted for individual and lexical variability in pitch production.

Results. To illustrate the range of individual variation, Figure 1 presents the tone systems (citation form) of two male speakers—one younger and one older—from each time period. The High tone exhibits a range of shape variants, including domey rise-fall, scoopy rise-fall, straight rise and scoopy rise. Domey rise-fall, exemplified in the left panel of Figure 1, may be considered a highly conservative variant, aligning with descriptions of the High tone from the early 20th century [3]. In the 1990's study [2], the domey variant appeared in the average contour plots of older males, but not in the plots of older females or younger males, suggesting curvature as an important parameter of variation.

The best GCA model for the 2020's data indicate that generation, sex, style, and duration significantly affect the shape parameters of High. Figure 2 shows the predicted trajectories across these factors, illustrating their interaction. Regarding curvature (quadratic term), younger speakers show more positive (scoopy) curvature ($\text{poly2} \times \text{Young } \beta = +106.6^{***}$), and the generational difference is larger for males than for females ($\text{poly2} \times \text{Young} \times \text{Male } \beta = +45.42^{***}$). However, the generational difference diminishes in connected speech ($\text{poly2} \times \text{Young} \times \text{sentence } \beta = -99.33^{***}$). These results suggest that though domey rise-fall is preserved by older males under certain conditions, younger speakers favor a contour with an initial scoop, though it remains uncertain whether this age-related pattern will lead to scoopy rise becoming the dominant variant over time.

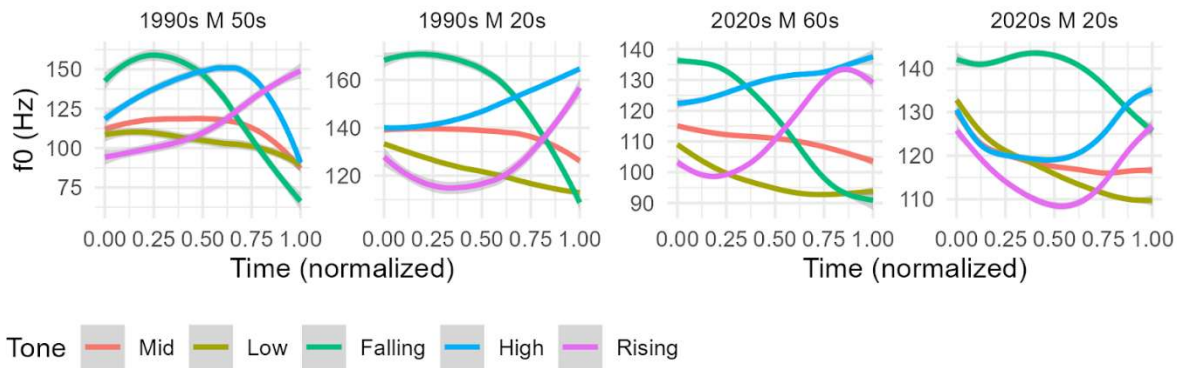


Figure 1: Selected tone systems of older and younger male speakers from the 1990's and 2020's. Left two panels: 1990's f0 trajectories redrawn from [1], representing the average contour over 5 tokens per tone per speaker. Right two panels: 2020's empirical f0 trajectories of citation form using geom_smooth, method=gam.

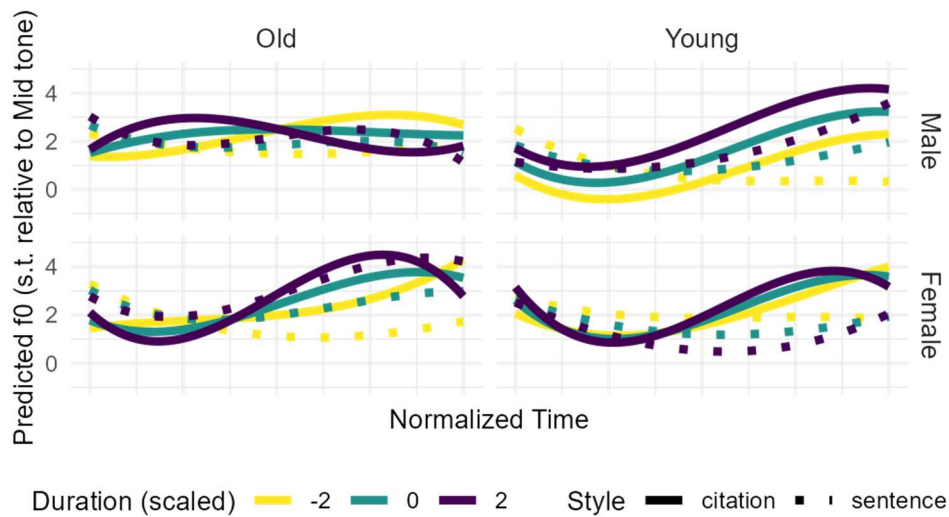


Figure 2: GCA model predicted f0 values for Thai High tone by generation and sex, colored by duration (scaled, SD) and line type by Style.

References:

- [1] E. C. Zsiga, "Modeling diachronic change in the Thai tonal space," *University of Pennsylvania Working Papers in Linguistics*, vol. 14, no. 1, pp. 395–408, 2008.
- [2] J. Gandour, S. Potisuk, S. Ponglorpisit, and S. Dechongkit, "Inter- and intraspeaker variability in fundamental frequency of Thai tones," *Speech Communication*, vol. 10, no. 4, pp. 355–372, Nov. 1991, doi: 10.1016/0167-6393(91)90003-C.
- [3] J. Gandour, S. Potisuk, and S. Dechongkit, "Tonal coarticulation in Thai," *Journal of Phonetics*, vol. 22, no. 4, pp. 477–492, Oct. 1994, doi: 10.1016/S0095-4470(19)30296-7.
- [4] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, Sep. 2017, doi: 10.1016/j.csl.2017.01.005.
- [5] P. Boersma and D. Weenick, *Praat: doing phonetics by computer*. (2024). Accessed: Nov. 18, 2024. [Online]. Available: <https://www.praat.org>
- [6] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392–410, Sep. 2017, doi: 10.1016/j.csl.2017.01.002.
- [7] J. Kirby, R. Puggaard-Rode, S. Maspong, and F. Burroni, "Effects of coda consonants on preceding vowel F0," in *Speech Prosody 2024*, ISCA, Jul. 2024, pp. 324–328. doi: 10.21437/SpeechProsody.2024-66.

Going beyond average contours in the study of tone variation and change:

A case study from Thai Tone 4

Teerawee Sukanchanon¹, Francesco Burroni^{1,2}

¹Center of Excellence in Southeast Asian Linguistics, Faculty of Arts, Chulalongkorn University, Thailand

²Institute for Phonetics and Speech Processing, Spoken Language Processing Group, LMU Munich, Germany

Compared to the study of variation and change in the production of consonants and vowels, our understanding of variation and change in tonal systems remains limited [1, 2]. A tonal system for which drastic changes have been reported in the last one hundred years is that of Central Thai [1, 3-5]. The most dramatic of these changes has been reported for Tone 4 (T4) which has purportedly evolved in a few steps from a High-Falling tone realization (42) in the early 1900s, to a mid rising tone (34) in the 1990s (**Fig. 1**). Evidence in favor of tone changes, like the ones assumed for T4, has often come from the inspection of single tokens or averages across generations [1-4]. However, average-based approaches are problematic because they overlook structured variation *within* tone categories. Without a firmer understanding of said variation and its structure across generations, definitive conclusions about variation vs. diachronic change cannot be drawn. To tackle this issue, we propose an integrated methodological approach that can help us understand within-category variation and map it out across generations combining traditional statistical analysis, functional data analysis, and clustering. We apply our methodology to the realization of Thai T4 to assess the claims of diachronic changes across generations. To preview our findings, we observed the existence of variants for the production of T4 *within each generation* that map to *hypothesized diachronic stages*. In contrast with previous suggestions, the distribution of these variants has remained fairly stable across generations, suggesting that the extent of diachronic change to T4 may have been overestimated.

Methods. To study variation in the production of Thai T4, we analyzed data from two Thai speech corpora: THAIMIT [5] and a subset of the C-COS corpus. **Subjects:** 42 speakers who are native speakers of Standard Thai were categorized into three age groups: G1 ($n=14$; age range=21-36; $\mu=28.4$; $\sigma=5.1$), G2 ($n=21$; age range=38-54; $\mu=46.1$; $\sigma=5$), and G3 ($n=7$; age range=58-85; $\mu=72.7$; $\sigma=9.2$). **Materials:** The words we selected are open/sonorant-coda syllables (to ensure F0 tracking over the entire syllable) produced in pre-pausal position (to avoid tonal coarticulatory effect). A total of 1,242 tokens were analyzed. F0 of each token was extracted with a neural pitch tracker in MATLAB in 10 ms time-step. F0 was then z-scored by participant. **Analyses:** 3 analyses were conducted: (1) An analysis of categorical variants informed by previous literature, (2) a functional principal component analysis (FPCA) to identify modes of variation in the data, and (3) an unsupervised dynamic time-warping (DTW) based hierarchical clustering (HC) analysis. We adopt DTW-based HC because, unlike other methods (e.g. k-means), DTW does not require time-normalization, thus, it allows us to detect duration-based clusters. For (1), tokens were categorized as *withDip* or *noDip* based on the presence of a slight fall at trajectory offset, a key generational difference noted in previous studies [1,4]. A mixed-effects logistic regression assessed the effect of speaker generation on variant distribution. For (2), linear mixed-effect models were applied to the first three PC scores (capturing 97% of the total variance) to examine generational shape variation. For (3), pairwise DTW-based distances between non-time-normalized trajectories generated a distance matrix, which was then subjected to HC. The number of clusters was determined *via* dendrogram inspection and cophenetic correlation.

Results. For the categorical analysis, the model showed that generation is not a significant predictor of the *withDip* vs *noDip* variants. Unlike what was found in [4], where the proportion of *withDip* and *noDip* variants was 80:20 and 30:70 respectively for older and younger groups, our data showed roughly the same proportion of 30:70 for all generations, with a gradual increase in number of *noDip* variants towards younger speakers (G1) (**Fig. 2**). As for the FPCA, linear models showed that, out of 3 PC scores, only PC2 is significantly different between younger speakers (G1) vs. the two older groups (G2 and G3). The higher values in PC2 for G1 indicate lower height of the first half of T4 trajectory and higher offset values (**Fig. 3**). Finally, DTW-based HC showed different contour shapes that seem to reflect diachronic stages hypothesized in previous work. In line with the categorical analysis, distributions of different variants appear to be similar across generations: the *noDip* variant (Cluster 4) has the highest frequency in all generations, but its frequency is slightly higher in the youngest speaker group (G1) (**Fig. 4**). This finding also dovetails with the FPCA findings where higher PC2 score of G1 indicates a lower trajectory up the midpoint and higher offsets.

Conclusions. We have proposed a series of methods to investigate within-category tone variation and change at different levels of granularity and based on increasingly less assumptions about the data (features based on previous literature → FPCA → DTW-based hierarchical clustering). We suggest that our approach is systematic and broadly applicable to other case studies of tonal variation and change. Relying on this approach, we found that what has been labeled categorical diachronic changes in average shapes for T4 appears in fact to be *gradual change in the frequency of different tonal shapes* within a tonal category that is then reflected in different averages across generations, as some have hypothesized [1, 3]. An aspect not fully established in previous work. Even so, the distribution of different shapes has remained fairly similar across generations, with only some variants becoming more frequent in younger speakers. Changes to Thai T4 are, thus, not as drastic as suggested by average-based analyses. Equipped with this approach, future research can investigate in more detail variation and change in the tonal system of Thai and of other languages.

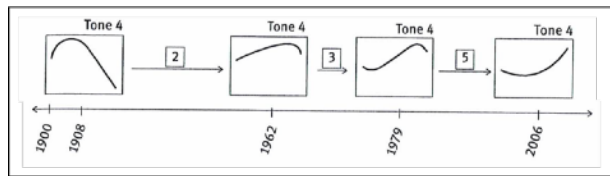


Fig. 1 T4 changes across the past century. (adapted with permission from [1])

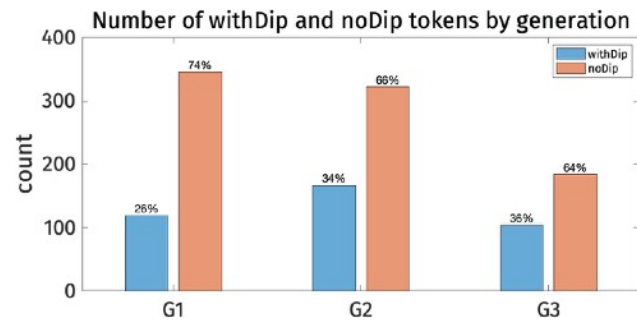


Fig. 2 Distribution of *withDip* and *noDip* variants by generation

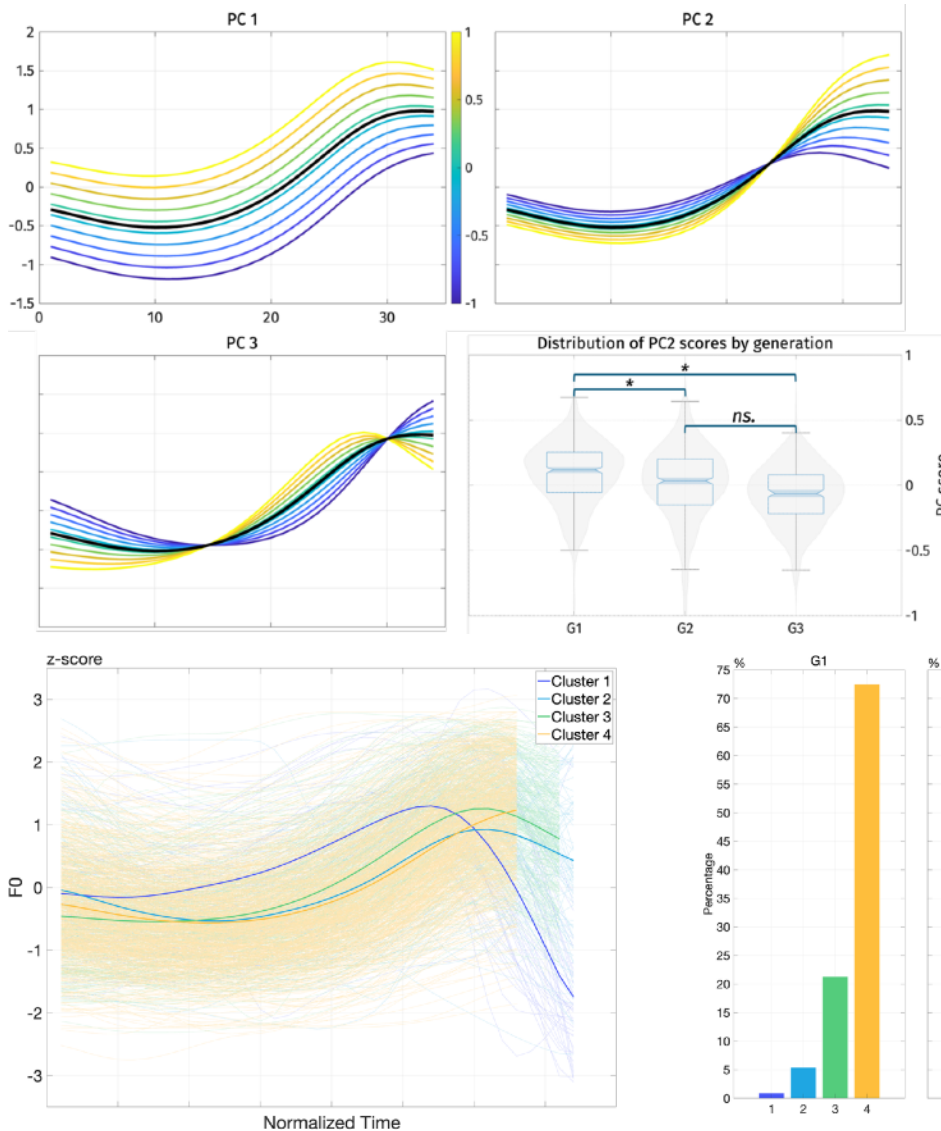


Fig. 3 Illustration of main modes of variation captured by PC1 (upper left), PC2 (upper right), and PC3 (bottom left) of T4 trajectories.

(Bottom right) Box plots of score distribution for PC2 by generation. Note the difference between G1 vs. G2 and G3.

Fig. 4 (Left) Mean and raw T4 trajectories by cluster based on DTW hierarchical clustering. Note the existence of clusters that resemble the diachronic stages posited in Figure 1. (Right) Distribution of T4 trajectories in different clusters for each generation (G1: left panel, G2: middle panel, G3: right panel)

References

- [1] Pittayaporn, P., *Phonetic and systemic biases in tonal contour changes in Bangkok Thai*. 2018, Mouton de Gruyter Berlin.
- [2] Yang, C. and Y. Xu, *Crosslinguistic trends in tone change: A review of tone change studies in East and Southeast Asia*. *Diachronica*, 2019. **36**(3): p. 417-459.
- [3] Pittayaporn, P. *Directionality of tone change*. 2007. Saarbrücken, Germany: Saarland University.
- [4] Teeranon, P. and R. Rungrojsuwan, *Change in the Standard Thai high tone: An acoustic study*. *MANUSYA: Journal of Humanities*, 2009. **12**(3): p. 34-44.
- [5] Thepboriruk, K. *Bangkok Thai Tones Revisited*. *Journal of the Southeast Asian Linguistics Society*, 2010. **3**: p. 86-105.
- [6] Chanchaochai, N., Cieri, C., Debrah, J., Ding, H., Jiang, Y., Liao, S., Liberman, M., Wright, J., Yuan, J., Zhan, J., & Zhan, Y. *GlobalTIMIT: Acoustic-Phonetic Datasets for the World's Languages*. *Interspeech 2018*, 2018. p. 192-196.

The evolution of uptalk in Standard Southern British English

Jiseung Kim, Alanna Tibbs, Amalia Arvaniti

Radboud University

jiseung.kim@ru.nl; alanna.tibbs@ru.nl; amalia.arvaniti@ru.nl

Uptalk is the use of a pitch rise at the end of statements instead of a standard fall. Studies on Standard Southern British English (SSBE) from the early 2000s report highly variable forms [1], no systematic distinction between rises on questions and statements [2], and considerable gender differentiation in shape, function or frequency [1, 3, 8]. Thus, uptalk at that time showed characteristics of an innovative feature [6, 7]. We hypothesize that the form and usage of uptalk would eventually stabilise, diverge from that of other types of rises, and show less gender-based differences [7].

To investigate the current status of uptalk in SSBE, we collected data from 30 SSBE native speakers (19 F; median age 21; 25 White British, 5 mixed), who were recorded in an Oxford recording studio. Participants silently read 36 pairs of context-response sentences, evenly divided into 6 pragmatic conditions, representing typical uptalk functions, *Confirmation Request*, *Negotiation*, *Uncertainty* [cf. 3, 8], as well as *Sarcasm* [cf. 8], *Polar Question*, and *Listing* (see Table 1). In each trial, a confederate read aloud the context and the participant the response without being prompted as to what tune to use. The targets in the responses were utterance-final words with penultimate stress, except that in *Listing* the target was only the first of three list items. We predicted that the responses to the first three categories above would end in a rise, if uptalk is now consistently used.

Of a total of 1080 utterances from 30 speakers, 977 from 29 speakers were retained for analysis. One speaker's 36 utterances were excluded because the speaker produced 1/3 of all utterances as falls; another 67 utterances from the remaining data were excluded because they were also produced as falls (on average 2.63 per speaker, with a maximum of 7 falls from a speaker). Target F0 was extracted using Praat [9] and normalized for speaker by converting it from Hz to semitones. Time normalization was done using landmark registration, a curve alignment process in which temporal landmarks, here the stressed vowel onset and offset, are used as reference points where the intervals separated by the landmarks are warped [10]. The 977 processed F0 curves were submitted to Functional Principal Component Analysis (FPCA), which identifies the dominant modes of variation within the data in the form of Principal Component (PC) curves [10]. Additionally, the time-normalised F0 curves were analysed in two Generalized Additive Mixed-effects Models (GAMMs); one with pragmatic category and the other with gender as fixed intercepts and smooth terms. Speaker was included as factor smooth (random slope and intercept) in the first GAMM. We used the *fade* [11] (for FPCA) *mgcv* [12] and *itsadug* [13] (for GAMMs) packages in R [14].

The FPCA showed that PC1, PC2 and PC3, which together accounted for 94% of curve variance, reflected three main differences among the rises, that respectively have to do with the start of the rise and the extent to the rise excursion (PC1), the scaling of the accented vowel (PC2), and the overall shape, as a scooped rise or a concave rise-plateau (PC3); see Fig. 1a. The GAMMs confirmed these differences and showed that the rises varied by pragmatic category (Fig. 1b), though not by gender. The effect of pragmatic category was supported by a series of Linear Mixed Models (LMMs) with the PC1, PC2, and PC3 coefficients as dependent variables, pragmatic category as a fixed effect, and speaker and item as random intercept. By and large, both GAMMs and LMMs showed that *Polar Question* differed the most from all other functions, showing a rise-fall-rise pattern. In *Listing*, F0 rose early and ended in a plateau. *Confirmation Request*, *Negotiation*, *Uncertainty*, and *Sarcasm* were similar to each other, except that *Sarcasm* was scaled lower. In short, typical uptalk functions (*Confirmation Request*, *Negotiation*, *Uncertainty*) had a more comparable shape, with a rise starting on the accented vowel.

In conclusion, young British speakers consistently used final rises in several pragmatic contexts associated with uptalk. Critically, the forms of the final rises produced in the contexts of typical uptalk functions were stable relative to those reported in earlier studies [1], systematically diverging from the rises of *Polar Question* [2] and *Listing* [14], suggesting that uptalk has stabilised and diverged from other rises since earlier reports [1, 2], similarly to other varieties in which uptalk is considered established [1, 5]. In addition, we did not observe gender-based differentiation reported in previous studies [1, 3, 8]. Further, the rises were used in a scripted speech production task, suggesting that uptalk is not exclusive to casual speech, at least for this speaker demographic. In short, uptalk is becoming a stable intonation feature among young British speakers.

Table 1. Sample stimuli; the target words are bold and underlined.

Pragmatic Categories	Context	Response
Confirmation Request	What name is your appointment under?	<u>Alanna</u> ?
Uncertainty	What time did you get back last night?	<u>Eleven</u> ?
Negotiation	What should we make for dinner tomorrow?	<u>Lasagna</u> ?
Polar Question	I visited England recently.	Was it <u>rainy</u> ?
Listing	Who did you go to the pub with?	I went with <u>Melinda</u> , Malena, and Yolanda.
Sarcasm	Why is David Beckham supporting <u>Qatar</u> ?	Because <u>money</u> ?

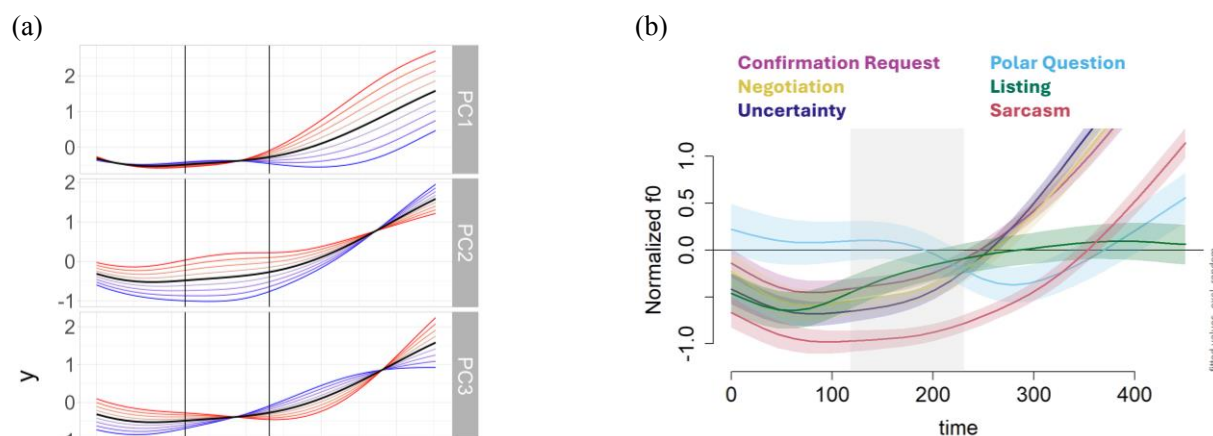


Figure 1. In (a), the first three PCs; the lines are PC curves with higher (red) and lower (blue) than the mean (black) PC coefficients. Vertical gray lines denote the onset and offset of the accented vowel. In (b), smooth curves by pragmatic category; the shaded interval denotes the accented vowel.

References

- [1] Barry, A. S. 2008. *The form, function and distribution of high rising intonation: A Comparative Study of HRT in Southern Californian and Southern British English*. VDM Verlag.
- [2] Shobbrook, K. & J. House. 2003. High rising tones in Southern British English. *15th ICPHS*, 1273–6. <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/>
- [3] Levon, E. 2016. Gender, interaction and intonational variation: The discourse functions of High Rising Terminals in London. *Journal of Sociolinguistics* 20: 133–163.
- [4] Levon, E. 2020. Same difference: the phonetic shape of High Rising Terminals in London. *English Language and Linguistics* 24(1): 49–73. doi:10.1017/S1360674318000205
- [5] Ritchart, A. & A. Arvaniti. 2014. The form and use of uptalk in Southern California English. *Speech Prosody* 7. https://www.isca-archive.org/speechprosody_2014/ritchart14_speechprosody.pdf
- [6] Trudgill, P. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in society* 3(2): 215–246.
- [7] Labov, W. 2010. *Principles of linguistic change: Cognitive and cultural factors*. Oxford: Wiley.
- [8] Arvaniti, A. & M. Atkins. 2016. Uptalk in Southern British English. *Speech Prosody* 8, 153–7. https://www.isca-archive.org/speechprosody_2016/arvaniti16_speechprosody.html
- [9] Boersma, P. & D. Weenink. 2024. Praat: doing phonetics by computer [Computer program]. Version 6.4.24, retrieved 1 December 2024 from <http://www.praat.org/>
- [10] Ramsay, J. 2005. Functional Data Analysis. In B.S. Everitt & D.C. Howell (eds), *Encyclopedia of statistics in behavioral science*. <https://doi.org/10.1002/0470013192.bsa239>.
- [11] Ramsay, J. O., Graves, S., & Hooker, G. 2020. *fda: Functional Data Analysis*. R package ver 5.1.5.1.
- [12] Wood, S. 2017. *Generalized Additive Models: An introduction with R*. Chapman and Hall/CRC.
- [13] van Rij, J., M. Wieling, R. Baayen & H. van Rijn. 2022. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package version 2.4.1.
- [14] R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, [Online]. Available: <http://www.r-project.org/>.

Diachronic changes in the intonation of Italian newsreading

Michelina Savino¹, Simon Wehrle² and Martine Grice²

¹*Dept. of Education, Psychology, Communication, University of Bari, Italy*

²*IfL-Phonetics, University of Cologne, Germany*

michelina.savino@uniba.it, {simon.wehrle, martine.grice}@uni-koeln.de

Keywords: diachronic change, intonation, broadcast reading style, speech rate, Italian

Background. The prosodic characterisation of the broadcast reading style across different languages has received increasing attention in recent years (e.g. [1, 2, 3, 4, 5, 6], but less attention had been paid to its evolution over time. For Italian, two preliminary studies [7, 8] comparing the prosodic features of newsreading over a forty-year time span (late 1960s vs. 2005), converged in observing a faster speaking rate in the “modern” newsreading style (longer IPUs and fewer pauses). Results on F0 were less clear. In particular, [8] found no difference regarding the intonation style of newsreading across the two eras, as measured in terms of wiggleness (the rate of F0 slope changes per second), and spaciousness (the size of F0 excursions, a measure related to pitch range) [9, 10]. Instead, higher mean F0 values were observed in the modern newsreading style—but this result was restricted to one single 2005 reader. In both of these preliminary studies, the data under analysis were obtained from a publicly available Italian corpus of original newsreadings by 6 newsreaders (all male) from the late 1960s, along with renditions of the same texts read by 2 journalists (one male, one female) in 2005 [11]. This elicitation method might have influenced the naturalness or genuineness of the “modern” newsreading style, as the two 2005 speakers were requested to read news reporting on events that had occurred decades before. To address this limitation, we replicated the analysis of [8] on a dataset consisting entirely of original newscasts, comparing the intonational style of the Italian newsreaders of the 1960s [11] with those of a comparable number of newsreaders recorded in 2024.

Materials and method. We analysed data from all 6 newsreaders (all male) from the original 1960s corpus, and from 13 (7 male, 6 female) newsreaders recorded in 2024 (*RAI* online archive). To maintain similarity of news content, recordings broadcast on the same day or, where this was not possible, in the same week, were selected. We followed the procedure described in [8] for annotating the recordings and for hand-correcting and processing F0 contours. In total, 1557 IPU (minimum pause duration 200 ms) entered into the analysis. Bayesian hierarchical regression modelling (with speaker as a random factor) was used for statistical analysis.

Results. For intonation style, visual inspection shows a trend for higher spaciousness in the 2024 compared to the 1960s newscasts; see Fig. 1. Compared to the 1960s readers, 5 out of the 13 present-day readers had higher mean values of spaciousness, and none had lower mean values. However, mean values by group are similar (2024 mean = 8.02, SD = 2.33; 1960s mean = 7.64, SD = 1.93), and Bayesian modelling suggests that there is no robust group-level effect ($\delta = 0.49$; 90% CI [-0.41, 1.39]; $P(\delta > 0) = 0.82$). Wiggleness did not seem to differ by era. The analysis of mean F0, conducted only for male speakers (as there are only male speakers in the 1960s data), shows a similar picture. The 2024 data accounted for 4 out of the 5 highest by-speaker means, and group means were also higher for 2024 (142 Hz, SD = 33) compared to the 1960s (119 Hz, SD = 29); see Fig. 2. Bayesian modelling does not provide unambiguous evidence for this observation—reflecting in part the reduced sample size after the exclusion of female speakers—but still indicates a moderate effect of group ($\delta = 13$; 90% CI [-5.18, 30.99]; $P(\delta > 0) = 0.89$). Finally, the previously attested higher speech rate in modern newsreading is confirmed via the metric of IPU duration. Mean durations were more than twice as long in the 2024 (mean = 5.93 sec., SD = 3.34) compared to the 1960s readings (mean = 2.84 sec., SD = 1.37), and all 2024 speakers produced longer mean IPU than all 1960s speakers; see Fig. 3. Bayesian modelling provides unambiguous evidence for the effect of group for speech rate ($\delta = 1.99$; 90% CI [1.61, 2.43]; $P(\delta > 0) = 1$).

Discussion. Our analysis confirms preliminary crosslinguistic observations on modern newsreading styles as being characterised by faster speech rate. In the current corpus of Italian, we observe a tendency in the 2024 newsreaders to produce larger F0 excursions (higher spaciousness values), a pattern that may be partially, but not fully, attributable to these speakers’ production of longer IPU [9]. A trend is also detectable for an overall higher pitch level in the modern era compared to the 1960s. These observations may in part be driven by the changed scenario in modern broadcasting, which is characterised by fierce competition among public and private TV broadcasting companies. Faster delivery and high pitch attract more attention than slow delivery and low pitch [12]. Thus, using fast delivery and higher pitch in newsreading may help to attract and also maintain the attention of viewers in an effort to discourage them from switching to another channel.

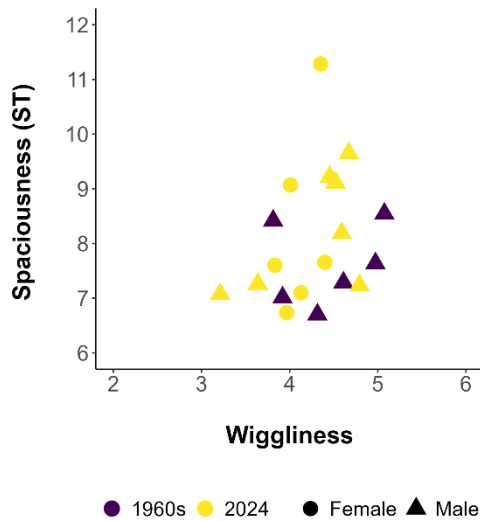


Figure 1: Wiggleness and spaciousness values by speaker, era, and gender.

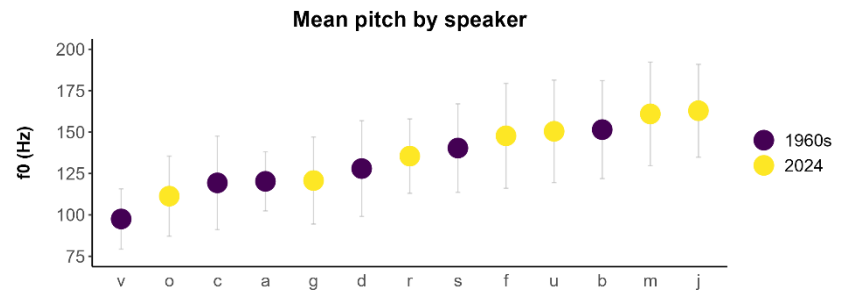


Figure 2: Mean F0 (in Hz) by speaker and era (male speakers only).

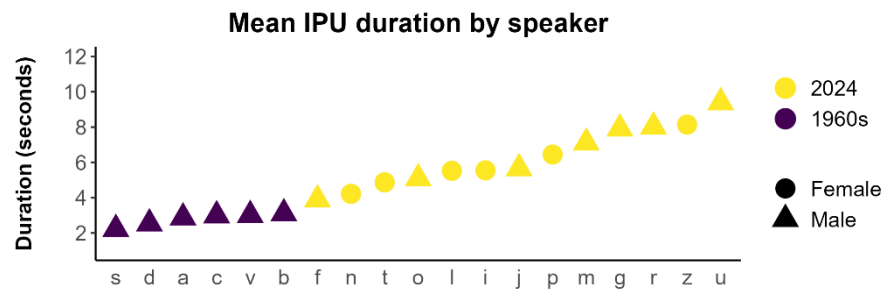


Figure 3: Mean IPU duration (in seconds) by speaker, era and gender.

References

- [1] C. Cotter, "Prosodic aspects of broadcast news register", in *Proceedings of the Annual Meeting of the Berkley Linguistic Society*, February 1993, pp. 90-100.
- [2] E. Strangert, "Prosody in public speech: Analyses of a news announcement and a political interview", in *Proceedings of INTERSPEECH 2005*, Lisbon, September 2005, pp. 3401-3404.
- [3] L. Castro, B. Serridge, J. Moraes, and M. Freitas, "The prosody of the TV news speaking style in Brazilian Portuguese", in *Proceedings of the Third ISCA workshop on Experimental Linguistics*, Athens 2010, pp. 17-20.
- [4] J. Di Napoli, "Functions of pause in Italian television news broadcasts", in *Proceedings of the 1st International Seminar on the Foundations of Speech – Breathing, Pausing and the Voice*, December 1-3, 2019, Sønderborg, Denmark, pp. 109-111.
- [5] P. Pik Li Mok, H. S. H. Fung, J. Li, "A preliminary study on the prosody of broadcast news in Hong Kong Cantonese", in *Proceedings of Speech Prosody*, Dublin, 2014, pp. 1072-1075.
- [6] E. Rodero and L. Cores-Sarria, "Best prosody for news: a psychophysiological study comparing a broadcast to a narrative speaking style", *Communication Research*, vol. 50, no. 3, pp. 361-384, 2023.
- [7] M. Pettorino, E. Pellegrino, L. Salvati, M. Vitale, A. De Meo, "La voce dei media. Trapianti ritmico-intonativi per un'analisi diacronica dell'italiano parlato", in *Atti del XVI Congresso Internazionale della Società di Linguistica Italiana*, 2012, pp. 777-791.
- [8] M. Savino, S. Wehrle and M. Grice, "The prosody of Italian newsreading: a diachronic study", in *Proceedings of Speech Prosody*, Leiden, The Netherlands, 2024, pp. 836-840.
- [9] S. Wehrle, *Conversation and intonation in autism: A multi-dimensional analysis*. Berlin: Language Science Press, 2023.
- [10] S. Wehrle, "A brief tutorial for using Wiggleness and Spaciousness to measure intonation styles", OSF; retrievable at osf.io/5e7fd, 2022.
- [11] <https://parlaritaliano.studiumdipsum.it/it/650-corpus-di-parlato-telegiornalistico-anni-sessanta-vs-2005>
- [12] C.-H., Hsu, J. P., Evans, and C.-Y. Lee, "Brain responses to spoken F0 changes: Is H special?", *Journal of Phonetics* vol. 51, pp. 82-92, 2015.

Does Time Change in Time?

Geographical and Generational Variation of Pre-boundary Lengthening in German

Nadja Spina¹ & Alfred Lameli¹

¹Research Center *Deutscher Sprachatlas*

Keywords: Pre-boundary lengthening, prosody, variation and change, dialects, German

Background: Pre-boundary lengthening (PBL), the lengthening of segments immediately preceding prosodic boundaries, is a major correlate of prosodic phrasing ([1]). Language-specific variation in PBL is well studied, but only few studies on cross-regional variation in PBL exist, and corpus studies are lacking. German dialects are endangered due to increasing contact with Standard German, leading to the emergence of regiolects, which are closer to Standard German and cover larger geographical areas ([2]). This diachronic change is well studied for segmental dialect features but has scarcely been investigated with regard to prosody. The present study therefore conducted a corpus analysis of regional variation in PBL. It compared different speaker generations to explore diachronic change according to the *apparent time hypothesis* ([2, 3, 4]), which states that, if diachronic change is at work, older generations represent an earlier language stage and younger generations a later stage. Synchronic variation between generations is thus interpreted as diachronic change. ([2, 4, 9]).

Corpus Analysis: Speech recordings from the REDE-corpus (project “Regionalsprache.de”, www.regionalsprache.de [5]), a corpus of dialect recordings from 148 locations in Germany, were selected for analysis. Data from two Alemannic and two Central Bavarian locations were selected, and all speakers provided for each location were included in the analysis (9 AL, 11 CB). 7 old (~ 65+ years), 8 middle-aged (~ 50 years) and 5 young (~20 years) speakers heard 40 sentences that consisted of several intonation phrases (IPs), such as in (1), once in Standard German and once in dialect. Their task was to repeat the sentences, one time targeting their native dialect (dialect register) and another time Standard German (regiolect register). Each speaker thus produced 80 sentences (1600 recordings in total). Segment and word boundaries were automatically annotated using WebMAUS [6]. Syllable boundaries, syllable positions and main stress within each word were manually annotated in Praat. The distance of each segment and syllable from an IP-boundary was annotated using an automated procedure. Phrase-final words were manually annotated by labelling boundary tones after the GToBI convention [7]. The duration of each segment was measured. Linear mixed effects models accounting for SEGMENT DURATION as a function of BOUNDARY, REGION and GENERATION were fitted to the data in R. SPEAKER, LEMMA and SENTENCE were included as random effects.

Results and Discussion: Figure 1 shows a highly significant PBL-effect ($p < 0.001$) for segments in the ultimate syllable of disyllabic words (e.g., *WOchen*, ‘weeks’) in both dialects across all generations. Segments in the (stressed) penultimate syllable were less significantly lengthened by young, even less so by middle-aged and insignificantly by old speakers. Figure 2 shows the difference between final and non-final mean segment duration in percent. Segments are clearly lengthened more in ultimate than in penultimate syllables. This increase in PBL towards the boundary confirms experimental results for Standard German ([8]). Figure 2 reveals striking generational differences. In general, there seems to be a continuum of PBL, with less PBL at a dialect pole and more PBL on a standard pole. Considering the dialect data, young Alemannic speakers produce much more PBL than old and middle-aged speakers. That is, they abruptly admerge towards standard. In Central Bavarian, however, all three generations produce roughly the same amount of PBL, adverting successively towards standard. This is in line with a stable diachronic dialect intensity in Bavarian that [(9)] found for the REDE-data. Compared to the regiolect data, young speakers produce PBL to the same extent in both registers in Alemannic. In Bavarian, though, the extent of PBL considerably differs in young speakers. This suggests that, due to the stable dialect intensity, young speakers have a wider repertoire of dialect- and standard-targeted extents of PBL and shift between different speech situations [(9)].

- (1) [Im Winter]_{IP} [fliegen die trockenen Blätter]_{IP} [durch die Luft herum]_{IP}.
 ‘In winter, dry leaves fly about through the air.’
 [Es hört gleich auf]_{IP} [zu schneien]_{IP}, [dann wird das Wetter wieder besser]_{IP}.
 ‘It will soon stop to snow, then the weather will become better again.’
 [Der gute alte Mann]_{IP} [ist mit dem Pferd durchs Eis gebrochen]_{IP} [und in das kalte Wasser gefallen]_{IP}.
 ‘The good old man has broken through the ice with his horse and fallen into the cold water.’

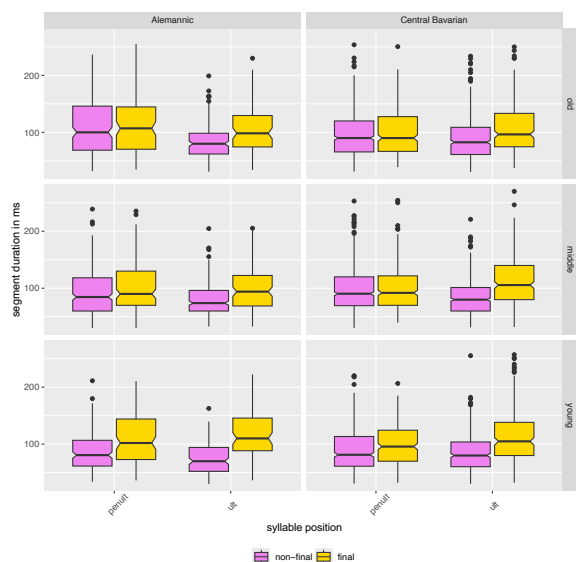


Figure 1: PBL-effect in milliseconds compared across regions and generations in disyllabic words (e.g., *WOchen*, ‘weeks’) across speakers

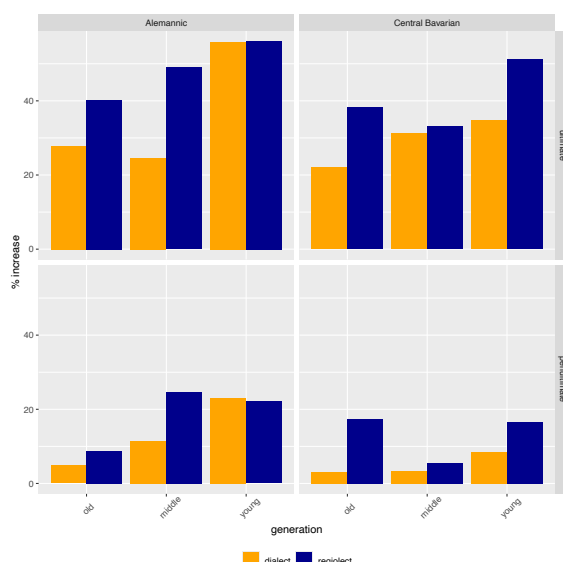


Figure 2: Strength of PBL-effect in percent compared across regions and generations in disyllabic words (e.g., *WOchen*, ‘weeks’) across speakers

- [1] Frazier, L., Carlson, K. & Clifton, C. 2006. Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences* 10/6, 245–249.
- [2] Schmidt, J. E., Herrgen, J. 2011. *Sprachdynamik. Eine Einführung in die moderne Regionalsprachenforschung*. Berlin: Schmidt.
- [3] Labov, W. *Principles of Linguistic Change. Volume 1: Internal Factors*. Oxford, Cambridge: Blackwell.
- [4] Lameli, A. 2004. *Standard und Substandard. Regionalismen im diachronen Längsschnitt*. Stuttgart: Steiner.
- [5] Schmidt, J. E., Herrgen J., Kehrein, R., Lameli, A., & Fischer, H. 2020 (Eds.): *Regionalsprache.de (REDE III). Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Marburg: Forschungszentrum Deutscher Sprachatlas.
- [6] The Munich Automatic Segmentation System (MAUS), web service, version 3.12, URL: <https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic>.
- [7] Grice, M. & Baumann, S. 2002. Deutsche Intonation und GToBI. *Linguistische Berichte* 191: 267–298.
- [8] Schubö, F. & Zerbian, S. 2020. Phonetic content and phonological structure affect pre-boundary lengthening in German. *Proceedings of the 10th International Conference on Speech Prosody* (Tokyo, Japan), 111–115.
- [9] Lameli, A. 2025. Gesprochenes Deutsch in den Regionen. Eine Standortbestimmung für die Bundesrepublik Deutschland. In Dannerer, M., Deppermann, A., Proske, N. & Weber, T. (Eds): *Gesprochenes Deutsch. Struktur, Variation, Interaktion*. Berlin, Boston: De Gruyter.

Introduction. In intonational research, clustering has recently been developed as a tool for analysis of F0 contours [1]. It can facilitate data exploration without imposing top-down categorisation, and is thus particularly suited for assessment of patterns in un(der)-documented language varieties, or for patterns and categories which defy easy classification [1,2]. Given the myriad of clustering techniques and related clustering parameters available, it is important to assess the impact of analyst choices. In line with [3], we compare how two clustering algorithms (k-means (KM) & hierarchical agglomerative clusters (HC)) capture focus distinctions in Afro-Mexican Spanish, a marginalized and under-researched variety. We ask: (RQ1) How evenly distributed are trajectories across clusters for each method?; (RQ2) To what extent do algorithms agree on the optimal number of clusters?; and (RQ3) How do emergent distinctions map to focus (broad v. narrow)?

The data. 11 participants read aloud 12-syllable sentences containing target words in narrow and broad focus. The target syllable contained the tonic vowel /a/ preceding /l/, /n/, or /s/. Syllable aperture was manipulated through the inclusion of /t/, e.g., *manto* (/ˈman.to/, ‘cloak’) vs *mano* (/ˈma.no/, ‘hand’). Target words received the 3rd and penultimate pitch accent. F0 was sampled at 20 time-normalised intervals within each syllable (40 across the tonic and post-tonic syllable). F0 trajectories were z-scored within speaker with 1,467 contours analysed (Panel D). **Parameters of analysis and evaluation.** KM was done using [3], using Euclidean distance with Gower adjustment. HC was performed with the stats package in base R [4], using Euclidean distance as a distance metric and complete linkage as linkage criterion, these HC settings have been used by default in [1,2]. In cluster assessment, for both KM & HC we assess 2-6 clusters and evaluate them according to RMSD (which when larger indicates better separation) and within- and between-cluster variance (lower within- and higher between-cluster variance indicates compact and well-separated clusters). We plot the metrics across the *number of clusters* to identify *elbows* in the plots, i.e. increases in cluster number associated with minimal or no improvement in these measures. We also examine the distribution of trajectories across clusters, for both algorithms.

Results. There is disagreement across algorithms: for KM 4 clusters is optimal, and for HC there is no clear *elbow*, with data skewed at Cluster 4. Examining the counts of contours per cluster, only 6 contours belong to HC4 with contours more evenly distributed for KM4 (Panel A). Analyses were re-run with the 6 extreme trajectories removed (N = 1,461) with RMSD and variance metrics suggesting 4 clusters as optimal for both algorithms (Panel C). Both with and without the extreme cluster trajectories, contours remain more evenly distributed for KM than for HC (e.g., KM4 incl. extreme contours, N = 199; KM4 exc. extreme contours, N = 197). Notably, while 4 clusters is considered optimal, the actual shape of the four clusters differs across methods (Panel B). In comparison, the 3-cluster solution is more comparable in terms of emergent cluster shapes. We thus consider the mapping between emergent clusters and focus for both 3 and 4 cluster solutions. This reveals similarities across the algorithms with narrow focus conditions are overwhelmingly accounted for by one cluster: a sharp tonic rise and steep post-tonic fall (KM4: 56.02%, HC4: 75.26%) (Panels B & E). This pattern holds for with both 3 and 4 clusters (Panel E). Outwith narrow focus, algorithms show variation in the cluster(s) associated with broad focus. For HC, broad focus may be represented by either a low plateau (Cluster 2, 37.73%) or a rise and fall (Cluster 3, 40.88%). For KM, however, broad focus is most likely associated with just the low plateau (Cluster 2, 43.62%) (Panels B & E).

Discussion. Results highlight similarities and differences in KM & HC. Excluding extreme trajectories, both methods signal 4 clusters as the optimal solution with narrow focus conditions accounted for by Cluster 3 (a tonic peak and post-tonic fall). Differences arise however in the inclusion of extreme trajectories. Whilst KM shows uniformity, with 4 clusters optimal regardless of extreme trajectories, HC is more susceptible to influence from these extreme datapoints, resulting in very small clusters which make evaluation metrics harder to interpret. In the present case, this suggests that HC requires more care given to outlier inspection and removal than KM. Whilst this sort of data cleaning is an important tool, it should be guided by prior knowledge such that important datapoints are not unnecessarily removed. Given the unexplored nature of Afro-Mexican prosody, this is not possible and removing datapoints could lead to the removal of important, previously undocumented variation. We further maintain that, having examined the extreme trajectories, these were not F0 tracking errors but rather a small number of high rising trajectories in the upper frequency ranges. They may therefore be contours rarely used in the variety, or instead part of a particular speaker’s inventory. It is therefore important to include them in the phonological analysis and interpretation of the clustered contours for the wider documentation purposes. Secondly, whilst algorithms converge in the mapping of narrow focus, they differ for broad focus, i.e., there is not a binary dichotomy between broad and narrow focus for either method. Instead, intermediary categories emerge. Whilst variation in broad focus is expected for non-Afro Mexican Spanishes [5], whether there is similar mapping onto these intermediary categories & how the methods differ in this requires further investigation. Moreover, initial observations indicate idiosyncratic behaviour, with certain speakers showing range variation in focus mapping (narrow focus associated with high pitch range, and broad low). We therefore propose running speaker-specific clustering models and examining the makeup of the clusters in the intermediary categories, principally to assess the role of segment and syllable structure which has shown to be significant in peak alignment in this variety [6]. We will also examine the impact of using different linkage criteria for HC.

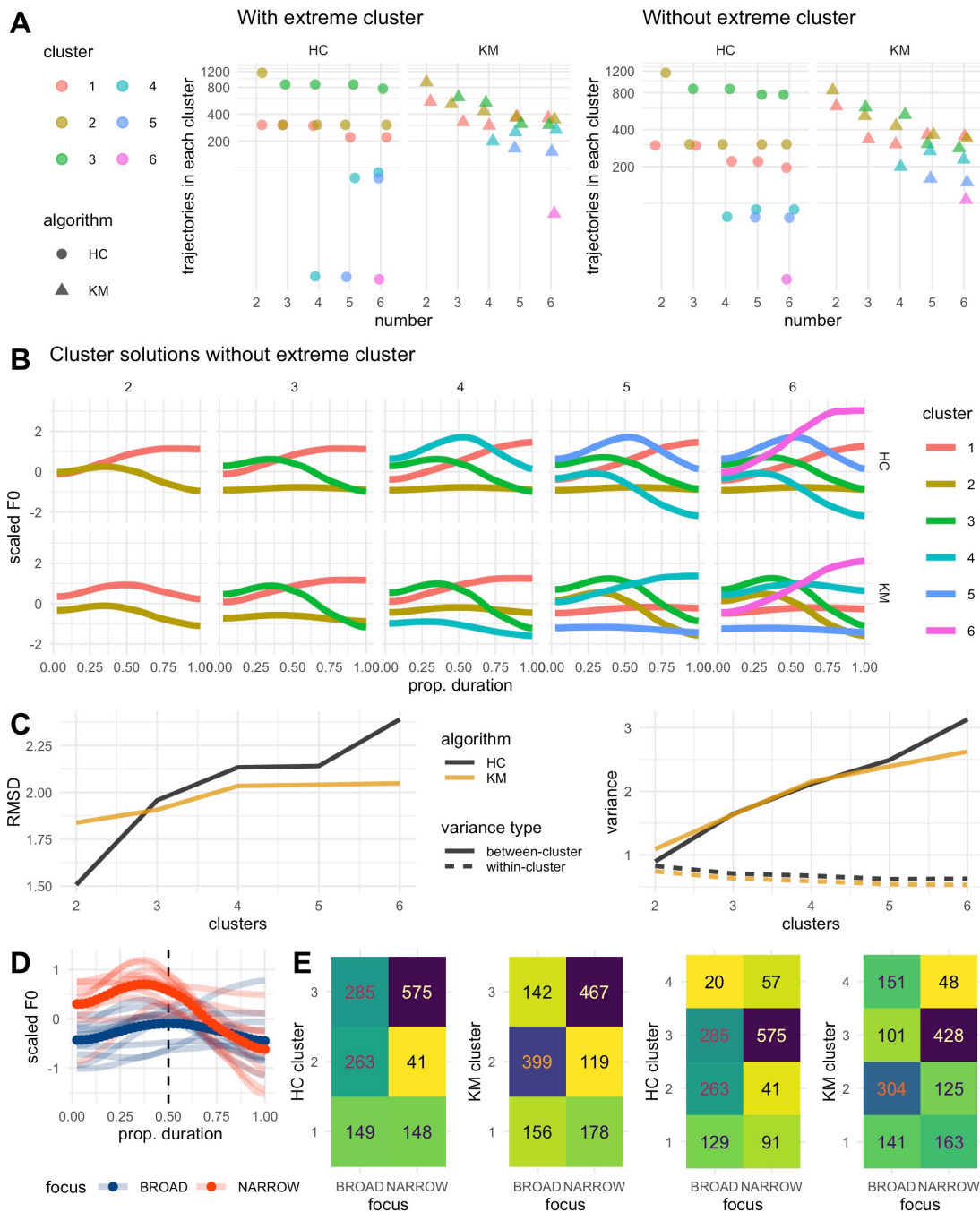


Figure: A: The counts of trajectories within each cluster (y axis, logarithmic scale) for cluster solutions containing 2-6 clusters (x axis), for both methods (HC and KM, labelled at top). At left the inclusion of extreme trajectories leads to the appearance of a small cluster (6 trajectories) from 4 clusters onwards for HC. With these extreme trajectories removed at right, this occurs at 6 clusters for HC. Note that k-means maintains relatively evenly distributed clusters in both cases. **B.** cluster means for 2-6 clusters (columns) and both algorithms (rows). **C.** Evaluation metrics for the clustering solutions without the extreme trajectories. The numbers identifying clusters are the same as in panel B. **D.** The focus conditions over normalised time with speaker means (light lines) and overall means. **E:** Heat maps showing the relation between focus condition (x axis) and cluster for both methods and 3 and 4 cluster solutions. The number of trajectories for each cluster and focus combination is shown in each cell. The numbers identifying clusters are the same as in panels B and C.

References: [1] Kaland, C. (2023). Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *Journal of the International Phonetic Association*, 53(1), 159-188. [2] Steffman, J., Cole, J., & Shattuck-Hufnagel, S. (2024). Intonational categories and continua in American English rising nuclear tunes. *Journal of Phonetics*, 104, 101310. [3] Kaland, C., Steffman, J., & Cole, J. (2024). K-means and hierarchical clustering of f0 contours. In *Proc. Interspeech 2024* (pp. 1520-1524). Genolini, C., & Falissard, B. (2010). KmL: k-means for longitudinal data. *Computational Statistics*, 25(2), 317-328. [4] [25] R Core Team, "R: the R project for statistical computing," 2022, version 4.2.1. [5] de la Mota, C., Butragueño, P. M., Prieto, P., and Fabra, P. (2011). Mexican Spanish Intonation Mexican Spanish Intonation. In Prieto, P. and Roseano, P., editors, *Transcription of Intonation of the Spanish Language*, pages 319-359. LINCOM Europa, Munich. [6] Marchini, G. (2024). Peak alignment in Afro-Mexican Spanish: an exploratory analysis. In *Speech Prosody 2024 Conference Proceedings*, Leiden: Netherlands.

Key words: intonation, language documentation, methods

The dynamical structure of the nuclear tune: The phrase accent

Khalil Iskarous¹, Thomas Sostarics², Jennifer Cole²

¹University of Southern California, ²Northwestern University

This study investigates dynamic pitch properties in the interval between the nuclear accented syllable and the end of the intonational phrase in Mainstream American English (MAE), e.g., the underlined syllables in “Only Melanie ran a mile”. In the Autosegmental-Metrical account, pitch in this region is governed by the *phrase accent*, which follows the nuclear pitch accent and can precede the boundary tone [1, 2, 3]. The question we ask is whether there is grouping structure within the nuclear tune, which combines two of the three pitch events (pitch accent, phrase accent, and boundary tone) into a coherent structure, similarly to how a foot combines two syllables. There are 3 hypotheses: A) The three pitch events are separate entities with no internal grouping; B) Pitch accent and Phrase accent cohere together into a unit separate from the boundary tone; C) Phrase accent and Boundary tone cohere together into a unit, separate from the pitch accent. The motivation for our question is that aggregation of cognitive objects into groups has been argued to be a fundamental aspect of neural computation [4], therefore it is important to understand the organizational structure of a nuclear tune. To test these hypotheses, we apply the measure of phase coherence from dynamical systems theory [5] to pitch trajectories from an intonation imitation task [5] to probe how listeners-speakers structure the nuclear tune by testing the *stability* of timing of pitch events. **Methods:** Auditory stimuli were based on three tunes (H*L- H%, H*L-L%, and L*H-H%) produced by 2 speakers, resynthesized to create four versions of each tune that vary in the slope and shape of the pitch movement following the accentual target (Figure 1): One with a sharp inflection (fall or rise) and sustained pitch (low or high) immediately following the accentual target; two with shallower and later inflections, and one with a linear interpolation (fall or rise) to the phrase-final pitch. These were resynthesized on short nuclear intervals (4 syllables: *Only Melanie ran*) or long ones (6 syllables: *Only Melanie ran a mile*). Participants (N=37) imitated 144 tunes on novel sentences. Variation in the pitch inflection points from earlier to later acts as a temporal perturbation of the produced timing of pitch events. We are interested in patterns of coordinated stable timing in imitations of our variable stimuli as evidence for internal grouping of pitch events in the nuclear interval, where . timing between events was measured as lag between velocity maxima/minima [6] of pitch events. **Predictions:** Hyp. A predicts lack of stable relative timing of pitch events; Hyp. B predicts phase coherence in a Pitch accent-Phrase accent group, measured as a low-variability phase lag and short duration between these two pitch events; Hyp. C predicts phase coherence of a phrase accent-boundary tone group. **Results:** Figure 2 (left) shows the mean F0 velocity profiles, and Figure 2 (right) shows the phase lags between Phrase accent and Pitch accent (red) and Boundary tone and phrase accent (blue) for each of the tunes with vertical lines marking medians of the distributions. The lag between pitch and phrase accents is consistent across tunes and about half the lag between phrase accents and boundary tone. In contrast, the boundary tone lag is more variable across tunes and exhibits higher interquartile ranges within each tune. Velocity profiles also show a period of near 0 velocity, or equilibrium, following the phrase accent peak velocity. From a dynamical point of view, this pattern suggests that there is a complex dynamical action composed of the pitch *and* phrase accent events that reaches equilibrium (i.e., finishes) long before the boundary tone. **Discussion:** The results support Hyp. B. Pitch accents and phrase accents are more tightly coordinated, based on phase coherence, suggesting a structure within the tune that groups the pitch accent and phrase accent into one dynamical unit.

References

[1] Pierrehumbert, J. B. (1980). The phonology and phonetics of English intonation (Doctoral dissertation, Massachusetts Institute of Technology). [2] Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255–309. [3] Grice, M., Ladd, D. R., & Arvaniti, A. (2000). On the place of phrase accents in intonational phonology. *Phonology*, 17(2), 143–185. [4] Taylor, P., Hobbs, J.N., Burrioni, J., Siegelmann, H.T. (2015). The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions. *Scientific Reports*, 5:18112 | DOI: 10.1038/srep18112. [5] Jantzen, K.J. and Kelso, J.S. (2007). Neural Coordination Dynamics of Human Sensorimotor Behavior: A Review. In Viktor K. Jirsa and A.R. McIntosh (eds.) *Handbook of Brain Connectivity*. Springer Verlag. [6] Browman, C. P., & Goldstein, L. (1985). Dynamic modeling of phonetic structure. In V. A. Fromkin (Ed.), *Phonetic Linguistics* (pp. 35–53). New York, NY: Academic Press.

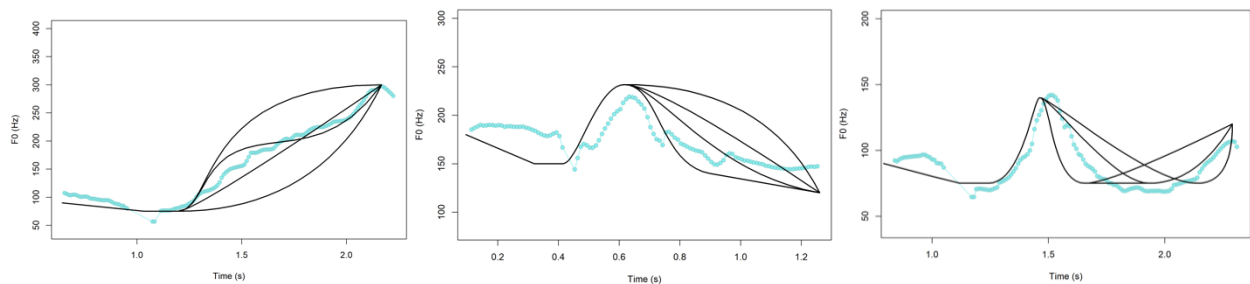


Figure 1: Resynthesized trajectories for stimuli: Rises, Falls, and RFRs. Green contour shows original pitch contour of the source recording used for resynthesis.

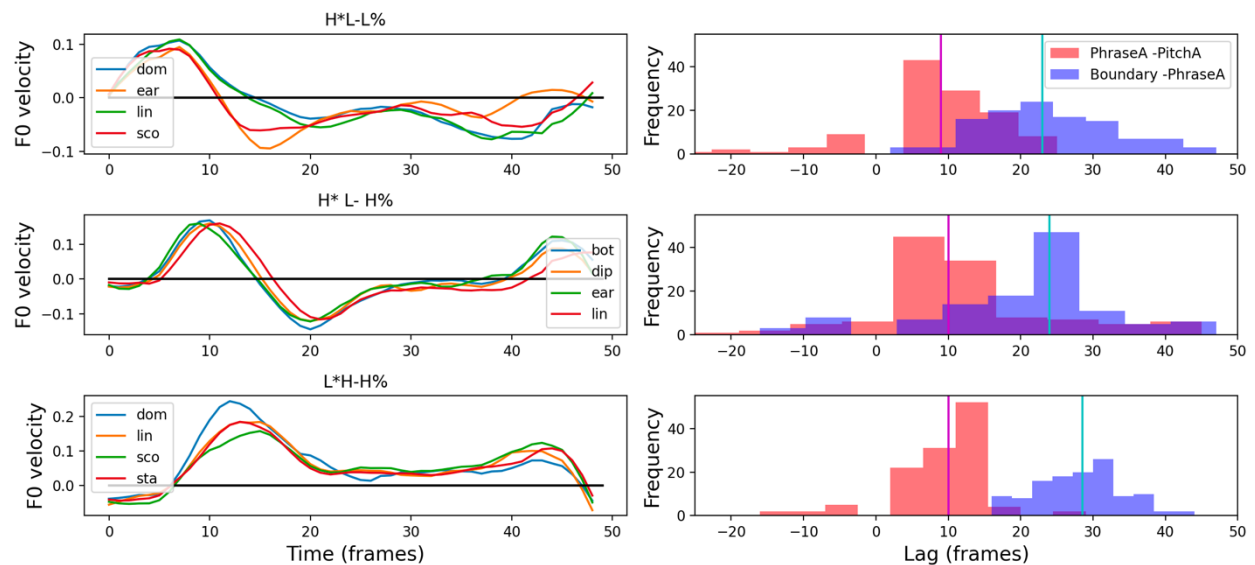


Figure 2: (Left) Mean F0 Velocity production profiles for the four resynthesized trajectories for each tune. (Right) Lag distributions between velocity extrema.

Informativity and the actuation of tonal coarticulation in Taiwan Southern Min and Taiwan Mandarin

Po-Hsuan Huang, Stephanie S. Shih
University of Southern California
{pohsuan, shihs}@usc.edu

3rd International Conference on Tone and Intonation, Herrsching, Germany, May 16–18, 2025

Keywords tonal coarticulation; informativity; actuation; Taiwan Southern Min; Taiwan Mandarin

Introduction & background The distribution of tonal coarticulation (TC) and its nature have long been studied and debated. Typologically, carryover effects are found to be assimilatory and stronger, and anticipatory effects dissimilatory and weaker (Chang & Hsieh, 2012). Discrepancies across languages, however, have also been attested (Peng, 1997; Wang, 2002; Brunelle, 2009; Huang, 2023). This raises the question of whether TC is subject to language-invariant biomechanical needs (e.g., Shen, 1992; Huang) or to some extent conditioned by phonological constraints (e.g., Brunelle, 2009). In Brunelle, Northern Vietnamese, which uses both F0 and laryngealization for tonal contrast, was found to exhibit higher TC magnitudes than Southern Vietnamese, where only F0 is used, supporting the phonological account. In our previous study (Huang, 2023) of Taiwan Southern Min (TSM) and Taiwan Mandarin (TM), on the other hand, comparable magnitudes of TC were found, despite the larger tone inventory in TSM and the subsequent higher possibility of perceptual confusion. In this study, we provide perspectives from information theory (Shannon, 1948) and show that while past studies of TC focus mainly on production and perception, functional-cognitive aspects are also involved. We thus suggest that the lack of production difference in Huang (2023) is likely due to neglect of these aspects. Specifically, a sizable number of studies (e.g., Stokes & Surendran, 2005; Seyfarth, 2014; Cohen Priva, 2017; Kamierski, 2023) have shown that information-theoretic constraints provide an explanation for systematic variation in speech. Crucially, Cohen Priva (2017) found that informativity outperforms other constraints in predicting the actuation of lenition across languages, with less informative targets more readily lenited. Since TC, like lenition and other kinds of systematic variation, involves the pull between articulatory costs and contrastive needs, it is likely to also be conditioned by the local informativity of the coarticulating tone pairs. In this study, we put forth evidence from TSM and TM and argue that informativity is at least partially responsible for the discrepancies in TC attested in the literature. In addition, we propose that such cognitive constraints interplay with linguistic constraints, such as tone inventory sizes.

Methodology The data in this study comes in two parts: 1) the F0 data of coarticulated tone pairs in TSM and TM and 2) the informativity of these tone pairs collected from a TSM corpus and a TM corpus. Linear mixed-effects models (LMMs) were then used for statistical analyses. **F0 data** Eleven TSM native speakers and fourteen TM native speakers (15 females; 20–27 y.o., mean=22.44) participated in the production experiment. In TM, all (4) tones were investigated. In TSM, checked tones were excluded due to the inherently shorter durations, leading to 5 chosen tones. For both languages, one disyllabic word was chosen for each tone pair. There were a total of 16 (4×4) TM and 25 (5×5) TSM tone pairs. There were 10 repetitions for each word. The participant was instructed to produce the stimuli in a relaxed manner. The audio was recorded at a sampling frequency of 4.41k Hz. Syllable boundaries were manually labeled in *Praat* (Boersma & Weenink, 2018), and F0 values were extracted using *Parselmouth* (Jadoul et al., 2018). Within each syllable, F0 values were sliced into 11 portions. The means of the last F0 portion in the preceding syllable and the first F0 portion in the following syllable were taken as their respective onset/offset. **Informativity** To estimate the informativity of each tone pair, one corpus was built for each language. The TSM corpus was taken from the TAT.MOE Corpus (Liao, 2022), with word boundaries and tones determined with the Tâi-lô orthography provided. Disyllabic words were chosen as the final corpus, comprising 306,001 words. The TM corpus was built with 21,668 posts on a Taiwan BBS forum PTT. The texts were first segmented with *CKIP segmenter* (Tsai & Chen, 2004). Disyllabic words were chosen as the final corpus, comprising 3,717,386 words. *pinyin* was used to label the tones of the words. Atonal characters were discarded. The informativity of the tone pairs was then calculated based on Cohen Priva’s (2017) formula with and without consideration of tone sandhi. We call these two types of informativity the *underlying* vs. *surface informativity*. **Statistical analyses** Two cross-language comparison models (underlying vs. surface informativity) were fitted. The target tone value was taken as the predicted value, and the context tone value and informativity and their interaction were taken as predictors, with participants and syllable onset/offset segment types as the grouping variables for the random intercepts/slopes. Four within-language models were also fitted to investigate the position (carryover vs. anticipatory) effects.

Results Significant positive effects of context tone values on target tone values were found in both the underlying ($p = 0.002$) and surface ($p = 0.020$) informativity models, suggesting assimilatory TC effects. Crucially, informativity was found to be inversely correlated with the degree of TC (both p ’s < 0.001). Linguistic differences were also attested: informativity effects were stronger in TSM (both p ’s < 0.001). Within-language models and follow-up models revealed that carryover effects are stronger in both TSM (both p ’s < 0.001) and TM (underlying: $p = 0.004$; surface: $p < 0.001$). Lastly, while informativity effects were found in both positions in TSM for underlying informativity (both p ’s < 0.001) and in the carryover position for surface informativity (p ’s < 0.001), they were only found in carryover positions in TM (both p ’s < 0.001). Illustrations are shown in Figures 1 and 2.

Discussion The inverse correlation between informativity and the magnitude of TC in the two languages shows that the actuation of TC is conditioned by functional constraints other than purely phonetic-phonological factors. Speakers in both languages, in general, avoid strong TC on more informative pairs and allow less informative ones to bear more coarticulation. However, linguistic differences were also found: informativity effects are stronger in TSM, while TM speakers are less sensitive to informativity. In Cohen Priva (2017), it is proposed that the informativity of an element indirectly entails the cost a language pays to maintain its faithfulness. Likewise, in Huang (2023), TSM has been seen to have comparable magnitudes of TC as TM despite the larger tone inventory, which is taken as evidence that biomechanical needs constrain TC. However, we argue in this study that, together, the results suggest that while the global articulatory costs that TSM and TM speakers pay are comparable, a more fine-grained local allocation of such costs is also at work. In TSM, with the larger tone inventory and the subsequent higher perceptual pressure, speakers strategically allocate costs based on how informative a tone pair is. On the flip side, in TM, perceptual confusion is less probable, and a fine-grained allocation of costs is of less importance.

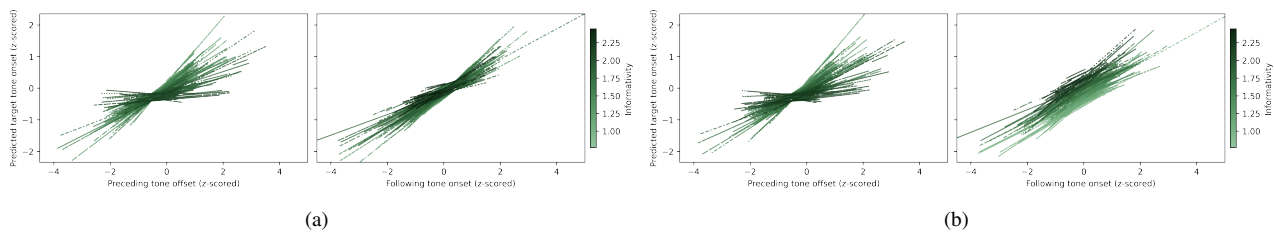


Figure 1: Fitted target tone values in carryover (left)/anticipatory (right) positions in TSM under the effects of underlying (a) and surface (b) informativity. The more informative tone pairs (darker lines) generally have smaller slopes (less TC).

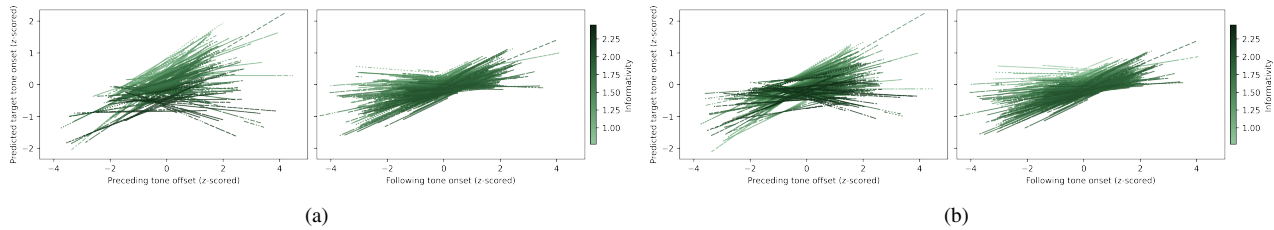


Figure 2: Fitted target tone values in carryover (left)/anticipatory (right) positions in TM under the effects of underlying (a) and surface (b) informativity. The more informative tone pairs (darker lines) generally have smaller slopes (less TC), though this is less pronounced in anticipatory positions.

References

- Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer. Version 6.2.14* [Computer Software]. Retrieved from <http://www.praat.org/>
- Brunelle, M. (2009, 01). Northern and Southern Vietnamese tone coarticulation: A comparative case study. *Journal of the Southeast Asian Linguistics Society*, 1, 49–62.
- Chang, Y.-C., & Hsieh, F.-F. (2012). Tonal coarticulation in Malaysian Hokkien: A typological anomaly? *The Linguistic Review*, 29, 37–73. doi: 10.1515/tlr-2012-0003
- Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language*, 93(3), 569–597.
- Huang, P.-H. (2023). *Perception and Production of Coarticulated Tones in Taiwan Mandarin and Taiwan Southern Min* (Master's thesis). National Taiwan University, Taipei, Taiwan.
- Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71, 1–15. doi: 10.1016/j.wocn.2018.07.001
- Kamierski, K. (2023). The role of informativity and frequency in shaping word durations in English and in Polish. *SSRN*. (Available at SSRN 4496718)
- Liao, Y.-F. (2022). *TAT_MOE corpus*. (Available online at <https://tgg1.naer.edu.tw>. Retrieved on Oct. 9, 2024.)
- Peng, S.-H. (1997). Production and perception of Taiwanese tones in different tonal and prosodic contexts. *Journal of Phonetics*, 25(3), 371–400. doi: 10.1006/jpho.1997.0047
- Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1), 140–155. doi: 10.1016/j.cognition.2014.06.013
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423. doi: 10.1002/j.1538-7305
- Shen, X. (1992). On tone sandhi and tonal coarticulation. *Acta Linguistica Hafniensia*, 25(1), 83–94. doi: 10.1080/03740463.1992.10412279
- Stokes, S. F., & Surendran, D. (2005). Articulatory complexity, ambient frequency, and functional load as predictors of consonant development in children. *Journal of Speech, Language, and Hearing Research*, 48(3), 577–591. doi: 10.1044/1092-4388(2005/040)
- Tsai, Y.-F., & Chen, K.-J. (2004). Reliable and cost-effective PoS-Tagging. *IJCLCLP*, 9(1), 83–96.
- Wang, H. (2002). The prosodic effects on Taiwan Min tones. *Language and Linguistics*, 3, 839–852.

Two-stage tonal encoding in Mandarin word production: Evidence from computational simulations

In Mandarin word production, there are two contrasting views regarding the mechanisms of tonal encoding. The two-stage model assumes that the lexical tone is selected first at the early stage of production, and then integrated with the atonal syllable at the later stage [1,2], while other researchers proposed that the lexical tone is retrieved only at the later stage of production [3]. Studies have used phonologically priming tasks to address this issue and reported facilitation effects on naming latencies under both the homophonous prime condition (both syllable- and tone-related) and the syllabic prime condition (syllable-related and tone-unrelated) but interference effects under the tonal prime condition (syllable-unrelated and tone-related) compared to the unrelated prime condition (both syllable- and tone-unrelated) [1,4,5]. There are two tentative explanations for this interference effect. One is that speakers prepare the tonal frame in advance with a dummy atonal syllable but later revise the dummy content to the target syllable [4]. The other is based on the competition mechanism whereby the tonal prime may activate a different existing syllabic frame and cause stronger competition during the later tone-to-syllable integration [1]. Further verification is needed to understand the nature of tonal encoding in Mandarin word production. Thus, this study performed computational simulations to uncover the mechanisms underlying the facilitation and interference effect on naming latencies observed in previous primed picture naming studies, which intended to verify the two different theoretical accounts on tonal encoding in Mandarin spoken word production.

Two different models of tonal encoding were built respectively. One was based on the two-stage model of tonal encoding (Fig.1a~1e) and the other was based on the assumption that the tone is only retrieved at the later stage of speech production (Fig.2). We examined which model could better simulate the effects in the phonologically primed picture naming task, including the homophonous, syllabic, tonal, and unrelated prime conditions. Activation spread between different layers through bidirectional excitation connections. Nodes in the word layer were connected by bidirectional inhibitory links to implement lexical competition. The naming response was produced when activation of the syllable motor program node reached a specific threshold. The number of cycles needed to reach this threshold constituted the model's reaction time (akin to naming latency). After validating the two-stage model for Mandarin tonal encoding, we further built another two models to investigate the prerequisite for the interference effect under the tonal prime condition, based on the syllable competition mechanism (Fig.1d) and the dummy syllable mechanism (Fig.1e), respectively. The activation of the syllable motor program node changing over time was plotted as curves.

The late retrieval model revealed no differences between the homophonous and syllabic prime conditions or between the tonal and unrelated prime conditions (Fig.3), incompatible with the empirical findings. In contrast, for the two-stage model, we successfully simulated facilitation effects under both homophonous and syllabic prime conditions, with the former eliciting relatively larger facilitation effects than the latter (Fig.4, Fig.5), consistent with previous findings [1,4,5]. We also generated the interference effect under the tonal prime condition in the two-stage model with either the competition mechanism (Fig.4) or the dummy syllable mechanism (Fig.5), by decreasing the excitation rate between the tonal frame and the syllable motor program, indicating a relatively slower tone-to-syllable integration process. These results support the two-stage model of tonal encoding in Mandarin word production, while both the competition mechanism and the dummy syllable mechanism appear appropriate. Decreased activation between the tonal frame and the syllable motor program, which implies slower tone-to-syllable integration, appears to be a prerequisite for generating the interference effect of tonal overlap without shared syllabic information. This study contributes to a broader understanding of the two-stage tonal encoding model by highlighting the dynamic interaction between tonal and syllabic representations in Mandarin spoken word production. Future research could validate the current model using different tonal languages.

Keywords: Lexical tone; Tonal encoding; Word production; Mandarin Chinese; Computational simulations

References

- [1] Chen, X., & Zhang, C. (2025). Setting the “tone” first and then integrating it into the syllable: An EEG investigation of the time course of lexical tone and syllable encoding in Mandarin word production. *Journal of Memory and Language*, 140, 104575.
- [2] Roelofs, A. (2015). Modeling of phonological encoding in spoken word production: From Germanic languages to Mandarin Chinese and Japanese. *Japanese Psychological Research*, 57(1), 22–37.
- [3] O’Searghda, P. G., Chen, J. Y., & Chen, T. M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, 115(2), 282–302.
- [4] Chen, J. Y., Chen, T. M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language*, 46(4), 751–781.
- [5] Zhang, Q. (2008). Phonological encoding in monosyllabic and bisyllabic Mandarin word production: Implicit priming paradigm study. *Acta Psychologica Sinica*, 40(03), 253–262.

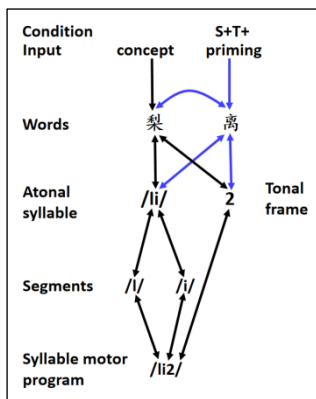


Fig.1a. Two-stage model (S+T+)

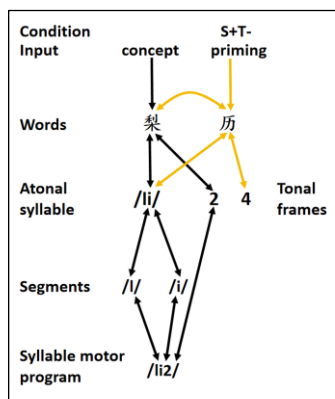


Fig.1b. Two-stage model (S+T-)

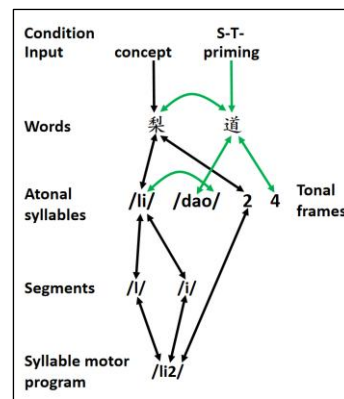


Fig.1c. Two-stage model (S-T-)

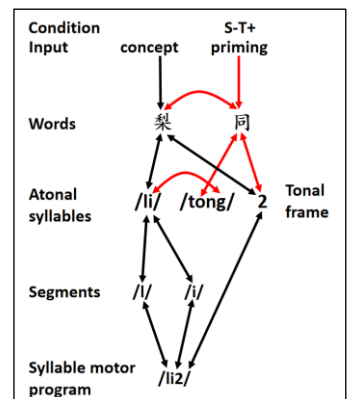


Fig.1d. Two-stage model (S-T+, competition)

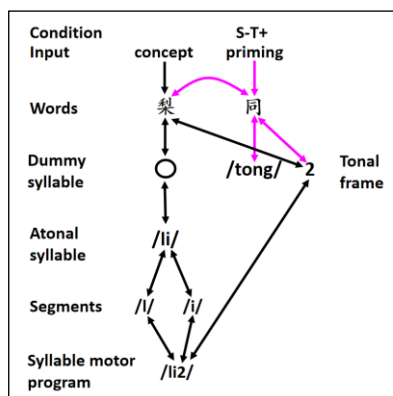


Fig.1e. Two-stage model (S-T+, dummy syllable)

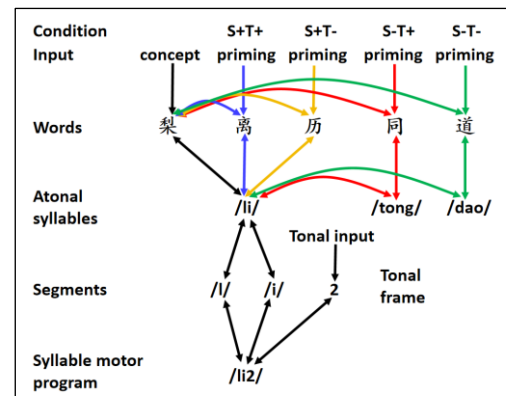


Fig.2. Late retrieval model with the tone only retrieved in the later stage

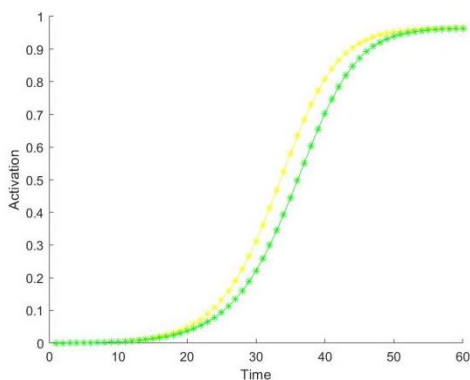


Fig.3. Results of the late retrieval model

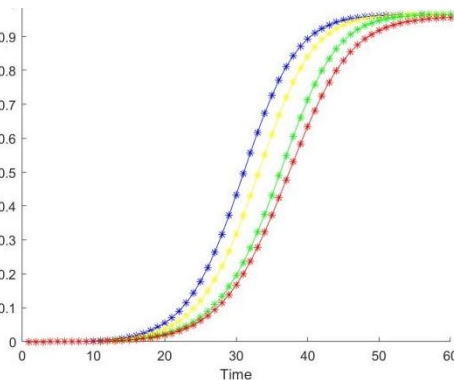


Fig.4. Results of the two-stage model (competition)

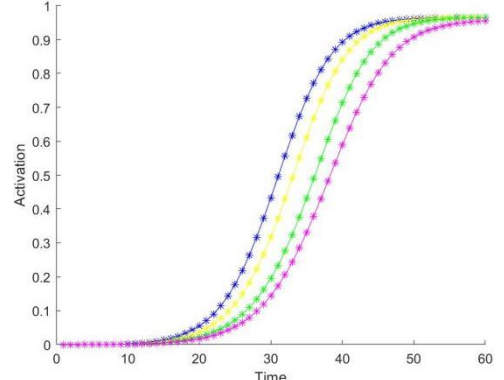


Fig.5. Results of the two-stage model (dummy syllable)

Note: The color of curves represents the corresponding condition: blue: homophonous prime (S+T+); yellow: syllabic prime (S+T-); green: unrelated prime (S-T-); red: tonal prime (S-T+) with competition mechanism; violet: tonal prime (S-T+) with dummy syllable mechanism.

In Fig.3, the yellow curve (S+T-) overlaps with the blue one (S+T+), and the green curve (S-T-) overlaps with the red one (S-T+).

Acoustic correlates of stress in Kupang Malay: a preliminary analysis

Kirsten Culhane¹ and Constantijn Kaland²

¹ New Zealand Institute of Language, Brain and behaviour, University of Canterbury ²Institute of Linguistics - Phonetics, University of Cologne

Kupang Malay is one of several Malay varieties spoken in eastern Indonesia. It is impressionistically described as having word stress falling on either the penultimate or ultimate syllable of the word. Minimal pairs such as /'baraŋ/ 'west' vs. /ba'raŋ/ 'heavy' and /'paraŋ/ 'machete' vs. /pa'raŋ/ 'war' are reported [1]. No claims have been made regarding pitch accents. At present, acoustic correlates of stress have not been investigated quantitatively.

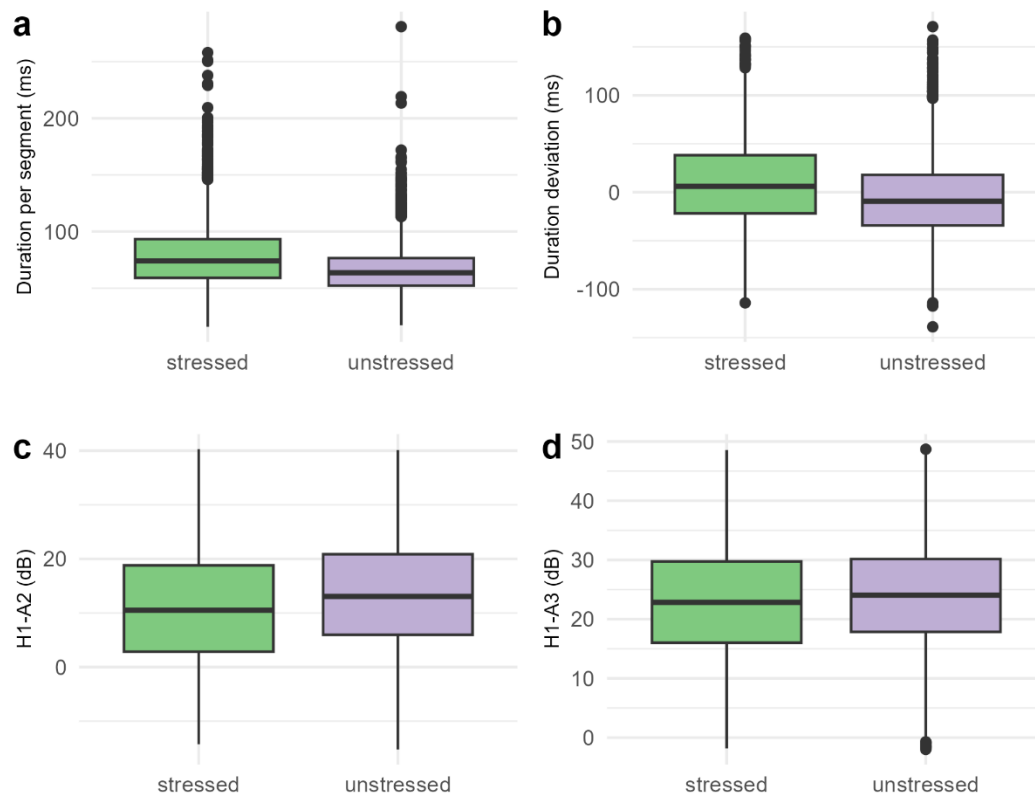
Little empirical work has been carried out on Malay varieties of Eastern Indonesia, and studies of these varieties are relevant to the broader discussion surrounding the nature of word prosodic systems of Malay. Stress claims vary considerably between varieties, and empirical studies show varied results: for some varieties demonstrate acoustic evidence for stress [2], while others do not [3]. "Stressless" varieties of Malay, however, have tended to receive more attention in the literature (e.g. Standard Indonesian [3, 4, 5]), and there is a pervasive assumption that varieties of Malay are typically stressless [6]. Given the lack of empirical work on the prosody of Kupang Malay, our study aims to fill this gap and contributes to the description of this underresearched language.

In this study we present a preliminary investigation of spectral tilt and duration as stress correlates in Kupang Malay. Duration was investigated because it appears to be the most consistent correlate of word stress cross-linguistically [7]. Raw duration was measured by taking the absolute duration of the entire syllable. From this, two derived measures of duration were computed: duration per segment and duration deviation. These measures were devised by [2] to be more accurate measures than raw duration, as they take syllable structure into account. Duration per segment is the duration of a syllable divided by the number of segments. Duration deviation is calculated by subtracting the mean duration of all syllables of a certain structure (e.g. CV, CVC) from the duration of a given syllable of that structure. Positive duration deviation indicates that a syllable is longer than the mean duration for syllables of that shape, while negative duration deviation indicates that a syllable is shorter than the mean duration for syllables of that shape. Spectral tilt measures capture intensity differences between low and high frequencies in the spectrum [8]. Typically, spectral tilt is shallower — i.e., there is less intensity roll-off towards the higher frequencies — in stressed syllables than in unstressed syllables [8]. Two measures of spectral tilt were taken, H1-A2 and H1-A3. H1-A2 refers to the intensity of the second formant (A2) subtracted from the intensity of the F0 (first harmonic; H1). H1-A3 refers to the intensity of the third formant (A3) subtracted from the intensity of the F0 (first harmonic; H1). Both spectral tilt measures were chosen as they were strong correlates of stress in Papuan Malay, which is closely related [2]. This study uses spontaneous data from 6 speakers (2 Male, 4 female).¹ Two of the recordings are first person story tellings. The other recordings include sections of storytelling, and discussion between participants. They are between 2 and 17 minutes long.

We find that syllables described as stressed (by [1], a L1 speaker) have longer duration (see Figure 1a,b) and shallower spectral tilt than unstressed syllables (Figure 1c,d). These findings were confirmed by our statistical analyses, where we examined the effects of stress on each measure (with various control predictors and random effects of speaker and word). Likelihood ratio tests (χ^2) were carried out, comparing the model to an intercept-only model. For all measures, we found a significant effect of stress. These findings provide first evidence for the production of word stress in Kupang Malay, confirming previous claims. They further demonstrate that prosody of Malay is diverse, with some varieties demonstrating acoustic evidence for word stress, while others do not.

¹ This is an admittedly low number of speakers. However, we note that the number of speakers in this study is higher than number of levels (five) typically considered to be sufficient for a variable to be treated as a random effect in linear mixed models (see [9]).

Figure 1: Boxplots of duration per segment (ms), duration deviation (ms), H1-A2 (db) and H1-A3 (dB) of syllables labelled stressed and unstressed



References

- [1] J. Jacob, "A sociolinguistic profile of Kupang Malay, a creole spoken in west Timor, eastern Indonesia," Special topic paper towards the requirements for the degree of Masters of Applied Linguistics. Faculty of Science, Information Technology and Education, Northern Territory University, Darwin, Australia, 2001.
- [2] C. Kaland, "Acoustic correlates of word stress in Papuan Malay," *Journal of Phonetics*, vol. 74, pp. 55–74, 2019.
- [3] C. Odé, "On the perception of prominence in Indonesian," in *Experimental studies of Indonesian prosody*, C. Odé, V. van Heuven, and E. van Zanten, Eds. Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië, Rijksuniversiteit Leiden, 1994, pp. 27–127.
- [4] A. Athanasopoulou, I. Vogel, and N. Pincus, "Prosodic prominence in a stressless language: An acoustic investigation of Indonesian," *Journal of Linguistics*, vol. 57, no. 4, pp. 695–735, 2021.
- [5] E. Van Zanten, R. Goedemans, and J. Pacilly, "The status of word stress in Indonesian," in *The phonological spectrum II: Suprasegmental structure*, V. J. v. H. Jeroen van de Weijer and H. van der Hulst, Eds. Amsterdam/Philadelphia: John Benjamins, 2003, pp. 151–175.
- [6] R. Goedemans and E. van Zanten, "No stress typology," in *Above and beyond the segments: Experimental linguistics and phonetics*, J. Caspers, Y. Chen, W. Heeren, J. Pacilly, N. O. Schiller, and E. van Zanten, Eds. John Benjamins, 2014, p. 83–95.
- [7] M. Gordon and T. Roettger, "Acoustic correlates of word stress: A cross-linguistic survey," *Linguistics Vanguard*, vol. 3, no. 1, pp. 1–11, 2017.
- [8] A. M. C. Sluijter and V. J. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, p. 2471–2485, 1996.
- [9] B. M. Bolker, "Linear and generalized linear mixed models," in *Ecological Statistics: Contemporary Theory and Application*, G. A. Fox, S. Negrete-Yankelevich, and V. J. Sosa, Eds. Oxford: Oxford University Press, 2015.

Form- and meaning-related factors in the production and perception of prominence

Janne Lorenzen & Stefan Baumann (*IfL Phonetik, University of Cologne, Germany*)

Keywords: Prosodic prominence, information status, German, production, perception

Prominence can be understood as a perceptual construct, making linguistic entities “stand out” in relation to their environment [1]. In German, an important function of prominence is to signal discourse meaning, e.g., information status (IS) [2]. Formal aspects of prominence include morphosyntactic and phonological structure as well as acoustic properties of speech.

In a *production* study, we investigated the influence of several pragmatic and syntactic factors on categorical and continuous prosodic correlates of prominence. We designed eight short stories (see (1) for an example), each including a target word (a common noun) occurring five times in different manifestations of referential and lexical IS, grammatical role, and position in the sentence (Table 1). Each of the four factors was predicted to either boost or attenuate prosodic prominence. 20 speakers (13f/7m, 18-31 years old) were recorded reading these stories in a listener-directed task.

The results indicate that both lexical and referential *newness* is typically marked with a nuclear accent (Targets 1 and 4). However, 25% of r-new but l-given words (Target 4) received a prenuclear or no accent (Figure 1A), suggesting that lexical IS had a stronger effect on nuclear accent placement than referential IS. Counter to our expectations, the initial position (Targets 3 and 5) attracted considerable phonetic prominence, typically being produced with a rising prenuclear accent (Figure 1B) and higher F0 scaling, longer duration, and higher intensity. This aligns with Bolinger’s [3] observation that utterances in West-Germanic languages are often produced with a prominent accent towards the beginning. Taken together, the four factors exerted an additive effect on prosodic prominence.

To investigate the contribution of continuous and categorical prosodic and non-prosodic cues on the *perception* of prominence, we conducted a prominence rating task, using recordings from the previously reported production task as stimuli. 65 participants (25f/39m/1nb, 18-71 years old) listened to 47 excerpts from the eight stories, consisting of three to five sentences each. They were asked to rate how *highlighted* the underlined target word occurring in the last sentence sounded using visual analogue scales. We included target words representing all five conditions (Table 1) produced with a variety of prosodic realizations.

A random forest analysis revealed that prosodic cues (especially phonological categories) contributed most to predicting prominence ratings (Figure 2A). Nuclear accented target words received higher ratings than prenuclear accented or deaccented words, rising and falling accents received higher ratings than high accents and those in turn received higher ratings than low accents and deaccented words (Figure 2B). In addition, target words produced with longer durations, more periodic energy, and higher values of F0 scaling were perceived as more highlighted. By contrast, non-prosodic cues contributed little to the perception of prominence. Among those cues, position was the most influential, with initial target words being rated as more highlighted than medial target words. The prevalence of prosodic over non-prosodic cues and especially the small effect of IS that we observed may provide evidence for the importance of the prosodic marking of IS. Alternatively, our task instructions may have led listeners to focus more on the prosodic aspects of the signal and less on its meaning [4].

While meaning crucially influences the position of the nuclear accent, which in turn is decisive for the perception of prominence, the meaning categories tested here were almost inconsequential for prominence perception. However, the initial position in a sentence turned out to be a more powerful attractor of prosodic strength, both in production and perception.

- (1) Um seinen Abschluss zu feiern, hat Julian einen Städtetrip unternommen. Er ist mit dem Zug nach Paris gefahren, hat im Hotel eingeecheckt und wollte dann die Stadt erkunden. An dem Tag hat Julian **eine Parade (Target1)** eingeplant. Er wusste, dass ein Feiertag war und ist sofort in die Innenstadt gefahren. Dort hat Julian **die Parade (Target2)** angeschaut. Die Stimmung in der Stadt war super und die Straßen waren voll mit feiernden Menschen. **Die Parade (Target3)** war sehr groß und sah professionell aus. Letztes Jahr hat Julian einen Freund in dessen Heimatstadt zum Karneval besucht. Dort hat sein Freund **eine Parade (Target4)** organisiert. Die war allerdings sehr klein und ziemlich langweilig. Dieses Mal war Julian aber richtig beeindruckt! **Die Parade (Target5)** wird er nie vergessen.

	Target1	Target2	Target3	Target4	Target5
Referential inf. status	New ↑	↓ Given	↓ Given	New ↑	↓ Given
Lexical inf. status	New ↑	↓ Given	↓ Given	↓ Given	↓ Given
Grammatical role	Object ↑	Object ↑	↓ Subject	Object ↑	Object ↑
Sentence position	Medial ↑	Medial ↑	↓ Initial	Medial ↑	↓ Initial

Table 1: Target word conditions and predicted influence on prosodic prominence (red = boosting, blue = attenuating).

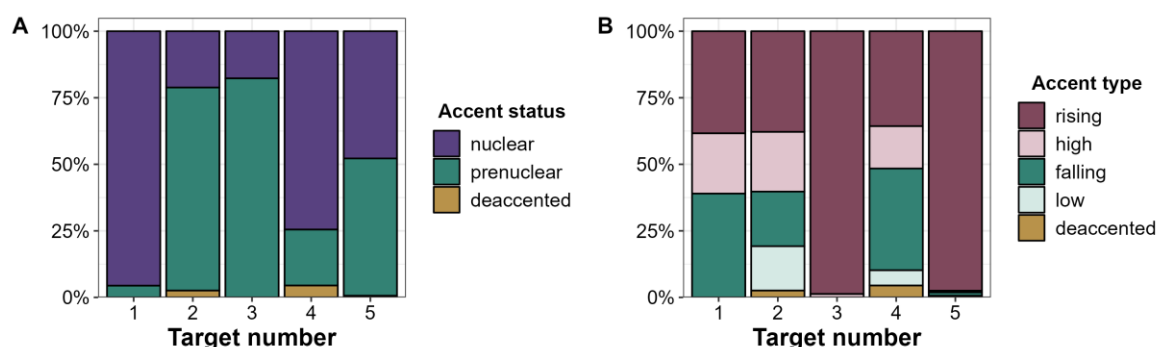


Figure 1: Accent status (A) and accent type (B) distributions by target word position.

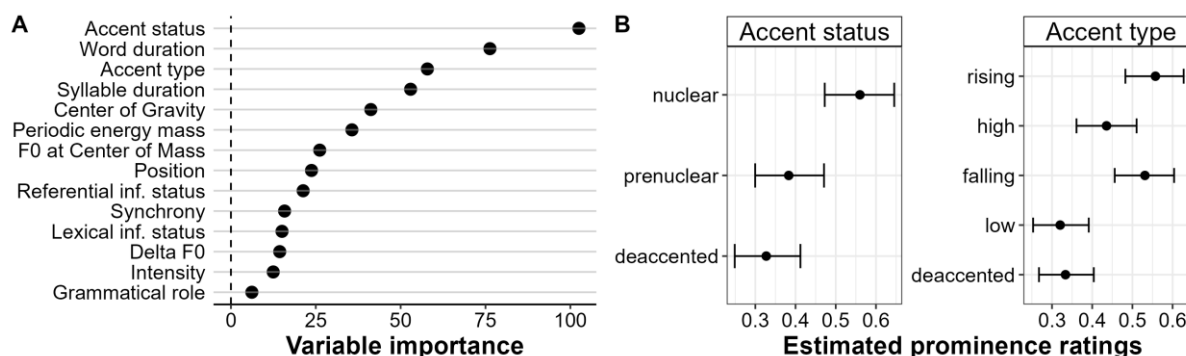


Figure 2: A: Variable importance ranking for prominence perception according to Random Forest Analysis. B: Estimated prominence ratings by accent status and accent type realizations.

References

- [1] Streefkerk, B. M. 2002. *Prominence. Acoustic and lexical/syntactic correlates*. PhD dissertation. University of Amsterdam.
- [2] Baumann, S., & A. Riester. 2013. Coreference, lexical givenness and prosody in German. In Hartmann, J., Radó, J., Winkler, S. (Eds.), *Lingua* 136, 16–37.
- [3] Bolinger, D. 1989. *Intonation and Its Uses*. Palo Alto: Stanford University Press.
- [4] Cole, J., J. Hualde, C. Smith, C. Eager, T. Mahrt & R. Napoleão de Souza. 2019. Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *J. Phon.* 75, 113–147.

Modal particle meets prosody: Speech act modification in declarative questions (?)

Sophie Repp & Heiko Seeliger

University of Cologne

The illocution of an utterance has prosodic reflexes in many languages. This holds for ‘basic’ illocutions, like assertion vs. question, but also for subtler illocutionary differences, like between ‘regular’ vs. surprise questions [1], or rejecting questions [10]. Prosody often ‘collaborates’ with other illocution-signaling means, like syntax or particles. Particles interact with prosody in languages like Cantonese, where sentence-final particles show illocution-related prosodic characteristics [6][7]. In this study, we examine the prosodic reflexes of the presence of a clause-internal modal particle (MP) in German, where MPs modify the illocution in subtle ways. For instance, the MP *wohl* ‘presumably’ signals weakened commitment [13][14]: In assertions, which express that the speaker is committed to the truth of a proposition, *wohl* signals a weaker commitment of the speaker; in questions, *wohl* indicates that the answer is expected to express only weak commitment of the addressee [14]. Crucially, this specific question meaning has been claimed to arise only in questions with an interrogative syntax, not in declarative questions: [13] claims that declarative questions with *wohl* only occur with assertion intonation (i.e., a final fall) and receive a question meaning via purely pragmatic reinterpretation: as a request to a very knowledgeable addressee to confirm the speaker’s conjecture that is expressed by the declarative. [11] show that the latter claim is too strong and they doubt the claim about the assertion intonation.

In this study, we examine what prosodic reflexes the meaning contribution of *wohl* has in assertions vs. questions with a declarative syntax (ASS vs. DQ). We present data from a production experiment with a 2×2 design (24 German speakers; 12 lexicalizations), which tested DQ and ASS with and without *wohl* (DQ_{wohl}/DQ_∅/ASS_{wohl}/ASS_∅). The task involved pseudo-dialogues between three speakers, see (1). Speakers 1/2 were prerecorded; participants took the role of speaker 3. In the target declaratives, punctuation indicated the intended basic illocution (‘.’ for assertion; ‘?’ for question), which was also backed by the context. The left context ended in a tag question (ASS_{∅/wohl}) or in a belief statement (DQ_{∅/wohl}). The right context revealed the following. ASS_∅: the speaker has first-hand knowledge of the truth of the proposition; ASS_{wohl}: the speaker has hearsay knowledge about the proposition, i.e., is only weakly committed; DQ_∅: the speaker requests confirmation of the proposition indicating interest in the answer; DQ_{wohl}: the speaker requests confirmation of a conjecture regarding the proposition, which is based on personal reasoning. Our analysis (Fig. 1) revealed that ASS_∅ / ASS_{wohl} show virtually no variation in boundary tones, irrespective of the presence of *wohl* (L-%); neither were there phonetic differences between ASS_∅ vs. ASS_{wohl}. DQ contained a variety of boundary tones: H-^H%, L-H%, H-%, L-%, H-L%. H-L% was produced by only three speakers (birth dialect areas: Baden (1 speaker), Riparian (2)): it is not found in Standard German. The distribution of the boundary tones differed between DQ_∅ and DQ_{wohl}. In DQ_{wohl} there were fewer high rises (H-^H%) than in DQ_∅, and more L-%. The dialect-specific H-L% was produced by all three speakers in DQ_∅, but by only one in DQ_{wohl}. Finally, DQ_{wohl} showed a lower mean F0 than DQ_∅ in the clause-final region (last 2 unaccented syllables) in final plateaus (H-%) and high rises. We also found differences between ASS and DQ in the prenuclear regions (cp. [9]), which we will address in the presentation.

Our findings indicate that the *wohl*-induced weakening of the commitment is not marked prosodically in ASS. In DQ, the addition of *wohl* has prosodic consequences, which we take to result from the modified illocution. We propose that DQ_{wohl} show fewer indicators of questionhood/interrogativity than DQ_∅, or put differently, DQ_{wohl} indicate that the speaker is less open towards and less interested in the answer, as is reflected in the fewer H-^H% and more L-% boundary tones, which have been argued to signal less open speaker attitude [4][5]. The rarer use of the speaker-specific question-typical H-L% boundary supports this assumption. Finally, interrogativity is typically associated with high pitch [3][8], so a lower mean F0 signals reduced interrogativity. The less questioning character of DQ_{wohl} plausibly results from the speaker requesting confirmation of a proposition towards which they are already biased (their conjecture); DQ_∅, in contrast, check contextual evidence without the speaker necessarily having a previous bias [12]. These results tie in nicely with findings for other ‘non-canonical’ speech acts, which differ from ‘text-book’ versions of assertions and questions: Rhetorical polar questions have been shown to occur more regularly with final plateaus than final rises (vs. information-seeking questions) [2]; rejecting questions show features both of rejections and questions [10].

(1) **Sample item** (Context abridged, and only in English)

Speaker 1: You know Susi, right? The antiquarian on Miller Street? She went on a mad shopping spree last week. She told you all, didn't she?

Speaker 2: Yes, her shop is far too small for all that stuff. She is thinking about renting a garage behind the ring road but she is worried about the transport. Alex, Susi talked with you about this, didn't she?

ASS_{ø/wohl}: She also wants to store her China there, right? (German tag: *oder?* 'or')

DQ_{ø/wohl}: I think, she wants to store her China there.

Speaker 3: Susi will da dann Vasen lagern. ./? ASS_ø / DQ_ø

(Alex) Susi will dann wohl Vasen Lagern. ./? ASS_{wohl} / DQ_{wohl}

Susi wants there / then then / wohl vases store

ASS_ø (ASS_{wohl}): 'Susi wants to store vases there/then, (I think).'

DQ_ø (DQ_{wohl}): 'Susi wants to store vases there/then, (you think)?'

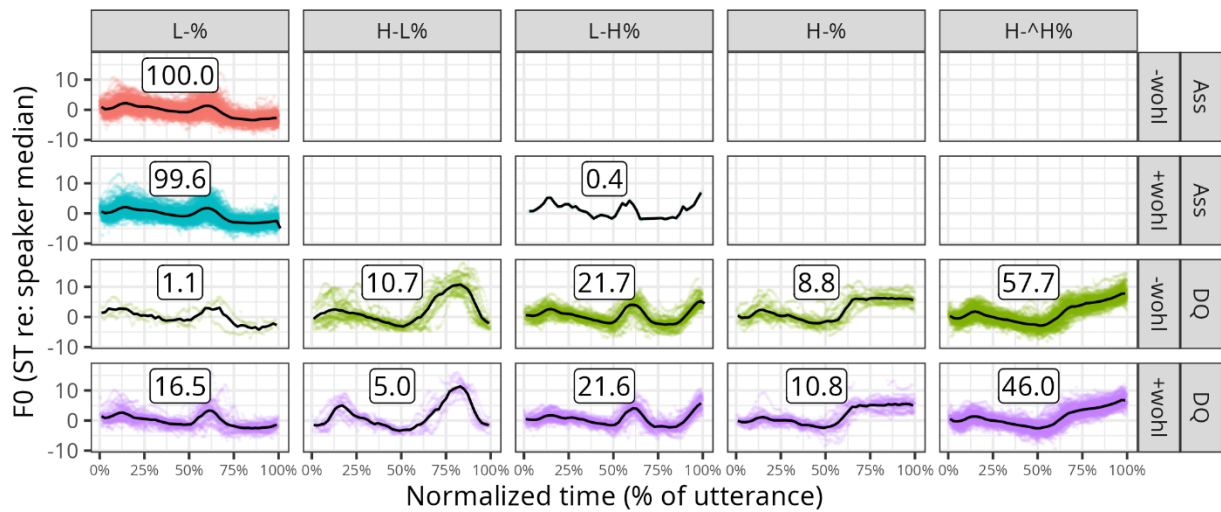


Figure 1: Time-normalized average F0 contours (black, with individual contours in colour) split up by boundary tone (columns) and experimental conditions (rows). The number is the percentage of contours in each condition.

References: [1] Asu, E. L., Sakhai, H., & Lippus, P. (2024). The prosody of surprise questions in Estonian. *JLinguistics*, 60(1), 7-27. [2] Braun, B., Dehé, N., Neitsch, J., Wochner, D., & Zahner, K. (2019). The prosody of rhetorical and information-seeking questions in German. *Lang&Speech*, 62(4), 779-807. [3] Haan, J. (2002). *Speaking of Questions*. LOT. [4] Kohler, K. (2005). Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions. *AIPUK* 35a, 125-142. [5] Kügler, F. (2003). Do we know the answer? Variation in yes-no-question intonation. In Fischer, Van de Vijver & Vogel (eds.) *Experimental Studies in Linguistics I. LiP*. 9-29 [6] Lee, J.H.N. (2021). Low boundary tone: Evidence from the acoustic differences between Cantonese sentence-final particles with low-falling tone. *Tone&Intonation* 1, 176-180. [7] Lee, J.H.N., Yip, K.-F., Liberman, M., Kuang, J. (2024). Some prosodic consequences of varied discourse functions in a Cantonese sentence-final particle. *SpeechPros* 632-636. [8] Michalsky, J. (2017). *Frageintonation im Deutschen. Zur intonatorischen Markierung von Interrogativität und Fragehaltigkeit*. De Gruyter. [9] Petrone, C., & Niebuhr, O. (2014). On the intonation of German intonation questions: The role of the prenuclear region. *Lang&Speech*, 57(1), 108-146. [10] Repp, S., & H. Seeliger (2023) Reject?! On the prosody of non-acceptance. *20th ICPhS*. Skarnitzl & Volín (eds.). 1355-1359. [11] Seeliger, H., & S. Repp. (2018). Biased declarative questions in Swedish and German: The syntax of negation meets modal particles (*väl* and *doch wohl*). In Dimroth & Sudhoff (eds.) *The Grammatical Realization of Polarity Contrast: Theoretical, empirical, and typological approaches*. Benjamins, 129-172. [12] Trinh, T. (2014). 2014. How to ask the obvious – A presuppositional account of evidential bias in English yes/no questions. *MITWPL* 71. [13] Zimmermann, M. (2004) Zum Wohl: Diskurspartikeln als Satztypmodifikatoren. *Ling Berichte*, 253-286. [14] Zimmermann, M. (2011). Discourse particles. In: von Heusinger, Maienborn & Portner (eds.) 2011, *Semantics* (HSK 33.2), de Gruyter, 2011-2038.

Toneme, is it different from underlying tone?¹

In the typological comparison of tonal systems, the primary focus is frequently on the inventory of tones. For instance, in WALS, languages are compared by the number of tones (Maddieson 2013). However, the term “tone” in tonology is polysemantic, leading to variations in the number of tones attributed to a tonal system depending on the interpretation: whether it refers to a surface pitch contrast, a pitch movement on a single TBU, or an underlying unit.

It was proposed elsewhere (Vydrin & Maslinsky 2025) to reintroduce the term “toneme” as the basic unit of tonology. To operationalize the notion of toneme, it is essential to provide a precise definition of its properties and establish a set of criteria. These criteria can be categorized into three subsets: a general criterion (the Persistence Criterion), contour criteria (Non-compositionality Criterion; Extensibility Criterion) and non-zero-ness criteria (Floating Criterion, Shared TBU Criterion, Activity Criterion, Tone Morpheme Criterion). Some criteria were suggested earlier (and coined by us): by Pike (1948) and Yip (2002) (Non-compositionality), by Hyman (2000) (Floating, Shared TBU, Activity), and adapted by us; the others have been proposed by us.

The analysis of tonal systems in terms of tonemes provides numerous advantages: it enhances the comparability of tonal systems based on their inventories and tonal density; it provides a clearer understanding of their organization and functioning; more precise formulation of tonal rules.

The question arises: toneme, is it just another term equivalent to the widely used term of underlying tone? It is true that there is considerable overlap between the two concepts. However, instances can be found where tones which are generated by processes — thus not considered underlying — exhibit characteristics of tonemes.

For example, in Eton (Northwestern Bantu), there are two tonemes, /L/ and /H/. Van de Velde (2008) postulates the existence of a “dissimilating high tone” on certain syllables. These syllables carry a low tone if preceded by a high-tone syllable (1), and they carry a high tone if preceded by a low-tone syllable (2); in phrase-initial position, they carry a high tone.

(1) Ñ-kúl wamà → ñkúl **wámò** ‘my slit drum’

(2) Ñ-pàn wamà → mpàn **wámô** ‘my arrow’

In fact, the “dissimilating high tone syllables” are toneless, they acquire their tone through the polarization rule. However, as illustrated in (1), the low tone generated by polarization becomes compressed due to the spreading of the preceding high tone, resulting in a tonal contour on the originally toneless syllable. According to the Shared TBU Criterion, the L tone must be recognized as a toneme. Besides, in (2), the high tone on *wa* generated by polarization spreads on the subsequent L-toned syllable (*wamò* → *wámô*). According to the Activity Criterion, this H tone must be recognized as a toneme too.

If we accept the possibility of toneme generation through rules, it allows for a clear distinction between tonal spreading and tone copying rules: in the case of tonal spreading, there is an extension of the span of an already existing toneme, whereas tone copying involves the generation of a new toneme.

Mwan (South Mande < Mande < Niger-Congo) demonstrates the existence of both tonal rules, which produce distinct effects.

The H tone copying rule is applied within a tonal phrase, whereby a H toneme is copied onto the subsequent low-toned or toneless syllable or foot, thereby entirely occupying it and deleting the original L toneme.

Conversely, the H tone spreading occurs in one specific context: when a direct object ends with a H tone and the following verb carries a grammatical L tone (i.e., the marker of the affirmative perfective), the H tone spreads to the right, compressing the L toneme without erasing it.

¹ This study is financed by the ERC Advanced grant “Theory of Tone”.

Bibliography

- Hyman, Larry M. 2000. Privative tone in Bantu. In *Symposium on Tone, ILCAA*. Tokyo.
- Maddieson, Ian. 2013. Tone. In Matthew S. Dryer & Martin Haspelmath (eds.), *World Atlas of Language Structures Online*. Leipzig: Max Plank Institute for Evolutionary Anthropology. <http://wals.info/chapter/13>.
- Pike, Kenneth Lee. 1948. *Tone languages: A technique for determining the number and type of pitch contrasts in a language, with studies in tonemic substitution and fusion* (University of Michigan Publications. Linguistics 4). Ann Arbor.
- Velde, Mark van de. 2008. *A grammar of Eton* (Mouton Grammar Library 46). Berlin: Mouton de Gruyter.
- Vydrin, Valentin & Kirill Maslinsky. 2025. Toneme as a basic unit of tonology and criteria for its identification. In *22 Old-World Conference in Phonology (poster)*. Amsterdam. <https://doi.org/10.5281/zenodo.14814108>.
- Yip, Moira. 2002. *Tone* (Cambridge Textbooks in Linguistics). Cambridge University Press.

Key words: toneme, underlying tone, toneme criteria, tonal rules, tone-bearing unit

Lexical pitch accents in Mian, a language of Papua New Guinea

Sebastian FEDDEN, Marija TABAIN

Keywords: lexical pitch accent, Papuan languages

Mian (also known as Mianmin or Miyanmin, ISO *mpt*) is a Papuan language of the Ok family, which is part of the larger Trans New Guinea (TNG) family (Fedden 2011). It is spoken by fewer than two thousand people in the north-west of Telefomin District in Sandaun Province, Papua New Guinea.

Mian has a total of five tonal melodies: L, H, LH, LHL, and HL. Words are lexically specified for one tonal melody, although in general the functional load of tone is low, in particular for verbs. Word stems can be described as either ‘unaccented’ or ‘accented’. Unaccented stems are either all L or all H, whereas accented stems have an accent which indicates where a complex tonal melody, i.e. LH, LHL, or HL, is to be aligned. The accent is the anchor point for the melody (Hyman 1978, Donohue 2003, Gussenhoven 2004). All tonal melodies occur on monosyllabic, disyllabic and polysyllabic words. However, there are differences according to word class, and according to native versus foreign vocabulary. For instance, the tonal melodies L, H, and LH are very common in nouns and adjectives, which tend to be mono- or disyllabic, whereas LHL and HL are quite rare in these words. By contrast, LHL and HL are very common in verbs, which tend to be polysyllabic, but rare in non-verbs.¹

The present study is based on dictionary recordings conducted in September 2023 by the second author. The speaker is Mr Jeremiah Eron. Recordings were made direct to computer, using a Dell Precision M4800 laptop running Windows 10, a Behringer U-Phoria UM2 Audio interface microphone amplifier, Audacity 3.3.3 audio software, an AKG condenser microphone (C5 model), XLR cable, and Audiotechnica headphones. Post-processing was done using Octra for finding word boundaries, MAUS for auto-alignment prior to manual correction, VoiceSauce and Snack software for extraction of f0 and other signal data, and R for data exploration and visualization.

Figure 1 shows GAM-smoothed and time-averaged f0 tracks for 1553 word tokens, plotted according to lexical tone. The top panel shows all data combined; and the bottom panel shows data according to number of syllables in the word (note that token numbers are low for certain cells, such as L tone in 5 syllables). The results confirm the largely auditory description of the system in Fedden (2011). The H tone remains high until the end of the word, and the L tone remains low throughout the word. The HL tone has an early peak and gradual drop-off. The LH tone has a very late and sudden peak, whereas the LHL tone has a slightly earlier peak that is not necessarily as steep. These patterns are remarkably similar across differing word syllables.

Future work will consider the interaction between lexical tone and vowel quality.

¹ There is an unusual interaction between accent location and vowel, in that the pharyngealized /a^ʕ/ vowel attracts the accent in polysyllabic words (Mian has six monophthong vowels, including /a^ʕ/, plus six diphthongs). In addition, Fedden (2011) describes a certain amount of vowel reduction in syllables that do not carry the tonal accent. This interaction between tone and vowel quality is not explored in the present paper, but is mentioned here as an indication of the somewhat unusual tonal system of Mian.

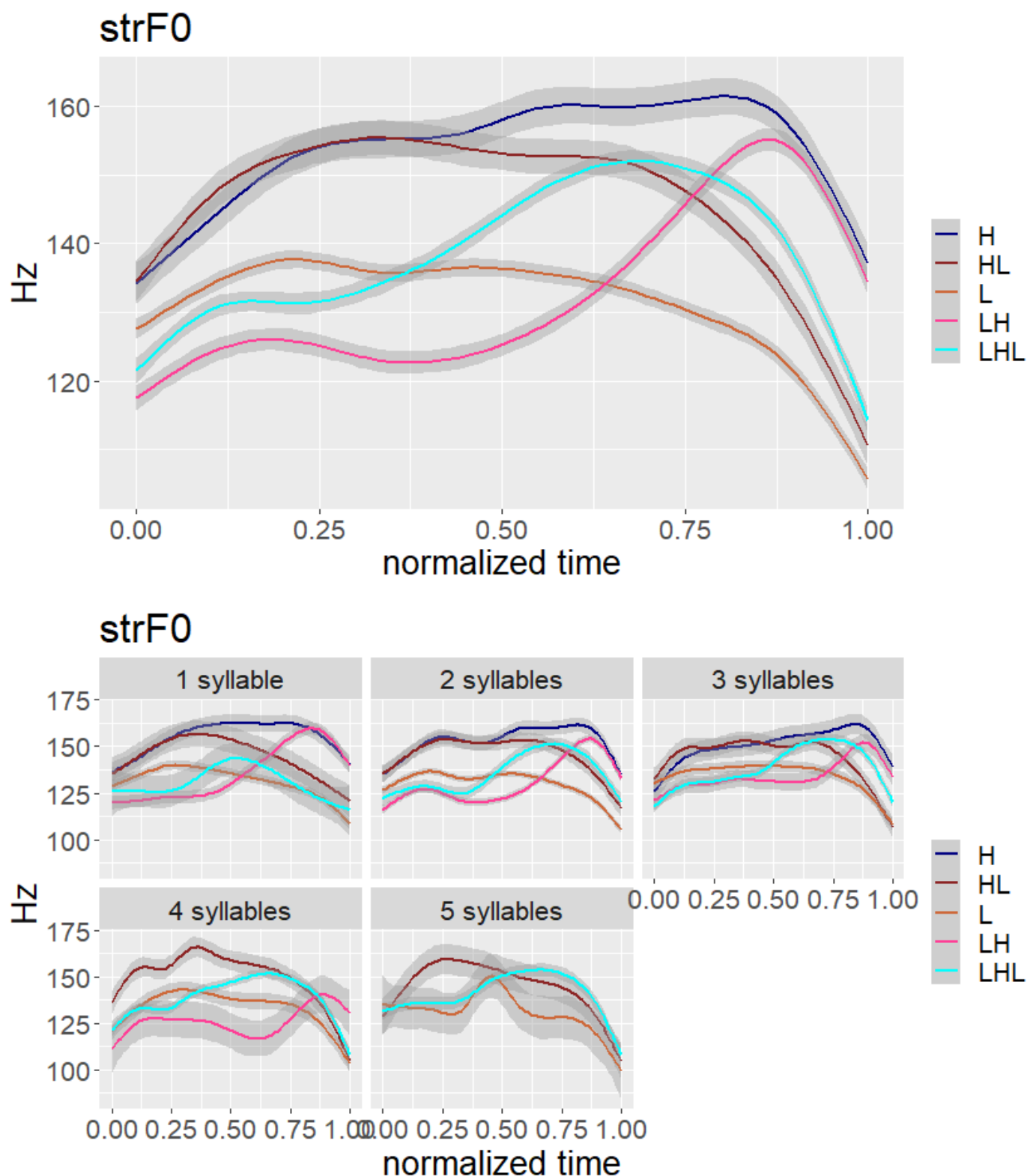


Figure 1 – f0 data for Mian lexical tones, based on 1553 tokens.

References

- Fedden, Sebastian. (2011). *A grammar of Mian*. Mouton Grammar Library 55. Berlin: De Gruyter Mouton.
- Hyman, Larry M. (1978). Tone and/or accent. In *Elements of Tone*, Donna J. Napoli (ed.), 1-20. Washington, DC: Georgetown University Press.
- Donohue, Mark. (1997). Tone systems in New Guinea. *Linguistic Typology* 1: 347-386.
- Gussenhoven, Carlos. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

The role of larynx height and tongue position in the Javanese register contrast

Marc Brunelle, Daniel Schweizer, Suzy Ahn, *University of Ottawa*

In Javanese, the Austronesian voicing contrast in onset stops has been replaced with a *register* contrast (also described as a contrast between *tense-lax*, *stiff-slack*, *light-heavy* or *clear-breathy* stops). Overall, *low register stops* have a longer lag VOT than *high register stops* and are followed by vowels with a breathier voice, a lower F1 and a lower f0 (Adisasmito-Smith 2004; Fagan 1988; Hayward 1995; Kenstowicz 2021; Thurgood 2004). While the acoustics of the Javanese register contrast have been studied extensively, their articulation is not as well understood. There is laryngoscopic evidence that the larynx is lowered during the production of low register stops (Brunelle 2010; Hayward et al. 1994) but available videos are hard to quantify and suffer from a low time resolution. It has also been proposed that low register stops could be produced with an advanced tongue root (Poedjosoedarmo 1986), following claims made regarding the production of the acoustically similar register contrast in languages of Mainland Southeast Asia (Gregerson 1976). Changes in tongue height could also account for formant differences between registers.

We conducted laryngeal and lingual ultrasound studies with five native speakers of Central and Eastern Javanese (4 women and 1 men) to determine the respective roles of larynx height, tongue root advancement and tongue body height in the register contrast. Laryngeal ultrasound videos were collected and quantified for larynx height using optical flow analysis following procedures similar to those described in Moisik et al. (2014) and Witsil (2019). Separate midsagittal lingual ultrasound videos were then recorded with the same participants and fanline coordinates were tracked across images to determine the position of the tongue root and tongue dorsum during target stops and adjacent vowels. The word list was composed of 30 target words containing 10 intervocalic target consonants (high /p, t, k/, low /p, t, k/, prenasalized low /nt, ŋk/ and non-contrastive /m, n, s/) before the three target vowels /i, a, u/.

Laryngeal ultrasound results indicate that the larynx is lower during the production of low register stops than during that of non-contrastive nasals (first row of Figure 1). In contrast, it is higher during high register stops. The maximum larynx height difference between high and low register stops ranges between 1 and 4 mm depending on environments and speakers and is normally achieved during the closure. Lingual ultrasound, on the other hand, yields little evidence for tongue root advancement during low register stops (second row of Figure 1). Tongue body height results show inconsistent patterns in different environments but differences between high and low register stops are at best limited (third row of Figure 1).

Overall, our results confirm that larynx height is an important articulatory strategy for realizing the Javanese register contrast. A low larynx could account for the relatively lower F1 found in vowels after lax stops (as suggested by Fagan 1988; Thurgood 2004) as well as for their breathier phonation and lower pitch (Brunelle 2010). There is on the other hand little evidence that the tongue position contributes significantly to formant differences between vowels following high and low register stops. While the diachronic source of Javanese register (< onset voicing) and its reliance on f0 (among other cues) are reminiscent of tone, the fact that the point of maximum contrast in larynx height between the two series is located during the closure suggests that the register contrast is still a consonantal property and has not become fully suprasegmental. This interpretation seems confirmed by the phonological behavior of register in *Walikan*, a Javanese word game in which phonemes are linearly reversed and in which register systematically follows onset consonants rather than vowels (Yannuar et al. 2022).

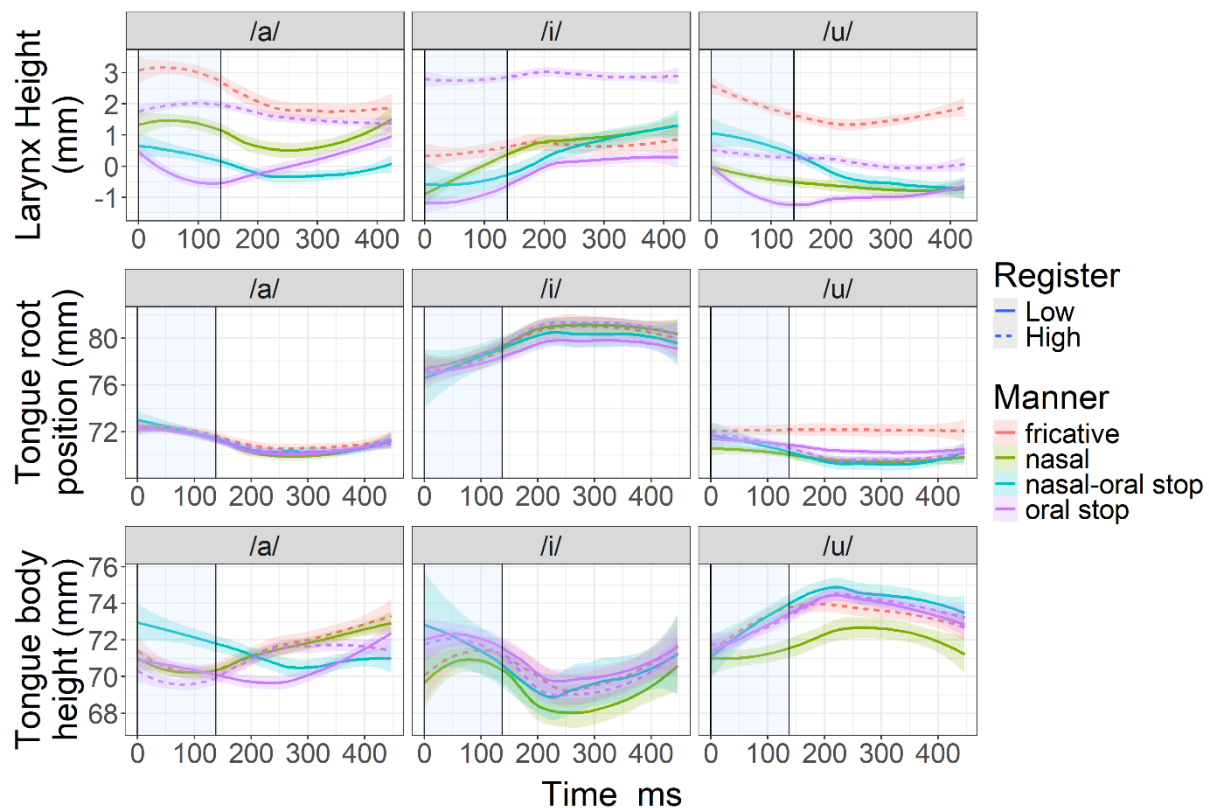


Figure 1: Normalized larynx height, horizontal tongue root position and tongue body height during the production of high and low register stops, by vowel and manner, averaged over the five speakers. Ribbons represent 95% confidence intervals. The blue shading corresponds to the onset, the white portion to the following vowel.

References

- Adisasmito-Smith, Niken. 2004. *Phonetic Influences of Javanese on Indonesian*. Ithaca: Cornell University Ph.D thesis.
- Brunelle, Marc. 2010. The role of larynx height in the Javanese tense ~ lax stop contrast. *Austronesian Contributions to Linguistic Theory: Selected Proceedings of AFLA*, ed. by R. Mercado, E. Potsdam & L. Travis, 7-24. Amsterdam/Philadelphia: John Benjamins.
- Fagan, Joel. 1988. Javanese Intervocalic Stop Phonemes. *Studies in Austronesian Ling.* 76.173-202.
- Gregerson, Kenneth. 1976. Tongue-root and Register in Mon-Khmer. *Austroasiatic Studies*, ed. by P.N. Jenner, L. Thompson & S. Starosta, 323-69. Honolulu: University Press of Hawaii.
- Hayward, Katrina. 1995. /p/ vs. /b/ in Javanese: the Role of the Vocal Folds. *Working Papers in Linguistics and Phonetics* 5.1-11.
- Hayward, Katrina, D. Grafield-Davies, B.J. Howard, J. Latif & Ray Allen. 1994. *Javanese Stop Consonants: The Role of the Vocal Folds*. London: School of Oriental and African Studies.
- Kenstowicz, Michael J. 2021. *Phonetic Correlates of the Javanese Voicing Contrast in Stop Consonants*. NUSA: Linguistic studies of languages in and around Indonesia.1-37.
- Moisik, Scott, Hua Lin & John H. Esling. 2014. A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound(SLLUS). *Journal of the International Phonetic Association* 44.21-58.
- Poedjosoedarmo, Gloria. 1986. The symbolic significance of pharyngeal configuration in Javanese speech: some preliminary notes. *Nusa* 25.31-37.
- Thurgood, Ela. 2004. *Phonation Types in Javanese*. *Oceanic Linguistics* 43.277-95.
- Witsil, AJ. 2019. *Imagefx: extract features from images*. R package version 0.3. 0.
- Yannuar, Nurenzia, Tom Hoogervorst & Marian Klammer. 2022. Examining Javanese phonology through word-reversal practices. *Oceanic Linguistics* 61.560-88.

Speech rate and prosodic phrasing effects in temporal processing: testing two Korean contrasts

Jeremy Steffman

Sahyang Kim

Taehong Cho

Sun-Ah Jun

The University of Edinburgh

Hongik University

Hanyang University

UCLA

Introduction: Rate-dependent speech perception is the phenomenon in which segment-internal temporal cues are perceived relative to temporal context. A fast preceding context makes subsequent cues (e.g., VOT) sound longer, while a slow context makes them sound shorter. These effects are typically understood via general auditory contrast mechanisms [1,2]. However, recent work suggests that prosodic phrasing, often cued by local durational changes (e.g., phrase-initial strengthening [2,3]), also generates similar perceptual effects. In this study, we test how speech rate context interacts with prosodic phrasing in Korean. Specifically, we examine two phonological contrasts: (1) between aspirated and fortis stops /k^h/-k*/ (/k^h/ with longer VOT and a shorter following vowel than /k*/) and (2) between plain /s/ and fortis /s*/ (/s/ with some aspiration following a frication noise and a shorter vowel, versus /s*/ with longer frication duration, no aspiration, and a longer vowel). Korean prosodic phrasing induces “domain-initial strengthening” [4,5], where the vowel in a CV sequence is lengthened phrase-initially. [3,6] showed these patterns have perceptual consequences. A prosodic contextual cue signaling an Intonational Phrase (IP) boundary (and domain-initial strengthening), caused the vowel in the upcoming CV sequence to be perceived as shorter than when with no IP boundary cue. That is, due to expected phrase-initial vowel lengthening after both /k*/ and /k^h/, a relatively longer vowel is needed to cue fortis /k*/ after an IP boundary than without one. This effect was observed even when the boundary was signaled only by F0, without preceding boundary-related lengthening [3], and when speech rate-related slowing down occurred near the boundary, typically expected to elicit a durational auditory contrast effect [6]. Our goal is to examine how these effects operate across different consonant manners, expanding to the similar fricative contrast /s/-/s*/ alongside the stop contrast /k^h/-k*/. To test the generalizability of these effects, we consider if and how the data support the hypothesis that prosodic phrasing modulates temporal processing, sometimes even defying durational auditory-temporal contrast effects.

Method: 56 speakers of Seoul Korean categorized two continua: (1) a 7-step VOT continuum categorized as /k*/ or /k^h/, in a carrier phrase (Fig. 1); and (2) a 7-step Fricative spectrum continuum categorized as /s*/ or /s/ in the same phrase. In Exp.1, there were four contextual conditions which crossed speech **rate** (normal vs. slow) and prosodic **phrasing** (no boundary/IP boundary). The rate manipulation was a linear expansion of the whole pre-target material (both distal and local contexts, shown in Fig. 1). The phrasing manipulations involved a localized lengthening of only the pre-target syllable, resulting in an intonational phrase (IP) boundary (see Figure 1). Each participant categorized both types of segmental contrasts (in separate blocks) in the 2 x 2 rate and phrasing conditions for a total of 140 responses per contrast. Responses were analyzed with Bayesian mixed-effects logistic regression (maximal by-participant random slopes). We report estimates and “probability of direction” (pd): when pd > 97.5, this indicates a credible effect. **Predictions:** *Canonical rate-based effects* are predicted. For stops, this means VOT is perceived as shorter when preceded by slow rate: increasing /k*/ responses. The prediction is different for fricatives, because frication is longer for /s*/ than /s/, a preceding speech rate slowdown may *decrease* /s*/ responses if frication sounds shorter to listeners (more like /s/). Prosodic phrasing effects, if they modulate durational perception in the vowel in CV [3,6], should be analogous in direction across both stop and fricative contrasts. That is, given that the vowel is longer after the fortis (/k*, s*/) than after the non-fortis (/k^h, s/), the perceived shorter vowel duration predicted after the IP boundary would result in decreased fortis responses. Notably, for the fricative contrast, this phrasal effect, if driven by preboundary lengthening, would not be distinct from a local durational auditory contrast effect (i.e., fewer /s*/ responses due to phrasing vs. simple lengthening). Thus, in Exp. 2, with 32 different participants, we examined the effect of a clearer prosodic boundary strengthened by a 200 ms pause preceding the fricative target, to further understand the nature of the phrasing versus auditory contrast effect.

Results. Predicted rate-based effects were observed in the *no boundary* condition for stops ($\beta = -1.67$, pd = 100), but not in the IP condition (pd = 51); also reflected in a two way interaction between rate and phrasing for stops (pd = 100; Fig. 2 Panel A). Replicating [6], we found phrasing-based effects for stops, only when speech rate is slow ($\beta = 1.19$, pd = 100), but not normal (Fig 2. Panel A). This is, critically, the opposite of a predicted contrast effect: the local rate slowdown (cueing an IP boundary) generates a shift which is contrary to the speech rate effect. For fricatives there was no evidence for a two way interaction (pd = 68), but rather an expected main effect of rate ($\beta = 0.65$, pd = 99, Fig. 2 Panel B) which was uniform across phrasing conditions. For fricatives, the phrasing effect was a main effect, which did not interact with rate ($\beta = 0.27$, pd = 99; Fig 2 Panel B). As the fricative effect could also be consistent with a rate-based contrast effect, we turn to the effect of phrasing and pause in Exp.2. Here credible ordinal differences were found between no boundary < IP < IP +pause conditions (pds > 99; Fig 2. Panel C). The unambiguous boundary marking via addition of a pause therefore suggests that listeners are sensitive to phrasing-based differences in this context. Notably, if phrasing played no role, the pause would create temporal separation between the target and its precursor, reducing the auditory contrast effect of pre-target lengthening, thus predicting the IP+pause condition to fall between the IP and no-boundary conditions, contrary to the findings.

In summary, this study contributes to our understanding of perceptual rate and phrasing effects, showing rate effects are clearly not always an “obligatory” [7] default. The presence of phrasing effects in the slow rate only for stops suggests that parsing prosodic structure may be easier at slower overall rates. On the other hand, the fricative effects were uniform across rates, indicating that interactions between phrasing and rate are contrast-specific. Unambiguous phrasing effects were found for the fricative contrast when pre-boundary lengthening was strengthened by a pause. However, without the pause added in Exp 2, we cannot conclusively separate phrasing from rate effects for the fricative contrast: a methodological point about how evidence for these interactions depends critically on stimulus choices. We will further discuss how these findings can be situated into the bigger picture of prosodic structure in speech processing.

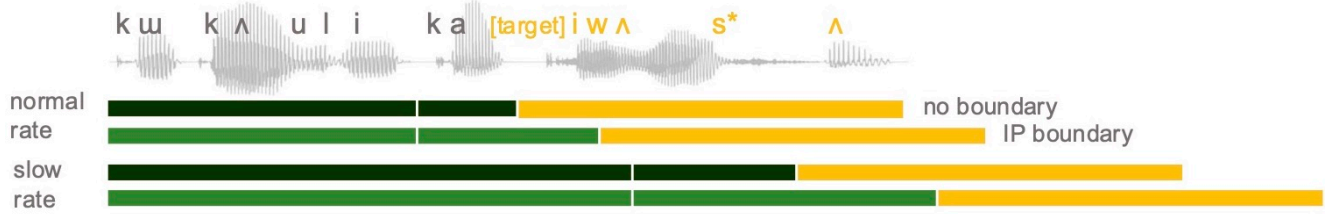


Figure 1: Schematic speech rate conditions and prosodic boundary conditions in the stimuli. The example shows the velar stop target ($/k^hiwʌs^*ʌ/$ vs. $/k^*iwʌs^*ʌ/$). The carrier phrase was the same for the other contrast.

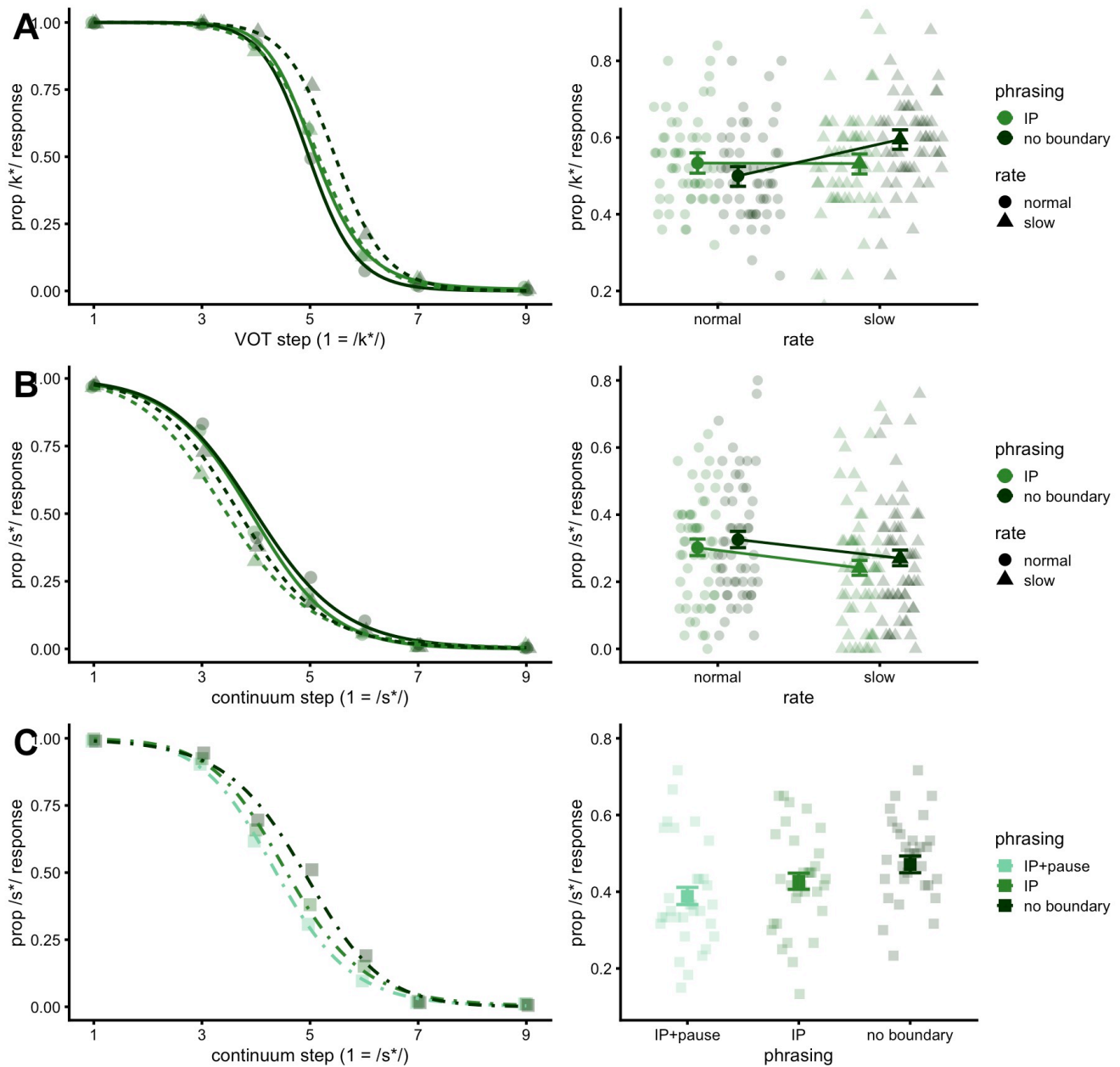


Figure 2: Categorization for the stop contrast (panel A) and fricative contrast (panel B). In each panel the categorization across the whole continuum is shown at left. At right, categorization is pooled across the continuum to highlight the context effects (error bars showing 95% bootstrapped CI) and points showing individual-participant responses. Panel C: The second experiment: the fricative contrast with an added pause.

Refs [1] Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics*, 58(4), 540-560. [2] Mitterer, H., Cho, T., & Kim, S. (2016). How does prosody influence speech categorization?. *JPhon*, 54, 68-79. [3] Steffman, J., Kim, S., Cho, T., & Jun, S.-A. (2022). Prosodic phrasing mediates listeners' perception of temporal cues: Evidence from the Korean Accentual Phrase. *JPhon*, 94, 101156. [4] Cho, T., & Jun, S. A. (2000). Domain-initial strengthening as enhancement of laryngeal features: Aerodynamic evidence from Korean. *UCLA WPP*, 57-70. [5] Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *JPhon*, 29(2), 155-190. [6] Steffman, J., Kim, S., Cho, T., & Jun, S. A. (2024). Speech rate and prosodic phrasing interact in Korean listeners' perception of temporal cues. In *Proc. SpeechProsody 2024* (pp. 1090-1094). [7] Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *J. of Experimental Psychology: Human Perception and Performance*, 14(3), 369.

An fNIRS investigation of Mandarin third-tone sandhi word production

Xiaocong Chen¹, Tai Yuan^{1,2}, Yiya Chen³, Fumo Huang¹, Caicai Zhang¹

¹*The Hong Kong Polytechnic University*, ²*The Stony Brook University*, ³*The Leiden University*

In Mandarin, Tone 3 (T3), a low-dipping tone in citation form, changes to a Tone 2 (T2)-like rising tone when followed by another T3, a well-known example of phonological alternations called the T3 sandhi [1][2]. Previous behavioral studies indicated that the underlying abstract tonal category and the context-specific rising tonal variant are both activated during the production of disyllabic words carrying the T3 sandhi [3][4]. Moreover, a recent EEG study [5] suggested that the T3 sandhi may undergo two stages of planning, an early retrieval of the abstract tonal category and the later specification and motor preparation of the context-specific tonal variant. However, the brain regions underlying these subprocesses remain largely unknown. Although the few existing neuroimaging studies [6][7] found that producing tonal sequences involving T3 sandhi induced stronger activation in right inferior frontal gyrus (IFG) than non-sandhi sequences, these studies did not tap into the production of real words carrying T3 sandhi, and the exact neural regions underlying the subprocesses of the T3 sandhi production remains unclear, warranting further research into the functional neuroanatomy underlying the third-tone sandhi. Thus, this study employed the functional near-infrared imaging (fNIRS) technique to investigate the brain regions for different subprocesses of T3 sandhi word production with a phonologically-primed picture naming task by manipulating different tonal relationships between primes and target words carrying T3 sandhi.

Forty-eight Mandarin native speakers (24 females) from Northern China participated in the experiment. Following [5], participants produced 36 disyllabic target words carrying T3 sandhi (e.g., 雨伞, *yu3san3*, 'umbrella') upon seeing a picture, preceded by a monosyllabic visual plus auditory prime. The primes shared the same syllable with the first morpheme of the target words while the tonal overlap between the prime and the target was manipulated. There were three types of primes: a T3 prime (sharing the underlying tonal category, e.g., 语 *yu3*), a T2 prime (sharing the surface tonal variant, e.g., 娱 *yu2*), and a control prime (with T1 or T4, e.g., 玉 *yu4*). Participants' naming latencies and fNIRS signals were recorded concurrently. Behavioral results showed that both T3 and T2 primes shortened naming latencies of the target T3 sandhi words compared to control primes (Fig. 1A), replicating previous findings [4][5]. The event-related fNIRS analyses of the oxygenated hemoglobin (oxyHb) signals showed that compared to control primes, T3 primes elicited reduced activation of the left middle temporal gyrus (MTG)/superior temporal gyrus (STG), the left and the right IFG and the right postcentral gyrus (postCG) (Fig. 1B). The decreased activation of left MTG/STG may indicate that T3 primes facilitated the retrieval of the abstract tonal category and/or lexical wordforms, whereas the less activation of left and right IFG and right postCG may indicate the reduction of articulatory efforts induced by T3 primes. In contrast, compared to control primes, T2 primes elicited reduced activation of the left IFG and the right middle frontal gyrus (MFG), but stronger activation of the left MTG/STG (Fig. 1C). Similarly, the decreased activation of left IFG and right MFG may indicate the facilitation of articulatory planning induced by T2 primes, whereas the increased activation of left MTG/STG might indicate the increased competition between T2 and T3 induced by T2 primes. In general, our study showed that T3 and T2 primes induced different neural activations within the bilateral frontotemporal brain network, possibly indicating different neural regions involved in the processing of the underlying category and the context-specific variant during the T3 sandhi word production.

Keywords: tone sandhi, fNIRS, word production, phonological encoding, Mandarin

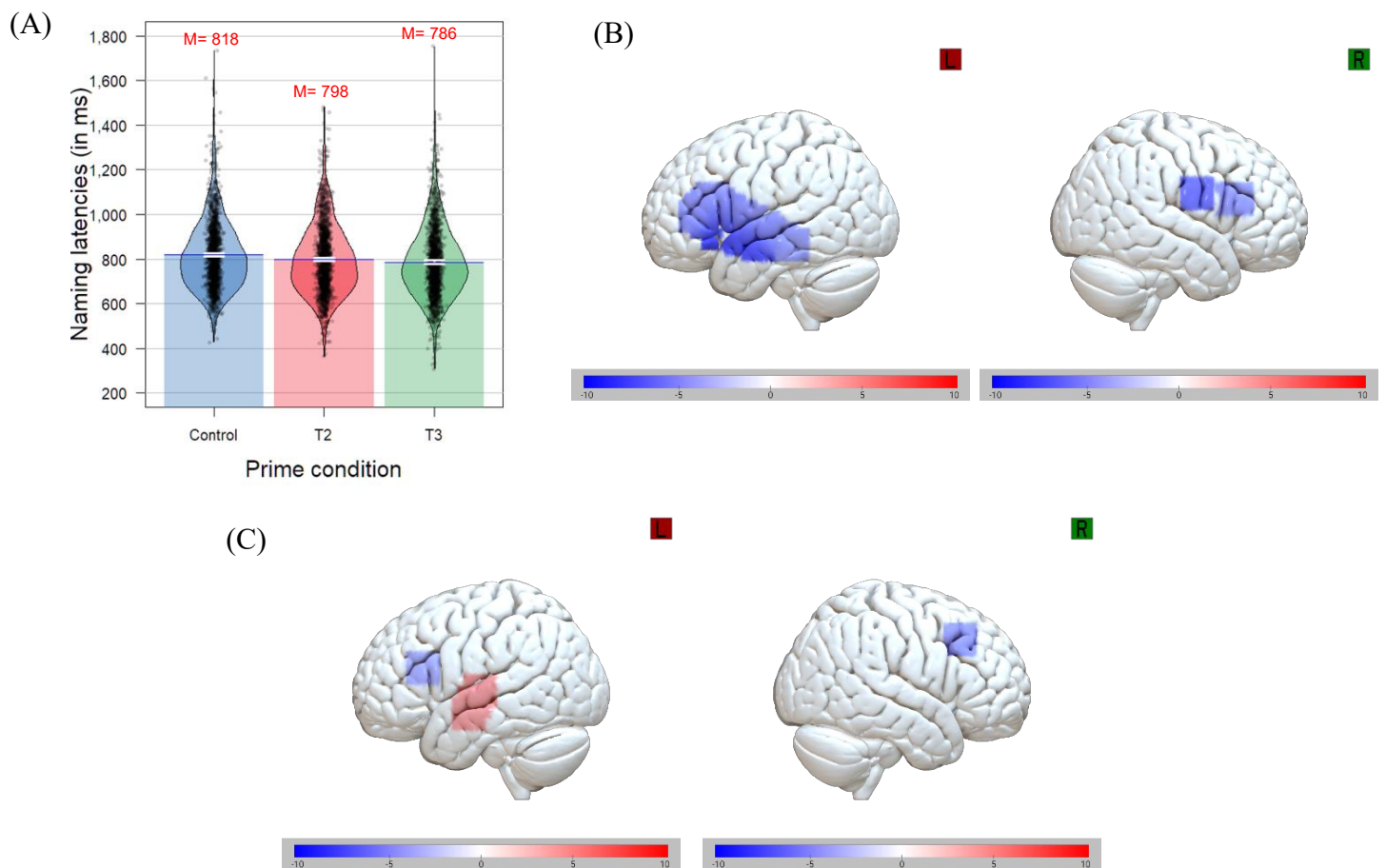


Figure 1. (A) The pirate plot of the mean naming latencies from three prime conditions; (B) oxyHb results (t-maps) for T3 priming effects shown in left and right sagittal views, with blue indicating significantly reduced activation of T3 prime condition than control prime condition (FDR-corrected $ps < .05$); (C) oxyHb results (t-maps) for T2 priming effects shown in left and right sagittal views, with blue indicating significantly reduced activation of T2 prime condition than control prime condition, and red indicating significantly larger activation of T2 prime condition than control prime (FDR-corrected $ps < .05$).

References

- [1] Chen, M. Y. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge University Press.
- [2] Zhang, J., & Lai, Y. 2010. Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology* 27(1), 153–201.
- [3] Chen, Y., Shen, R., & Schiller, N. O. 2011. Representation of allophonic tone sandhi variants. *Proceedings of Psycholinguistics Representation of Tone. Satellite Workshop to ICPhS*, 38–41.
- [4] Nixon, J. S., Chen, Y., & Schiller, N. O. 2015. Multi-level processing of phonetic variants in speech production and visual word processing: evidence from Mandarin lexical tones. *Language, Cognition and Neuroscience* 30(5), 491–505.
- [5] Chen, X., Zhang, C., Chen, Y., Politzer-Ahles, S., Zeng, Y., & Zhang, J. (2022). Encoding category-level and context-specific phonological information at different stages: An EEG study of Mandarin third-tone sandhi word production. *Neuropsychologia*, 175, 108367.
- [6] Chang, C. H. C., Lee, H.-J., Tzeng, O. J. L., & Kuo, W.-J. (2014). Implicit target substitution and sequencing for lexical tone production in Chinese: An fMRI study. *PLoS ONE*, 9(1), e83126.
- [7] Chang, C. H. C., & Kuo, W.-J. (2016). The neural substrates underlying the implementation of phonological rule in lexical tone production: An fMRI study of the tone 3 sandhi phenomenon in Mandarin Chinese. *PLoS ONE*, 11(7), e0159835.

Tenseness in Jiuhe Bai

Meihao Wan, Peggy Mok

The Chinese University of Hong Kong

wanmeihao@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

Introduction: Bai is a Sino-Tibetan language primarily spoken in the Yunnan Province, China. Traditionally, it is considered to have a tense vs. lax register contrast in its tone system [1]: T1, T3, T5 are tense tones; T2, T4, T6 are lax tones. Historically, the tenseness of Bai is derived from the voiced initials [2]. The most well-studied phonetic features distinguishing this register contrast in Bai are pitch and phonation type: tense tones are accompanied by nonmodal phonation which varies across speakers [3, 4]. However, previous studies have not observed differences in vowel quality which also plays an important role in the register contrast of other Sino-Tibetan languages, such as Hani [5] and Yi [6]. This study aims to investigate the phonetic cues of the register contrast in the production of Bai comprehensively.

Method: All data were collected in Jiuhe Township which belongs to Lijiang City, Yunnan Province. 10 native Bai speakers participated in the experiments, all of whom were born and raised in Jiuhe. The stimuli consisted of 14 monosyllabic minimal pairs or near-minimal pairs of the 6 lexical tones. Participants were asked to produce them three times in a random order. Therefore, there were 252 tokens (14 syllables \times 6 tones \times 3 repetitions) per participant. Acoustic analysis was conducted with VoiceSauce [7], extracting the relevant parameters for pitch, phonation and vowel quality (H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, CPP, HNR, Energy, SHR, f0, F1, F2).

Predictions: Jiuhe Bai is a variety of Bai, thus we hypothesized that the register contrast in Jiuhe Bai also involves distinctions in both phonation and pitch, similar to other Bai languages. Additionally, we hypothesized that there are vowel quality differences in the register contrast, with tense vowels having higher F1 values than those of lax vowels.

Results: Figure 1 shows the z-scored pitch contours of the tones in Jiuhe Bai using Generalized Additive Mixed Models (GAMM). T1 and T2 share similar pitch contours, while T3 has a sharper slope compared to both T1 and T2. T4 and T5 have the same ending, but the initial pitch of T5 is higher than that of T4. T6 has the highest pitch among all the lexical tones. Figure 2 shows the formants of the tense and lax vowels at vowel midpoint. The vowels with tense tones have higher F1 values than those with lax tones. A principal component analysis (PCA) was conducted to reduce the dimensionality of the spectral parameters of phonation. The first two PCs account for 59.1% of the variance in the data (PC1 = 30.7%, PC2 = 28.4%, Figure 3). The factor correlation loading for PC1 and PC2 are displayed in Figure 4. PC1 is mostly correlated with individual harmonics, such as H4*(0.40), A1*(0.38), H2*(0.38), H1*(0.36), PC2 is mostly correlated with periodicity and noise, such as HNR15(-0.36), HNR25(-0.36), HNR35(-0.32), and also related to spectral tilt measures, like H1-A3* (0.34), H1-A1*(0.32). A mixed-effect logistic regression model was employed to evaluate the contribution of different cues to the register contrast. Table 1 shows that the register contrast is significantly related to all three dimensions: F0, F1 and PC1, which pertains to phonation.

Discussion: The Jiuhe Bai register contrast exhibits differences in all three dimensions: pitch, vowel quality and phonation. Tense vowels have higher F1 values than those of lax vowels, suggesting a retraction of the tongue root in the language. This study provides insights into the interaction of multiple dimensions in the register contrast.

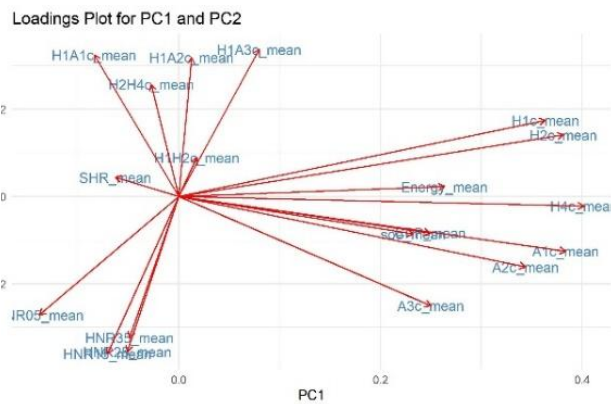
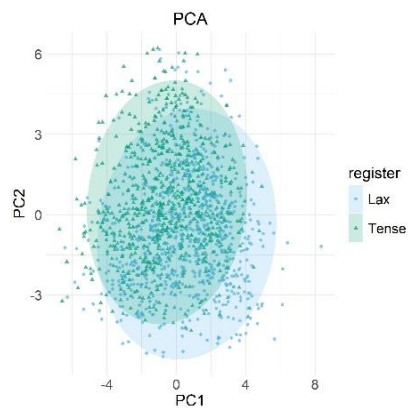
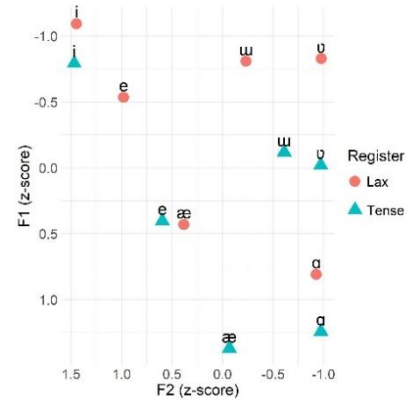
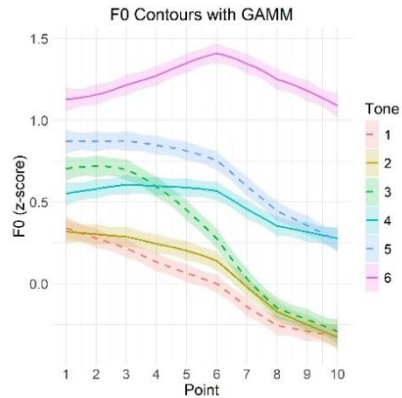
Figure 3: *PCA of the acoustic space*

Figure 4: *The loadings for PC1 and PC 2*

Table 1: *Results of mixed-effects logistic model*

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.02509	0.05592	-0.449	0.6536
PC1	-0.30248	0.07150	-4.231	2.33e-05***
PC2	0.01988	0.11101	0.179	0.8579
sF1_mean	0.50164	0.11023	4.551	5.34e-06***
sF2_mean	-0.09893	0.05902	-1.676	0.0937
strF0_mean	-0.46155	0.07881	-5.857	4.72e-09***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Keywords: tone, register, phonation, production, Bai

References: [1] Xu, L., & Zhao, Y. (1984). A introduction of Bai language [白语简志]. Beijing: The Ethnic Publishing House. [2] Wang, Feng. 2012. Language Contact and Language Comparison: The Case of Bai. Beijing: Commercial Press. [3] Wang, F. (2015). Variations of laryngeal features in Jianchuan Bai. *Journal of Chinese Linguistics*, 43(1), 434-452. [4] Li, X., & Wang, F. (2020). Phonation variation and strategy of tone distinction: the case of Meiba Bai. *Journal of Chinese Linguistics*, 48(2), 379–401. [5] Maddieson, I., & Ladefoged, P. (1985). “Tense” and “lax” in four minority languages of China. *Journal of phonetics*, 13(4), 433-454. [6] Kuang, J., & Cui, A. (2018). Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *Journal of Phonetics*, 71, 194-214. [7] Shue, Y.-L., P. Keating, C. Vicenik, K. Yu (2011) VoiceSauce: A program for voice analysis, Proceedings of the ICPhS XVII, 1846-1849.

Native listeners tolerate tonal pronunciation errors (bad maps) in Cantonese-accented Mandarin

Zhisheng CUI¹ & Eric Pelzl¹

¹*The Hong Kong Polytechnic University*

eddie.cui@connect.polyu.hk, eric.pelzl@polyu.edu.hk

Second language (L2) speakers often substitute unfamiliar sounds in their L2 with similar sounds from their native language (L1) inventory, a phenomenon known as a "bad map" [1] (categorical pronunciation error). Bad maps can impede word recognition for L1 listeners, though listeners can adapt with exposure (e.g., [2]).

We investigated the consequences of bad maps in the context of tonal languages. Although both Mandarin and Cantonese are Chinese languages, they are mutually unintelligible, and have different tonal inventories (four versus six lexical tones). Cantonese speakers learning Mandarin often struggle with specific tone contrasts, particularly Mandarin T1-T4 and T2-T3, due to L1 interference or other perceptual challenges (e.g., [3]; [4]). We tested whether native Mandarin listeners with extensive exposure to Hong Kong Cantonese-accented Mandarin (HK Mandarin) are more likely to accept tonal bad maps that are typical of HK Mandarin (e.g., the mispronunciation of Mandarin T1 as T4) compared to listeners with minimal exposure.

Participants ($n=47$) were all from northern China (16 males, 31 females; mean age = 20.34 years). They were divided into two groups based on their length of residence (LOR) in HK: those with 6 months or less of residence (minimal experience; $n = 23$; mean LOR = 0.19 years) and those with 1 year or more (extensive experience; $n = 24$; mean LOR = 1.91 years).

We conducted a lexical decision task using Mandarin stimuli produced by two speakers with strong HK Cantonese accents (established by a separate rating task). Experimental stimuli comprised 256 items, split between nonwords (32 typical nonwords, 32 atypical nonwords, 64 nonword fillers) and 128 real words. Typical nonwords contained typical HK Mandarin bad maps (e.g., T4-as-T1), while atypical nonwords contained atypical bad maps (e.g., T4-as-T2). Examples of all bad map types are shown in Table 1.

Descriptive results indicating how often participants accepted the stimuli as real words are shown in Figure 1. Data were fit using logistic mixed-effects regression. The best fitting maximal model [5] included fixed effects of Experience (minimal vs extensive) and the interaction of Word Status (nonword vs word) and Condition (typical vs atypical), with by-participant and by-item random intercepts and random slopes for Word Status. Model results (Table 2) showed higher acceptance rates for words versus nonwords, with both groups more likely to accept typical nonwords than atypical nonwords. Post-hoc comparisons indicated that an interaction between Word Status and Condition was driven by differences between typical and atypical nonwords ($z = 5.30$, $p < .001$). There was no difference between participants based on Experience.

Exploration of specific bad map types (Figure 2) showed that typical T2-as-T3/T3-as-T2 nonwords were consistently accepted at a higher rate than the atypical nonwords, while T1-as-T4/T4-as-T1 nonwords were not.

In summary, we tested whether Mandarin listeners would be more likely to accept nonwords in HK Mandarin speech if those nonwords contained tonal bad maps that are typical in HK Mandarin pronunciation. Superficially, our results appear to support this hypothesis, showing that Mandarin listeners were more likely to accept nonwords with typical bad maps more readily than nonword with atypical bad maps. However, exploratory analysis suggests this difference was driven largely by acceptance of T2-as-T3 and T3-as-T2 bad maps. This effect may be due to the typicality of this mispronunciation in HK Mandarin, but it could also be due to the mutual confusability of these tones. For T1-as-T4/T4-as-T1 bad maps, our data do not seem to support a typicality effect, and instead suggest participants may have been prone to accept all types of bad maps from the strongly accented speakers.

The high acceptance rate for atypical tonal bad maps raises questions about the role of typicality in processing tonal bad maps. Samuel and Larraza [2] found that acceptance rates for atypical segmental bad maps were low. Perhaps one explanation is that tonal bad maps are easier to resolve within the limited set of Mandarin tones, regardless of typicality, while segmental bad maps involve numerous possible substitutions, making them more challenging to process. Alternatively, it is possible that the "atypical" tonal bad maps we used in the present study do occur in natural HK Mandarin speech, despite not being documented. There is a need for research that systematically investigate the frequency distribution of tonal

bad maps in HK Mandarin.

Finally, our study found no effect of previous experience on lexical decision performance. This might indicate weaknesses in our simple use of LOR as a proxy for linguistic experience, leading to either an underestimate or overestimate of how much our participants have adapted to HK Mandarin. Alternatively, it may be that participants have not adapted to specific tonal bad map patterns, but instead have learned to ignore all types of tonal bad maps in HK Mandarin. An obvious addition for future research is the recruitment of L1 Mandarin participants outside of HK to make a stronger comparison based on experience with HK Mandarin.

Key Words: lexical tones, accented speech, tonal bad maps, Mandarin, Cantonese

Table 1: Example Stimuli. *Note:* words and nonwords were counterbalanced across experimental lists.

Condition	Word	Nonword	Bad Map
Typical	gui1ze2 规则 <i>rule</i>	gui4*ze2	T1-as-T4
Typical	xue2xiao4 学校 <i>school</i>	xue2xiao1*	T4-as-T1
Typical	da2an4 答案 <i>answer</i>	da3*an4	T2-as-T3
Typical	tiao4wu3 跳舞 <i>dance</i>	tiao4wu2*	T3-as-T2
Atypical	yu2chun3 愚蠢 <i>stupid</i>	yu4*chun3	T2-as-T4
Atypical	fang1mian4 方面 <i>aspect</i>	fang1mian2*	T4-as-T2
Atypical	fa1she4 发射 <i>launch</i>	fa3*she4	T1-as-T3
Atypical	jie2guo3 结果 <i>result</i>	jie2guo1*	T3-as-T1

Figure 1: Acceptance rates by experience and condition

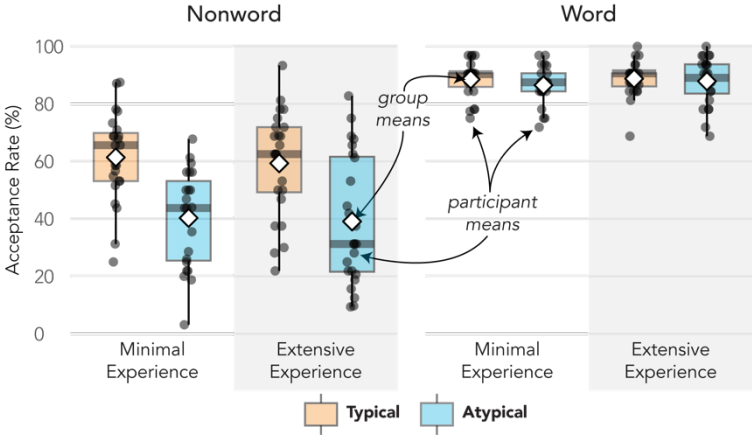
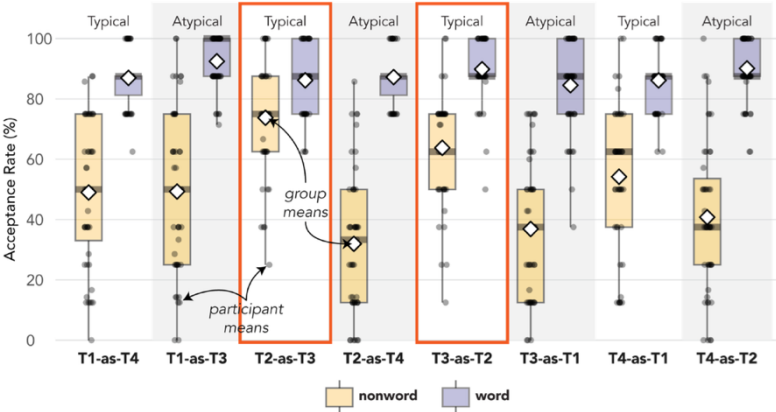


Table 2: Results of the Mixed-effects Model

Effect	Estimate	Std.Err	z-value	Pr(> z)
Intercept	1.529	0.177	8.645	<.001
Experience[Extensive]	0.136	0.224	0.606	.544
Word Status[Word]	3.053	0.213	14.334	<.001
Condition[Typical]	0.728	0.258	2.825	.005
Word Status:Condition	-0.965	0.336	-2.874	.004

Figure 2: Acceptance rates for specific types of tonal bad maps with T2-as-T3/T3-as-T2 highlighted



References:

- [1] Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, 119(1), 131–136. <https://doi.org/10.1016/j.cognition.2010.10.018>
- [2] Samuel, A. G., & Larraza, S. (2015). Does listening to non-native speech impair speech perception? *Journal of Memory and Language*, 81, 51–71. <https://doi.org/10.1016/j.jml.2015.01.003>
- [3] Leung, A. S. (2008). Tonal assimilation patterns of Cantonese L2 speakers of Mandarin in the perception and production of Mandarin tones. In *Proceedings of the 2008 Annual Conference of the Canadian Linguistic Association*
- [4] Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, 40(2), 269–279.
- [5] Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.

Prosodic variation in negative and positive polar questions

Johannes Heim¹ & Judith Schlenter² & Sophie Repp²

¹University of Aberdeen, ²University of Cologne

In pursuit of accounting for question bias, positive polar questions, PQs (1), and negative polar questions, NQs (2), have been compared in their prosodic profiles [1][2][3].

- (1) Are you exhausted? (2) Aren't you exhausted?

Key methodological challenges have been the relative infrequency of NQs in corpora of natural conversation, and the limited naturalness of speech in laboratory settings. To approach these challenges and strike a balance between data availability and naturalness, we systematically extracted samples from an American English TV SOAP corpus [4] and tested previous claims about the prosodic characteristics of PQs and NQs regarding nuclear contours and the accentuation of the clause-initial auxiliary, AUX. The existing literature reports variation within and between the two question types. PQs are often reported to have a final rise if information-seeking [5] but they can also occur with falls [6] or plateaus [7]. Sometimes these variations are associated with rhetorical or assertive interpretations, but it is an open issue whether falls exclude an information-seeking interpretation [8]. Further complication arises from cross-dialectal variation in question rises as they vary in nuclear pitch accents [8]. Accordingly, a wide range of contours are attested for PQs (L* H-H%, L* H-L%, H* H-H%, H* H-L%). NQs are equally heterogeneous, albeit with some candidates for potential markers of speaker bias. Depending on where the speaker is leaning toward (*p* or *Øp*), there may be a correlation of bias with the steepness of a rise [3][5]. The AUX with contracted negation is said to predominantly occur with a rising L+H* accent [10], which has been interpreted as marking polarity contrast. Yet in PQs, where AUX is often unaccented, there is also polarity contrast [5].

In our paper, we show that previous claims on the prosody of PQs and NQs requires further qualification. In a corpus study of the TV soap *The Bold and the Beautiful*, we identified information-seeking PQs and NQs produced by four actors (2m, 2f). Non-information-seeking candidates were excluded via transcript analysis, probing for an answer (expectation) in the conversation. The final data set ($n = 160$) contained 20 PQs and 20 NQs per actor. In all questions, the clause-initial AUX (+*n't*) was followed by a subject pronoun. We annotated AUX accents and nuclear contours (MAE-ToBI). Our results for the nuclear contours (Fig. 1a&b), which mostly were rising or contained plateaus, showed no overall differences between PQs and NQs except for somewhat fewer low rises (L* L-H%) in NQs vs. PQs. There were strong individual differences, indicating that individual speakers distinguish between question types, albeit with different preferences: the two male speakers seem to have opposite preferences (Fig. 1b). In the PQs, there was overall great contour variation, confirming earlier findings. AUX accentuation (Fig. 2a&b) was low in PQs (very low for male speakers), also confirming earlier findings. However, unlike previous reports, AUX in NQs did not have predominantly L+H*, but showed substantial variation: (i) there were just as many H* and L*+H accents; (ii) we observed an accent previously reported for German, a late-medial peak (L+H)* [7][10], which was followed by a fall on the subsequent unstressed syllable; (iii) there was what we mark as L*|H in Fig. 2a&b: a rise whose H tone is realized on the unstressed syllable without perceptible transition, resulting in a tonal *step*; (iv) unaccented AUX. The occurrence of L*|H did not correlate with lexical item, i.e., it did not correlate with the amount of segmental material of AUX. Our findings suggest that there is much more variation in the prenuclear region than was previously discussed for NQs and than is generally assumed for prenuclear accents. Our context analysis indicates that the choice of accent may be related to speaker bias: the steeper or later the rise, the stronger the speaker is biased toward their own belief. For nuclear contours, steepness of the rise and height of the final boundary tone (which in NQs tended to be higher) might serve a similar function. We will present corroborating data from a perception study.

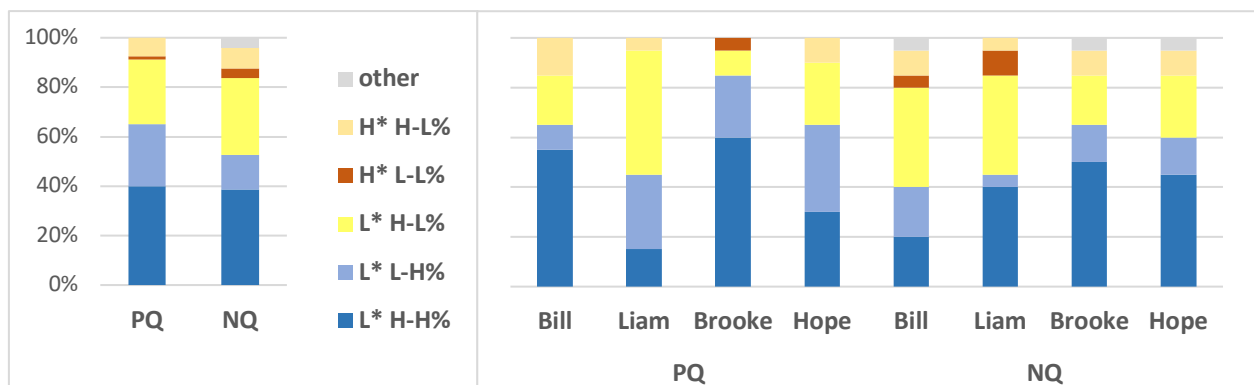


Figure 1a&b. Proportion of nuclear contours in PQs and NQs: total & by individual

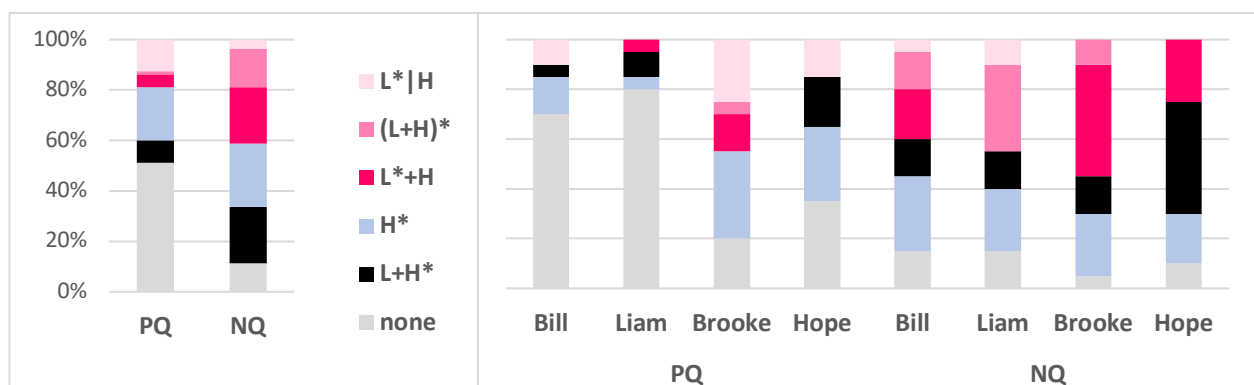


Figure 2a&b. Proportion of AUX accent types in PQs and NQs: total & by individual

- [1] Romero, M. 2024. Biased Polar Questions. *Annual Rev. of Linguistics*, 10(1), 279-302.
- [2] Domaneschi, F., Romero, M., & Braun, B. 2017. Bias in polar questions: Evidence from English and German production experiments. *Glossa* 2(1): 26. 1–28.
- [3] Arnhold, A., Braun, B., & Romero, M. (2021). Aren't prosody and syntax marking bias in questions? *Language and Speech* 64(1), 141-180.
- [4] Davies, M. 2011-. *Corpus of American Soap Operas*. <https://www.english-corpora.org/soap/>.
- [5] Hedberg, N., & Sosa, J. M. 2002. The prosody of questions in natural discourse. In *Speech Prosody 2002, International Conference (Aix-en-Provence)*.
- [6] Geluykens, R. 1989. R(a)ising questions: Question intonation revisited: A reply to Batliner and Oppenrieder. *Journal of Pragmatics* 13(4), 567-575.
- [7] Braun, B., Dehé, N., Neitsch, J., Wochner, D., & Zahner, K. 2019. The prosody of rhetorical and information seeking questions in German. *Language & Speech*, 62(4), 779–807.
- [8] Nilsenová, M. (2006). *Rises and falls. Studies in the semantics and pragmatics of intonation*. University of Amsterdam.
- [9] Armstrong, M., Piccinini, P., & Ritchart, A. (2015). The phonetics and distribution of non-question rises in two varieties of American English. *ICPhS*.
- [10] Zahner-Ritter, K., Einfeldt, M., Wochner, D., James, A., Dehé, N., & Braun, B. 2022. Three kinds of rising-falling contours in German wh-questions: Evidence from form and function. *Frontiers in Communication* 7, 838955.

Perception of Stress in Disyllabic Words in Changsha Xiang: The Effects of Syllable Duration and F0 Contour

ZHOU, Wangqian

The Chinese University of Hong Kong
zhouwangqian@link.cuhk.edu.hk

Lexical stress, when conceived of as a structural position that reflects prosodic headedness in the lexical domain, has been approached from the perspective of neutral/neutralised tones in Chinese [1, 2]. For instance, acoustically, Changsha tones are shortened and flattened in metrical weak positions [3, 4]. Although there have been a few studies on how metrical structures of a tone language influence its native speakers' perception of lexical stress in a second language [5, 6], as well as a number of studies on how Mandarin listeners used acoustic cues to identify and discriminate stress patterns in their native language [7, 8], the perceptual mechanisms underlying stress perception in other Chinese varieties like Changsha Xiang are insofar understudied.

To seek perceptual evidence for the exploitation of acoustic cues (syllable duration and pitch contour) in perceiving disyllabic word stress in Changsha Xiang, an AX discrimination task was used to test listeners' performance in acoustically manipulated conditions. One male and one female speaker of Changsha Xiang were recruited to produce the materials for the AX task. 12 pairs of real disyllabic words contrastive in stress position, like /'eiã34 teiau33/ ('banana') and /eiã34 'teiau34/ ('to intersect'), were produced with two embedding sentences, one providing the contextual cue to its stress placement, the other being a constant metalinguistic statement. In addition, 36 pseudowords with phonotactically illegal structures in Changsha Xiang ([raka], [roko], and [riki]) were constructed. Each pseudoword was elicited following the production of a real disyllabic word that served as a tonal reference. To further clarify stress positions, all pseudowords were also embedded in two declarative carrier sentences as either a verbal or a nominal element.

The word tokens were first segmented and prosodically analysed using ProsodyPro [9]. The acoustic analyses of the stimuli corroborated previous descriptions of the phonetic implementation of Changsha stress (see Figure 1). Three acoustic conditions were then contrived by manipulating the auditory stimuli so that they might retain or obscure the syllable duration and F0 contour as cues to stress. For the both cues condition, all tokens were normalised to 600 ms total duration and 66 dB intensity. For the duration only condition, a new set of tokens was created by resetting the syllables of each right-dominant word to 340 ms and 260 ms (left-dominant tokens' average timing). For the F0 only condition, both syllables were reset to 300 ms for all words. 30 native speakers of Changsha Xiang were involved to indicate if pairs of disyllables had either the same or different stress patterns.

Three-way repeated measures analyses of variance (ANOVAs) were conducted on participants' accuracy rate, reaction time, and sensitivity d' in responding to real words or pseudowords bearing different tones across different acoustically manipulated conditions. The results showed that while both syllable duration and F0 contour provided facilitation to the discerning of trochaic and iambic stress patterns in Changsha, the effect of pitch contour is more important than the effect of syllable duration. Specifically, accuracy and sensitivity were highest when both cues were present, followed by the F0 only condition, and lowest in the duration only condition (Figure 2 and Figure 3). Additionally, the pitch cue was more helpful in processing tones that exhibited larger acoustic differences from their unstressed alternants (Tone 6, Tone 1, Tone 2 and Tone 4), resulting in higher accuracy and sensitivity, as well as shorter reaction times. The pitch cue was also particularly effective in perceiving real words, which is likely due to the benefit from semantic knowledge in top-down processing.

The results challenged the Functional Load Hypothesis of stress marking [10, 11] by showing that pitch information remains paramount in perceiving lexical prominence despite its role in making lexical tone distinctions. This underscores the need to reevaluate prevailing theories regarding stress perception in tone languages. Moreover, the results also called for further acoustic, perceptual, typological and second language research. For example, they provided valid reference for research into Chinese listeners' perceptual learning of stress in other languages, and also the employment of cues in perceiving different levels of prominence.

Keywords: perceptual cue, metrical stress, tonal neutralisation, Changsha Xiang

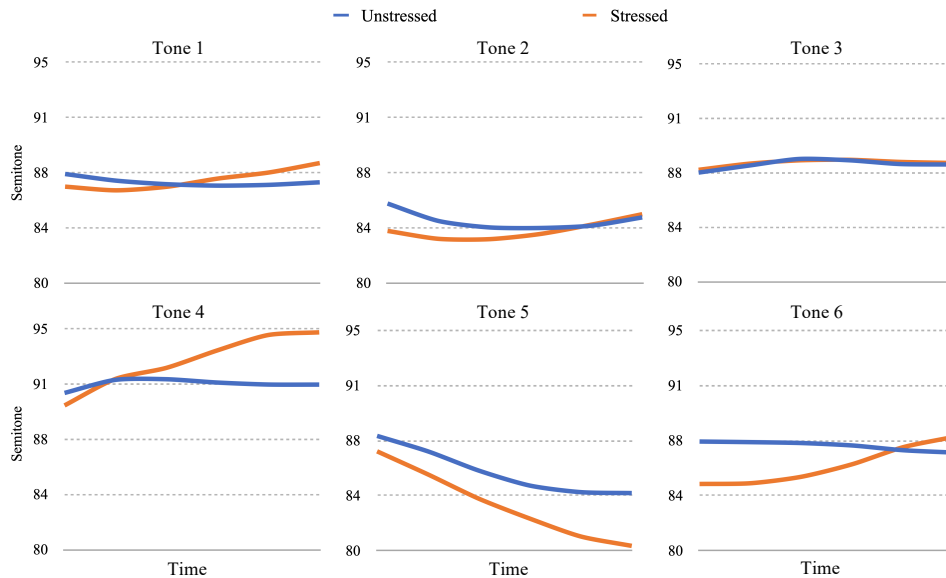


Figure 1. Time-normalised pitch contours of Changsha tones in the second syllable of right- dominant (stressed) and left-dominant (unstressed) disyllables in semitone

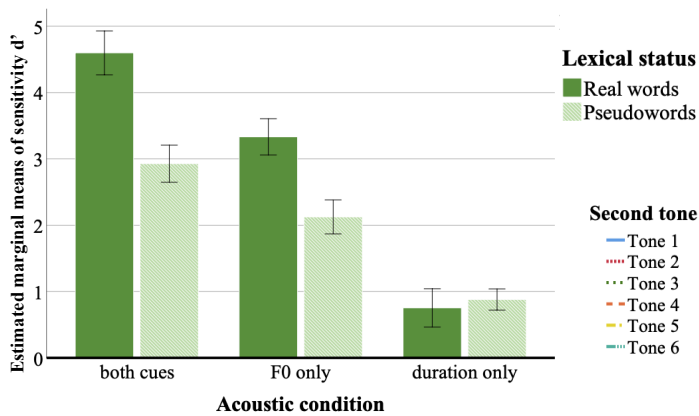


Figure 2. Estimated marginal means of sensitivity d' across lexical statuses and acoustic conditions

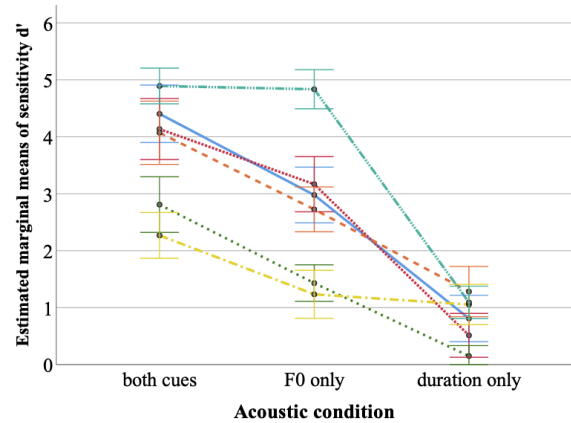


Figure 3. Estimated marginal means of sensitivity d' across acoustic conditions and second tone categories

Selected References

- [1] Duanmu, S. (2007). *The Phonology of Standard Chinese*. Oxford University Press.
- [2] Feng, S. (2017). *Prosodic Morphology in Mandarin Chinese*. Routledge.
- [3] Guo, C., & Chen, F. (2022). Phonetic realizations of metrical structure in tone languages: Evidence from Chinese dialects. *Frontiers in Psychology*, 13, Article e945973. <https://doi.org/10.3389/fpsyg.2022.945973>
- [4] Zhang, M. (2021). Positional tone sandhi in Changsha Xiang: A production study [conference presentation abstract]. In *1st International Conference on Tone and Intonation (TAI), Sonderborg, Denmark: Abstract book* (pp. 138-139). <http://dx.doi.org/10.13140/RG.2.2.31526.86087>
- [5] Yu, V. Y., & Andruski, J. E. (2010). A Cross-Language Study of Perception of Lexical Stress in English. *Journal of Psycholinguistic Research*, 39(4), 323–344. <https://doi.org/10.1007/s10936-009-9142-2>
- [6] Guo, X., & Chen, X. (2022). Perception of English Stress of Synthesized Words by Three Chinese Dialect Groups. *Frontiers in Psychology*, 13, 803008–803008. <https://doi.org/10.3389/fpsyg.2022.803008>
- [7] Wang, Y. (2004). The effects of pitch and duration on the perception of the neutral tone in Standard Chinese. *Acta Acustica*, 29(5), 453-461. <https://dx.doi.org/10.15949/j.cnki.0371-0025.2004.05.013>
- [8] Fan, S., Li, A., & Chen, A. (2018). Perception of lexical neutral tone among adults and infants. *Frontiers in Psychology*, 9, Article e00322. <https://doi.org/10.3389/fpsyg.2018.00322>
- [9] 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.
- [10] Berinstein, A. E. (1979). A cross-linguistic study on the perception and production of stress. *UCLA Working Papers in Phonetics*, 47, 1-59.
- [11] Van Heuven, V. J., & Turk, A. (2020). Phonetic correlates of word and sentence stress. In C. Gussenhoven, & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp.1501–1565). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198832232.013.8>

Tonal sound change in Kagoshima Japanese: Production of monosyllabic words

Lia Saki Bučar Shigemori, Yuki Asano
Institute for Phonetics and Speech Processing, Munich
lia@phonetik.uni-muenchen.de, yuk.asano@lmu.de

Keywords: varieties of Japanese, tonal sound change

The aim of this study is the analysis of the surface realisations of lexical pitch patterns by different speaker groups in two varieties of Japanese, Kagoshima Japanese (KJ) and Tokyo Japanese (TJ), with a focus on how the increasing influence of TJ affects the pitch patterns of younger KJ speakers.

Varieties of Japanese exhibit great differences in their word-prosodic systems [1]. Monosyllabic words in both KJ and TJ can be grouped into two groups regarding their accent patterns, however there is no general way to predict the accent pattern in one variety based on the other. Monosyllabic words in KJ produced in isolation are realized with a falling f_0 if they belong to the Type A pitch accent group and with a flat f_0 if they belong to the Type B group, regardless of whether the syllable is monomoraic (coda-less monosyllabic words with a short vocalic nucleus) or bimoraic (monosyllabic words with a long vowel, a diphthong, or a short vowel followed by a nasal coda). In TJ, monosyllabic words can have two lexically contrastive pitch accent patterns, namely the unaccented pattern, where there is no f_0 fall, and the accented pattern, where there is a fall after the first mora of the syllable. In case of monomoraic words, the contrast is produced when the word is followed by a particle, but is neutralized when produced in isolation. In bimoraic monosyllabic words, the contrast is not neutralized, but the majority of words is accented.

For KJ previous studies have reported an ongoing tonal sound change due to the influence of TJ, the variety the standard is based on and is most widely used in the media, in particular in younger speakers [2, 3]. This tonal sound change observed for KJ has been interpreted as a recategorization of the mapping between the lexical item and its pitch accent categories. In the case of monosyllabic words, the resulting outcome of the sound change in KJ is the neutralization of accent patterns in monomoraic and bimoraic monosyllabic words. Younger speakers of KJ have been reported to produce monomoraic words with the Type B pattern, which can be interpreted as a transfer of the TJ pattern, and bimoraic monosyllabic words with the Type A pattern, which can be interpreted as a transfer of the generalized TJ pattern¹.

While in the previous studies, the results were based on the author's perceptive impression, in the current study we want to replicate the study by directly analysing the f_0 of the recorded data. This might allow us to see additional fine grained differences.

35 KJ speakers and 35 TJ speakers read 32 monomoraic tokens and 14 bimoraic tokens in isolation. Table 1 shows the number of recorded tokens that belong to a certain accent group in KJ and TJ. Because the sound change in KJ has been described as being led by younger speakers who are more exposed to the standard, we divided the speakers into an older group (aged above 50) and a younger group. A discrete cosine transformation (DCT) [4, Chapter 21] has been carried out on the extracted f_0 contours, and the first coefficient (k_1), which reflects the steepness has been extracted. A k_1 of zero indicates a horizontal f_0 , while the higher the k_1 the steeper the fall of the f_0 .

The results, visualized in Figures 1 and 2, confirm what has been reported in previous studies: The traditional KJ accent pattern contrast is realized by the older KJ speakers for both monomoraic and bimoraic monosyllabic words. Type A words have a higher k_1 than Type B words. The younger KJ speakers neutralize the monomoraic words just like the TJ speakers. The k_1 mean is above 0 for both the TJ speakers and the younger KJ speakers, which might be due to the additional effect of sentence prosody. For bimoraic words, the contrast between Type A and Type B words is much less pronounced for the younger KJ speakers. TJ speakers do produce a clear contrast between accented and unaccented words. Although not visualized here, the data further revealed that the older KJ speakers seem to be affected by the TJ accent patterns, but in a contrary direction. Words that belong to the Type A group in KJ but are unaccented in TJ were produced with a higher k_1 , thus an even steeper slope. Possibly, older KJ speakers are well aware of the TJ accent patterns and hyperarticulated the falling f_0 in Type A words which are produced with a flat f_0 in TJ.

The findings provide a starting point to explore how suprasegmental features are categorized and stored when they differ in the language varieties and whether the extent or the way speakers are exposed to these effects also affects their awareness of the differences.

¹In the case of monosyllabic words the sound change looks like an imitation of the TJ patterns, but based on other words, previous literature shows that the surface tonal realization of KJ is retained by the innovative speakers.

Table 1: Number of tokens for the TJ accented/unaccented (and ambigue, in case both patterns are correct) and KJ TypeA/Type B combination, on the left for monomoraic and on the right for bimoraic words.

	accented	unaccented	ambigue		accented	unaccented
Type A	5	10	1	Type A	5	3
Type B	12	2	2	Type B	6	0

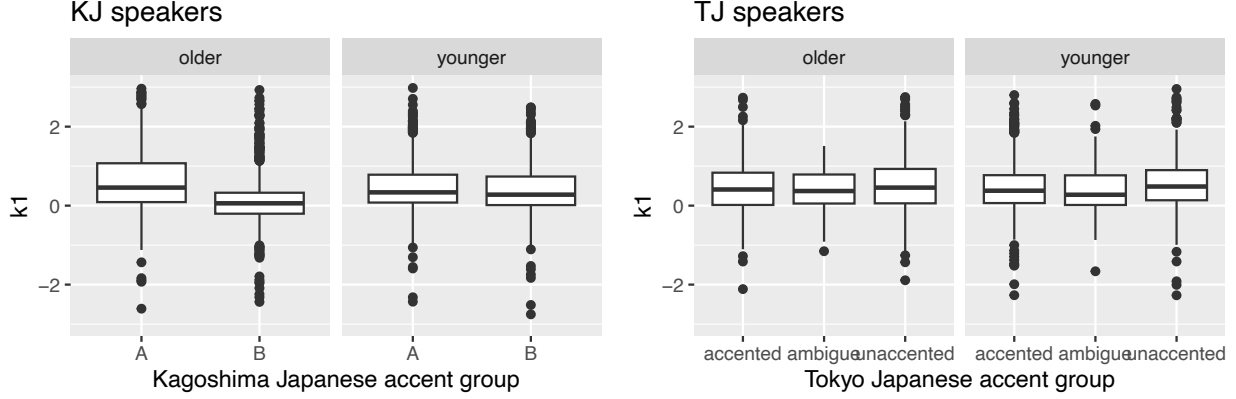


Figure 1: k1 values of the DCT analysis of the f0 contour for monomoraic words, KJ speakers on the left and TJ speakers on the right, separately for older and younger speakers. A higher k1 value indicates a steeper f0 fall, a k1 value of 0 indicates a flat f0 curve. The *ambigue* category stands for items for which both accent patterns are acceptable.

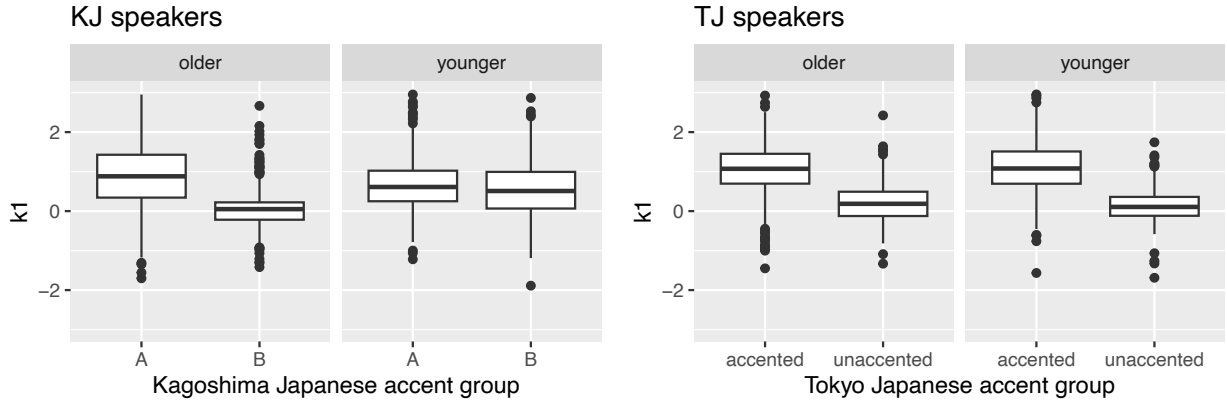


Figure 2: k1 values of the DCT analysis of the f0 contour for bimoraic words, KJ speakers on the left and TJ speakers on the right, separately for older and younger speakers. A higher k1 value indicates a steeper f0 fall, a k1 value of 0 indicates a flat f0 curve.

References

- [1] Yosuke Igarashi. “Typology of Intonational Phrasing in Japanese Dialects”. In: ed. by Sun-Ah Jun. Oxford university press, 2014. Chap. 15.
- [2] Haruo Kubozono. “Postlexical Tonal Neutralizations in Kagoshima Japanese”. In: *Tonal Change and Neutralization*. Ed. by Haruo Kubozono and Mikio Giriko. Berlin, Boston: De Gruyter Mouton, 2018, pp. 27–57. ISBN: 9783110567502. DOI: doi:10.1515/9783110567502-003.
- [3] Haruo Kubozono. “Bilingualism and Accent Changes in Kagoshima Japanese”. In: *Tonal Change and Neutralization*. Ed. by Haruo Kubozono and Mikio Giriko. Berlin, Boston: De Gruyter Mouton, 2018, pp. 279–329. ISBN: 9783110567502. DOI: doi:10.1515/9783110567502-011.
- [4] Raphael Winkelmann. *The EMU-SDMS Manual*. 2022. URL: <https://ips-lmu.github.io/TheEMU-SDMS-Manual>.

Word-level acoustic prominence in Siwi Berber: a paradigmatic and syntagmatic study

Constantijn Kaland & Valentina Schiattarella

Institut für Linguistik/-Phonetik – Universität zu Köln, Università degli studi di Napoli, L'Orientale

ckaland@uni-koeln.de, vale.schiattarella@gmail.com

Siwi is a Berber language (Afro-Asiatic, iso: siz) spoken in the Siwa and El Gara oases (Egypt). Over 30,000 people live in the main oasis of Siwa, the majority of whom speak Siwi as their first language. Siwi is claimed to have word stress that distinguishes verbs (stress on the first syllable of the stem) from nouns (stress on the last syllable when considered in isolation, or on the penultimate syllable, but only when preceded by a preposition). A shift of stress in nouns is claimed to signal definiteness (penultimate) and indefiniteness (ultimate), see [3-7, 9] for stress shift as a referential tracking device. Penultimate word stress is reported for Zwara Berber in Libya [2], and the definite/indefinite distinction seems to be found in some of the eastern Berber languages [10]. The current study is critical with the respect to the claims in the literature, in particular concerning the noun/verb distinction based on words in isolation. The literature on word stress has shown repeatedly that isolated words tend to be uttered with phrase intonation, giving them particular tonal patterns that might not relate to the (lexical) word level (e.g., [1]; [8]). In the current study, we therefore first *explore* what the acoustic information from three prosodic cues (f_0 , intensity, duration) may reveal about word prominence patterns in Siwi, by both paradigmatic and syntagmatic analyses, after which we *test* the definiteness hypothesis for a subset of nouns for which the definiteness was marked by a language expert, if they were unequivocally clear from the context. Crucially, our study is the first acoustic analysis of Siwi prosody based on a large set of spontaneous data.

Data for this study were collected during several field trips since 2011, mainly consisting of spontaneous data (monologues and dialogues of varying lengths). Short narratives of approximately 1.5 to 8 minutes in length from 7 speakers (2 female, 5 male, age range: 18-50 years) were selected for analysis. The narratives were transcribed with the help of native speakers and segmented on the syllable, word and (intonation) phrase level. Pre-final and final syllables in the phrase were excluded to avoid effects of phrase(-final) intonation. Furthermore, a subset of the most common syllable structures (CCV, CCVC, CV, CVC, CVCC, V, VC) and word classes (adverbs, nouns and prepositions) was taken as to obtain a representative set of syllables and acoustic measures. For each syllable, time-series f_0 and time-series intensity was measured (each 20 measurement points), as well as its duration (ms.). For f_0 , the first derivative (velocity) was taken as to abstract from absolute levels of fundamental frequency and to have an indication of the extent to which f_0 movements stand out. To obtain a single value representing f_0 velocity in the syllable, the median value of the absolute velocity values was taken. In this way, it did not matter whether velocity expressed the rate of rising or falling, as both could be cues to prominence. As for intensity, the median value from the time-series (in dB) was taken, as the time-series were shown to be highly sensitive to segmental material (e.g., highly fluctuating intensity around plosives). Then, for each prosodic cue, the acoustic measure was rescaled to account for 1) syllable structure, 2) position in the word (final or non-final), 3) speaker differences and 4) recording intensity. F_0 and duration values were median-centered, where the median was taken from all syllables with the same syllable structure and the same word position from the same speaker. Intensity was median-centered in the same way for syllable structure and word position, except that this was done for each *recording* separately, as to account for intensity differences between two recordings from the same speaker. Then, the values were scaled between 0 and 1 for each cue, and a prominence score per syllable was taken by summing the rescaled values (score range 0-3). In this way, each cue contributed equally to the prominence score (paradigmatic). Then it was checked which syllable had the highest prominence value in the word in belonged to (i.e., syntagmatic, likelihood range 0-1, averaged only if all syllables in the word were present).

The results (Table 1) show that in disyllabic words, the ultimate syllable tends to have the highest prominence score in almost all analyses. In triyllabic word, the penultimate syllable tends to be the one with the highest prominence scores. The exceptions are found for trisyllabic nouns and prepositions. Quadrisyllabic words show less consistent patterns, probably due to the small amounts of data (20 words in total). The co-occurrence of prepositions with a noun and possibly also a possessive in Siwi might have caused a different stress pattern. Overall, however, these observations are trends and prominence differences are in some cases too small to draw firm conclusions. The analysis of the nouns show that definiteness did not consistently affect the prominence scores. Disyllabic words always had the highest prominence on the ultimate syllable, whereas trisyllabic words had the highest prominence inconsistently on the penultimate or ultimate syllable (Table 2). Overall, the current study suggests that Siwi words have right-aligned prominence, either on the ultimate (=disyllabic) or both penultimate and ultimate (>disyllabic), corroborating [2]. We plan to incorporate more data and do perceptual verification of these results in the near future.

Table 1. Prominence scores (paradigmatic) and likelihoods of being the syllable with the highest prominence score in the word (syntagmatic) for each syllable in polysyllabic adverbs, nouns and prepositions (+noun/possessive). Empty cells represent insufficient data. Grey cells indicate highest values per word.

class	analysis	disyllabic		trisyllabic			quadrisyllabic			
		1	2	1	2	3	1	2	3	4
adverb (n=162)	paradigmatic	1.20	1.39	1.10	1.55	1.33				
	syntagmatic	0.34	0.66	0.14	0.71	0.14				
noun (n=186)	paradigmatic	1.18	1.29	0.96	1.22	1.22	1.11	1.06	1.01	0.96
	syntagmatic	0.45	0.54	0.09	0.46	0.45				
preposition (n=295)	paradigmatic	1.23	1.21	1.11	1.30	1.29	1.22	1.22	1.40	0.94
	syntagmatic	0.48	0.56	0.23	0.19	0.57	0.50	0.33	0.00	0.00

Table 2. Prominence scores (paradigmatic) and likelihoods of being the syllable with the highest prominence score in the word (syntagmatic) for each syllable in polysyllabic nouns that were marked for their definiteness in the discourse context. Grey cells indicate highest values per word.

definiteness	analysis	disyllabic		trisyllabic		
		1	2	1	2	3
definite (n=67)	paradigmatic	1.14	1.25	0.99	1.23	1.18
	syntagmatic	0.38	0.59	0.00	0.57	0.40
indefinite (n=56)	paradigmatic	1.07	1.25	0.99	1.22	1.32
	syntagmatic	0.44	0.56	0.25	0.25	0.50

References

- [1] M. K. Gordon and T. Roettger, “Acoustic correlates of word stress: A cross-linguistic survey,” *Linguistics Vanguard*, vol. 3, no. 1, p. 20170007, Aug. 2017, doi: 10.1515/lingvan-2017-0007.
- [2] C. Gussenhoven and W.-R. Chen, “Segmental intonation in Zwara Berber voiceless stressed syllable peaks,” in *Speech Prosody 2022*, ISCA, May 2022, pp. 283–286. doi: 10.21437/SpeechProsody.2022-58.
- [3] N. Louali, “L’accent en berbère: Catégorie grammaticale et démarcation syntaxique,” in *Nouvelles études berbères. Le verbe et autres articles*, K. Naït-Zerrad, R. Voßen, and D. Ibrizimow, Eds., in *Berber Studies*, Köln: Rüdiger Köppe Verlag, 2003, pp. 67–77.
- [4] N. Louali and G. Philippson, “L’accent en Siwi (Berbère d’Egypte),” in *Actes des XXVes Journées d’Étude sur la Parole (JEP 2004)*, B. Bel and I. Marlien, Eds., Fès: Association Francophone de la Communication Parlée, 2004, pp. 325–328.
- [5] C. Naumann, *Acoustically based phonemics of Siwi (Berber)*, 1. Aufl. in *Berber Studies*, no. 36. Köln: Köppe, 2013.
- [6] V. Schiattarella, “Accent On Nouns And Its Reference Coding In Siwi Berber (Egypt),” in *Nominal Anchoring: Specificity, definiteness and article systems across languages*, K. Balog, A. Latrouite, and R. van Valin, Eds., Berlin: Language Science Pres, 2020, pp. 149–170. Accessed: Dec. 13, 2024. [Online]. Available: <https://africarxiv.ubuntunet.net/handle/1/788>
- [7] L. Souag, *Berber and Arabic in Siwa (Egypt): A study in linguistic contact*. in *Berber Studies*, Vol. 37. Köln: Rüdiger Köppe Verlag, 2013.
- [8] V. J. Van Heuven, “Acoustic Correlates and Perceptual Cues of Word and Sentence Stress: Towards a Cross-Linguistic Perspective,” in *The Study of Word Stress and Accent*, 1st ed., R. Goedemans, J. Heinz, and H. Van der Hulst, Eds., Cambridge University Press, 2018, pp. 15–59. doi: 10.1017/9781316683101.002.
- [9] W. Vycichl, *Berberstudien & A Sketch of Siwi Berber (Egypt)*. in *Berber Studies*, Vol. 10. Köln: Rüdiger Köppe Verlag, 2005.
- [10] W. Vycichl and S. Chaker, “Accent,” *Encyclopédie berbère*, vol. I. Abadir – Acridophagie, pp. 103–106, 1984.

Do L1 Chinese speakers and L1 Spanish speakers use the same prosodic strategies to convey sadness and joy in Spanish?

Cristina Herrero-Fernández

Universidad Nebrija

Pernelle Lorette

Mannheim Universität

Miguel Mateo

Universidade Federal do Rio de Janeiro

Abstract

The current study focuses on the expression of sadness and happiness by L1-Chinese/LX-Spanish bilinguals living in Madrid, as it investigates whether L1 Chinese speakers use prosodic strategies to express sadness and joy in LX Spanish that deviates from the ones used by L1 speakers. A corpus of 100 paired utterances (100 sad speech utterances and 100 joyful speech utterances) produced by ten Chinese Spanish bilinguals living in Madrid was analyzed. The sentences shared identical lexico-grammatical content but differed at the prosodic level. For instance, the sentence “*Está lloviendo*” (“It’s raining”) was produced either with joy or with sadness. The analysis was carried out using the Melodic Analysis of Speech (Cantero, 2002) and the Prosodic Analysis of Speech (Cantero, 2019), two complementary models that focus on suprasegmental features such as pitch, intensity, and temporal structure. The melodic component was examined by identifying relevant pitch values and standardizing them to extract essential melodic contours. For intensity, a standard curve was constructed by cumulatively adding dB values from an initial reference point, producing an algorithm that captures the “melody of intensities”. Temporal analysis was approached by measuring distances between the intensity peaks of tonal centers. Later, 50 paired utterances produced by 5 L1 Spanish speakers were analyzed to compare the prosodic strategies used by both groups. Finally, perceptual tests involving 75 L1 Spanish speakers provided insights into how the emotional speech produced by Chinese L1 speakers in Spanish was perceived and which prosodic cues might be responsible for potential misperception. Utterances produced by L1-Chinese/LX-Spanish bilinguals were evaluated by L1 Spanish speakers in an online perception test. After hearing each recording, raters assessed the speaker’s perceived emotional state using a shortened version of the Two-Dimensional Affect and Feeling Space (2DAFS-18), which measures valence and arousal on a coordinate grid. Statistical analyses focused on comparing perceived vs. intended emotion (joy or sadness). Preliminary results showed that, in general, joy utterances were rated with positive valence, while most sadness utterances received negative valence scores—except two that scored slightly positive (Figure 1). Results show that Chinese L1 speakers did use prosody to try to convey emotions, since most of the paired utterances displayed distinct melodic contours for joy and sadness, with joy speech characterized by greater tonal variability and circumflex contours, and sad speech by flatter melodic patterns and descending inflections. However, some utterances shared identical melodic contours for both emotions, potentially leading to perceptual ambiguity. Other results point out that Chinese-L1/Spanish-LX bilinguals and Spanish L1 speakers use different prosodic strategies when trying to convey sadness and joy in Spanish. For instance, while L1 Spanish speakers show a preference for descending intensity profiles when trying to convey sadness (Figure 2), L1 Chinese speakers show less preference for this pattern and a greater presence of zigzag and ascending profiles in Spanish LX (Figure 3). Moreover, they don’t tend to lengthen the final segments of utterances as much as L1 Spanish speakers, which may contribute to a perception of slowness and solemnity, characteristic of sadness in Spanish (Hidalgo, 2020). Finally, the perception tests revealed that some utterances were not perceived as intended by the LX speakers (Figure 1), which might be related to the different prosodic strategies used by each group.

Keywords: Emotional Speech, LX Spanish, Chinese Speakers, Prosodic Strategies, Intercultural Communication

Figure 1. Notched boxplot displaying valence perception per speaker per intended emotion, with 400 representing neutral ratings.

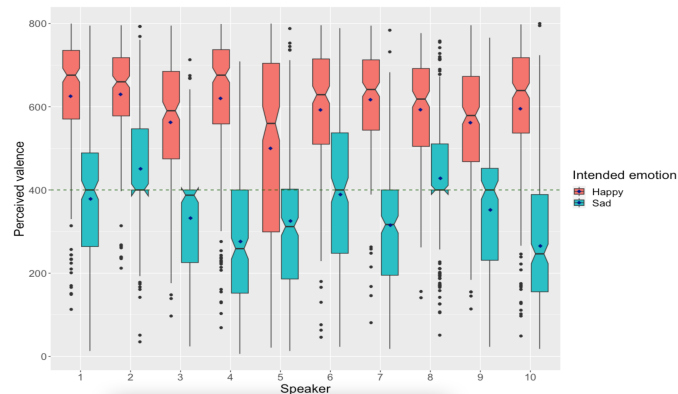


Figure 2. Rythmic and dynamic profile of sad speech utterance spoken by L1 Spanish speaker.

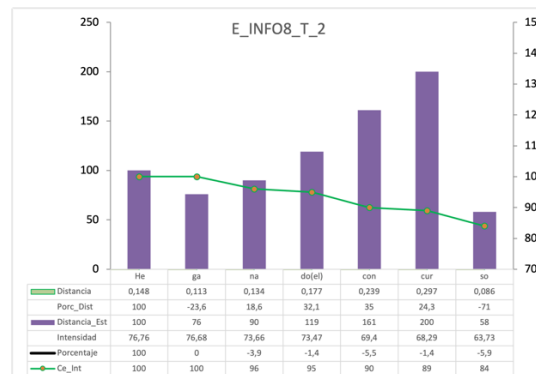


Figure 3. Rythmic and dynamic profile of sad speech utterance in LX Spanish spoken by Chinese L1 speaker.



References

- Cantero, F. J. 2002. *Teoría y análisis de la entonación*. Ediciones de la Universidad de Barcelona.
- Cantero, F. J. 2019. "Análisis prosódico del habla: más allá de la melodía". In *Comunicación Social: Lingüística, Medios Masivos, Arte, Etnología, Folclor y otras ciencias afines*, eds. M.R. Álvarez Silva, A. Muñoz Alvarado y L. Ruiz Miyares. Volumen II. Santiago de Cuba: Ediciones Centro de Lingüística Aplicada.
- Hidalgo, C. 2020. "Prosody and emotions in spontaneous Spanish speech: The case of sadness". *Revista Española de Lingüística Aplicada*, 33(1): 45-67.

Effects of mutual visibility on global pitch features

Maciej Karpiński¹, Bettina Braun²

¹Adam Mickiewicz University in Poznań, ²University of Konstanz

Mutual visibility has been shown to impact communication processes, including speech perception and comprehension (Rosenblum 2008), emotional expression (Bavelas et al. 2014), or turn-taking (Kendrick et al. 2023). Following Lindbloom's H&H hypothesis (1990) and relevant empirical studies (e.g., Wagner & Bryhadyr 2017; Żygis & Fuchs 2019), we expect more extensive usage of speech prosody, e.g., the average pitch level, range and variability in blocked visibility, which may additionally be affected by speaker and interlocutor gender (cf. Levitan et al. 2012).

We analyzed 42 recordings of task-oriented dialogues from the MultiCo corpus (Karpiński et al. 2023), comprising 14 female-female, 13 male-male, and 15 mixed pairs (aged 20-25 years). They were asked to build a tower out of imaginary blocks of their own design, in alternating order, not using any physical requisites. The time was limited to five minutes, and they had to remember the details so that they could draw the tower at the end of the session. Participants stood facing each other at a fixed distance of approximately 4 meters. Visibility was manipulated between subjects: Half of the pairs could see each other, and half could not; the only difference was the presence of a barrier (blend) to block visual contact. Head-on microphones, preserving a fixed distance and position relative to the mouth, were used.

We used the transcriptions and segmentation on the lexical level to construct Interpausal Units (IPU, separated by more than 250 ms of silence) and sorted the data into three periods (first 30 sec, central 3.5 minutes, last minute). Prosody was analyzed in the central 4206 IPU. For each IPU, we extracted the number of words, duration (as co-variate), median *f*₀ and IQR in Hz and *f*₀-range in st (using filtered autocorrelation, in the range btw. 120-600Hz for females and 80-500Hz for males), as well as the mean intensity in dB. Given the skewed IPU lengths, we removed IPU with more than 20 words (N=54, 1.2% of the data). IPU with missing values were also removed (N=668, 16.1%). The variables were analyzed using linear mixed-effects regression models with *IPU duration* (scaled log), *speaker gender* and *interlocutor gender* (both sum-coded) and *visibility* (contrast-coded) as fixed factors in interaction and *speaker* and *content* as random effects (random intercepts and random slopes for duration for *speaker*). Non-significant terms were removed ($\alpha=0.05$). To check the models' validity, data points with residuals larger than 2.5 were removed and the models were refitted. Variance inflation factors were below 6. For **median *f*₀**, results showed a significant effect of visibility (4.9Hz higher in the no-visibility condition, $p<0.001$), which interacted with duration and speaker gender ($p<0.01$, Fig. 1). **IQR** was not affected by visibility ($p>0.5$). ***f*₀-range** showed an effect of visibility (0.47st higher in the no-visibility condition, $p<0.01$), in addition to interactions with IPU duration, gender and interlocutor gender ($p<0.005$, Fig 2). Since IPU duration varied within subjects, the differences in median *f*₀ and *f*₀-range across visibility conditions could not be attributed to group differences alone. Conversely, **mean intensity** was higher in the visibility condition ($p=0.06$), suggesting that the increase in average *f*₀ and *f*₀-range in the no-visibility condition was not due to a general Lombard effect. Finally, visibility interacted with IPU duration and gender ($p<0.01$), Fig 3.

In future work, we plan to test whether pitch features correlate with gesture movement and whether there is differential accommodation across visibility conditions. Furthermore, will annotate intonational phrasing and pitch accents to be able to analyze effects of visibility more locally.

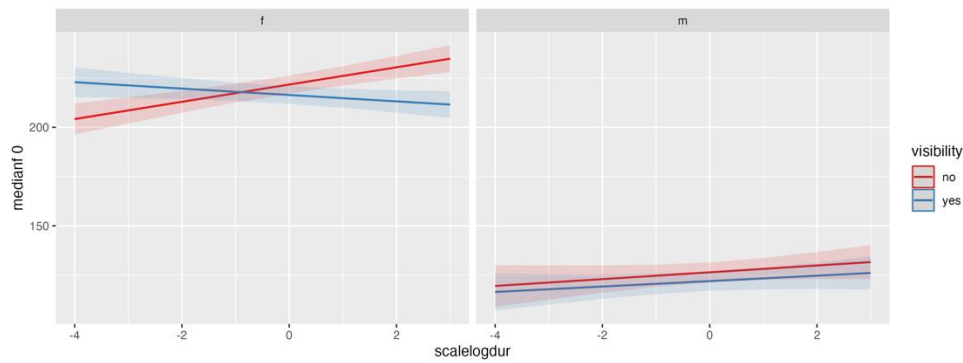


Fig 1. Interaction between IPU duration, speaker gender and visibility for medianf0 (Hz).

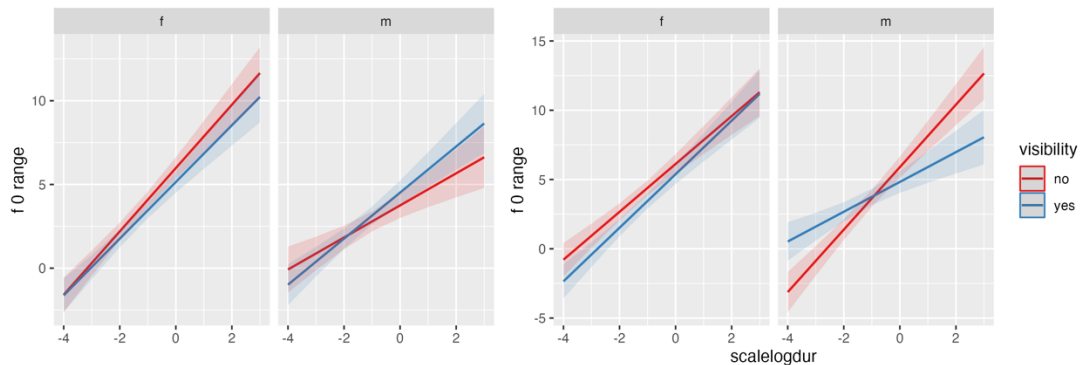


Fig 2. Interaction between IPU duration, speaker gender, interlocutor gender (left two panels: female interlocutor, right panels: male interlocutor) and visibility for f0-range (st).

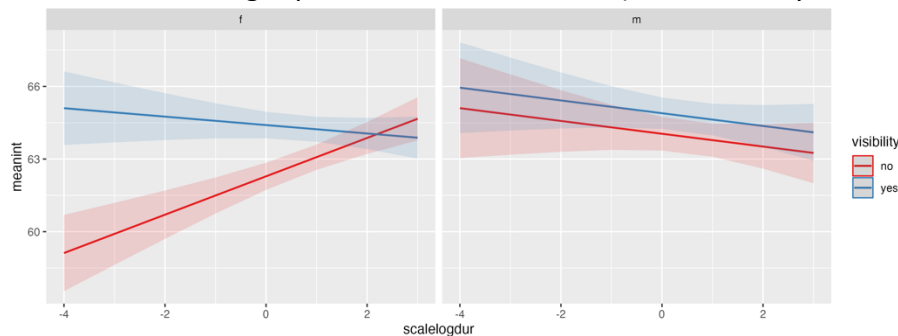


Fig 3. Interaction between duration, gender and visibility for mean intensity (dB).

References

- Bavelas, J., Gerwing, J. & Healing, S. (2014). Hand and Facial Gestures in Conversational Interaction. In: Oxford Handbook of Language and Social Psychology, pp. 111-130.
- Karpiński, M., Jarmołowicz-Nowikow, E., Klessa, K., Piosik, M., Taborek, J. (2023). DARIAH-PL MultiCo Multimodal Corpus. In: Z. Vetulani & P. Paroubek (Eds.) *Human Language Technologies as a Challenge for Computer Science and Linguistics*, WN UAM, pp. 136-140.
- Kendrick, K. H., Holler, J., & Levinson, S. C. (2023). Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B*, 378(1875), 20210473.
- Levitan, R., Gravano, A., Willson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. In Proc. Conf. NAACL: HLT (pp. 11-19).
- Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403-439). Dordrecht: Springer Netherlands.
- Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Curr. Dir. in Psych. Sc.*, 17, 405-409.
- Wagner, P., & Bryhadyr, N. (2017). What You See is What You Get Prosodically Less-Visibility Shapes Prosodic Prominence Production in Spontaneous Interaction. In *Proceedings of INTERSPEECH* (pp. 3226-3230).
- Žygis, M., & Fuchs, S. (2019). How Prosody, speech mode and speaker visibility influence lip aperture. In *Proceedings of the 19th ICPHS* (pp. 230-234).

A pitch accent for contrastive emphasis in Danish?

Nicolai Pharao, Department of Nordic Studies and Linguistics, University of Copenhagen
Rasmus Puggaard-Rode, Institut für Phonetik und Sprachverarbeitung, LMU München

Keywords: pitch, contrast, Danish, read speech

According to Grønnum's (2022) model of Danish intonation, contrastive emphasis in Danish is signaled by more extensive F0 contours associated with the stress group in emphasised words (and often also a reduction of the contour in surrounding stress groups in the same utterance). Importantly, in this model the contour associated with emphasised words is qualitatively the same as that found for non-emphasised words, with the difference limited to the pitch level and magnitude of the excursion. This is in contrast to other languages where contrastive emphasis or focusing may be signaled by a pitch accent with a different contour (Gussenhoven 2004:86-7). The model for Danish is based on read speech, but Dyhr (1995) and Tøndering (2004) have empirically verified the qualitative invariance of the stress group pattern in spontaneous speech, Dyhr (1995) also for words with contrastive emphasis.

This study is conducted on read aloud sentences contrasting two phrases in an interrogative frame (based on the design from Kirby & Ladd 2016). 12 speakers of Danish from Greater Copenhagen were recorded in a sound proof studio reading aloud a list of sentences following the mould of "Did you verb X or Y?" and each disyllabic target word occurred in both X and Y positions. The utterances thus invite the speaker to contrast the word in utterance medial position (X) with the word in utterance final position (Y). While such a construction could potentially elicit a higher degree of prominence on the words in utterance medial position, there should be no difference in F0 contour for these items, according to Grønnum's (2022) model. However, listening to the recordings gave the impression that some speakers placed emphatic stress on some utterance medial items. In this study, we test whether this auditory impression is related to differences in F0 contour shape.

Using the contour clustering tool developed by Kaland (2021), we conducted time series analysis of F0 in disyllabic words occurring in either medial or final position. Comparisons of the F0 contours are shown in Figures 1 and 2 overleaf. 308 contours from utterance medial tokens were eligible for clustering analysis. After removing 87 contours with a level contour due to the laryngealization typical of the Danish *stød* accent (Grønnum 2022), 221 contours in medial position were subjected to cluster analysis. Stepping down from 25 possible clusters (following Kaland (2021)), the final analysis revealed only two distinct contours, shown in Figure 1 overleaf. The predominant contour is a rise from a low tone in the accented syllable, but the opposite pattern is also found, i.e. a high tone on the accented syllable followed by a fall to the post-tonic. This high-low pattern is contra to the model in Grønnum (2022) and the findings in Dyhr (1995). Analysing the same words in final position yielded 193 contours available for cluster analysis, after removal of flattened contours due to *stød* (as well as eleven spurious contours) to ensure comparability. The resulting two clusters, shown in Figure 2, reveal a slight difference in level of the contours only, i.e. there is no difference in contour shape. While the analysis presented here suggests the option of a specific contour shape for contrastively emphasised words, it is important to note that only 6.7% (n = 33) of the analyzable contours were realized with this high-low pattern. Contours included in the high-low cluster were only attested for 3 out of the 12 speakers, and none of these speakers exclusively used the high-low pattern for emphasised words. In other words, the attested high-low pattern is at most an optional way to mark emphasis prosodically in contemporary standard Danish, albeit one that has not previously been attested. It suggests the possibility of a pitch-accent for contrastive emphasis developing.

Figures and references

Figure 1 – F0 contours in medial position

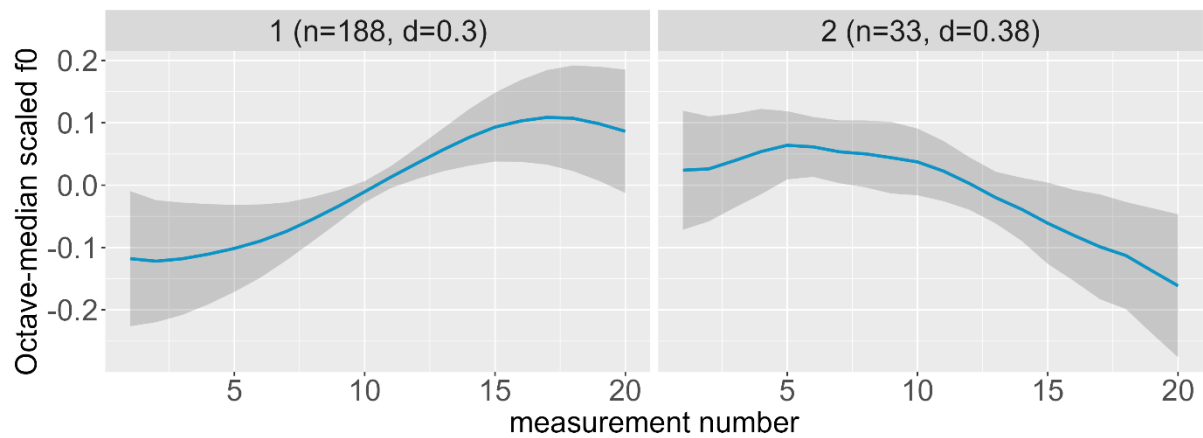
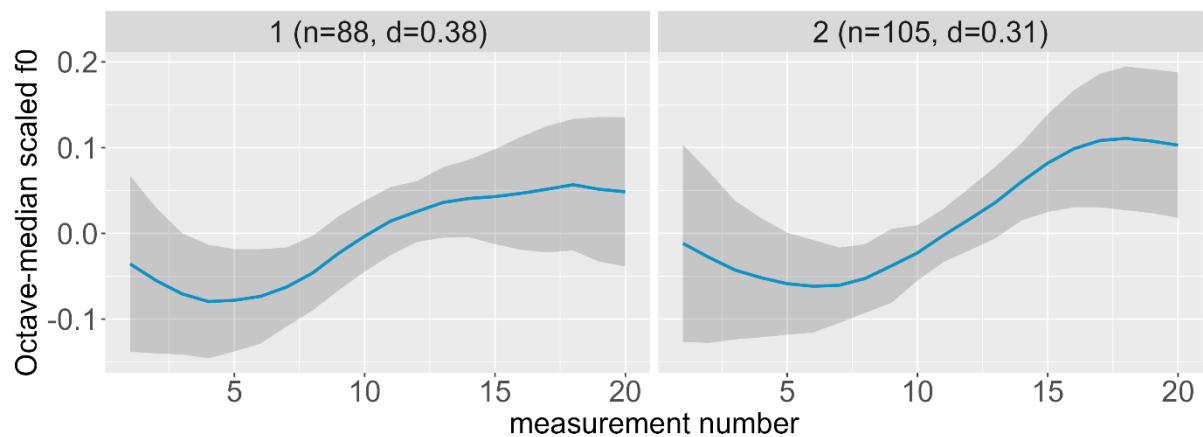


Figure 2 – F0 contours in final position



References

- Dyhr, N.-J. 1995. The fundamental frequency in Danish spontaneous speech with special reference to syllables boosted for emphasis. J. Rischel & H. Basbøll (eds.), *Aspects of Danish Prosody*, 49-67. Odense University Press.
- Grønnum, N. 2022. Modeling Danish intonation. J. Barnes & S. Shattuck-Hufnagel (eds.), *Prosodic theory and practice*, 85–116. The MIT Press. DOI: 10.7551/mitpress/10413.003.0005.
- Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.
- Kaland, C. 2021. Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *Journal of the International Phonetic Association*, 53 (1), 159–188. <https://doi.org/10.1017/S0025100321000049>
- Kirby, J. P. & D. R. Ladd. 2016. Effects of obstruent voicing on vowel F0: Evidence from “true voicing” languages. *J Acoust Soc Am* 140 (4): 2400–2411. <https://doi.org/10.1121/1.4962445>
- Tønndering, J. 2004. *Intonation og informationsstaus – belyst ved prominensmålinger*. Unpublished MA thesis, University of Copenhagen, Denmark.

The combined effect of focus and Beijing retroflex suffixation on tone

Yutong Wang¹, Mitko Sabev^{2,1}, Brechtje Post¹

¹University of Cambridge, ²Universität des Saarlandes

yw590@cam.ac.uk, msabev@lst.uni-saarland.de, bmbp2@cam.ac.uk

The intricate interplay between tone and intonation in Mandarin Chinese and its variants has been frequently discussed, e.g. [1-3], with a primary focus on the study of F0 realisation. This study explores a more complex interaction involving intonation and tone with Beijing retroflex suffixation (*erhua*). Retroflex suffixation in Beijing Mandarin (BM) involves adding a suffix, realised as [ə], to the rime to convey diminutiveness. In addition to inducing spectral centralisation in the suffixed rimes, retroflex suffixation has been shown to compress the F0 contour, lowering overall F0 values while preserving tonal shapes [4]. However, little is known about the realisation of retroflex suffixation in wider prosodic contexts.

This paper presents an acoustic experiment investigating the interaction of Beijing retroflex suffixation with tonal and focus prominence variations. Six native BM speakers (3 female, 3 male; average age = 26) were recorded producing both unsuffixed and suffixed words with different tones: high-level Tone 1, rising Tone 2, low-dipping Tone 3, falling Tone 4, under various focus conditions: broad focus, on-focus, post-focus, and pre-focus, as well as with neutral tone (NT) (on-focus only). Target words were embedded in the same carrier sentence, as shown in (1). Each target word was repeated three times by the speakers, and the carrier sentences were elicited by different prompt questions (presented in a quasi-random order) to manipulate focus (Table 1). NT targets were elicited in reduplicated phrases such as /'tʰu⁵¹.tʰu-DIM/ *tutu-er* ‘bunny’ 兔兔儿 and were produced under the on-focus condition.

(1) Carrier sentence

我 说 了 <target word> 三 遍
/wo:³ ʃuo¹ lɿ <target word> san¹ pien⁴/
I say PAST <target word> three time
'I said <target word> three times.'

The recordings of 1,392 utterances were segmented using [5], with boundaries manually adjusted. F0 values for each syllable were extracted at 20 equidistant points of the rime using the Praat script *ProsodyPro* [6] (Bark-transformed, with outliers removed). Generalised Additive Mixed Models (GAMMs) were built to analyse the effects of suffixation, tone, and focus on F0 contours, including a factor smooth to model the non-linear effect of the three for each speaker. Rime durations were measured and log-transformed (Figure 1).

Results show that F0 lowering due to suffixation does not always occur, and suffixation does not consistently lengthen the duration of syllables, suggesting that rime coalescence, rather than *er*-attachment to the rime, is employed by BM retroflex suffixation. Focus induces hyperarticulation of tonal targets, enhancing tonal contrast [7]. The results further reveal tone-specific susceptibilities to suffixation and focus variation. Post-focus tokens exhibit the lowest F0 values due to compression [8], while duration shortening is observed only in Tones 3 and 4. Lower F0 values after suffixation are found in Tone 3 except in pre-focus condition, whereas Tone 1, 2 and 4 tokens display minimal F0 variation attributable to suffixation. For NTs, the F0 of NT tokens derived from Tone 2 is significantly lowered after suffixation, while those from Tone 3 and 4 are raised; NTs from Tones 1 remain largely unaffected. The strong-weak alternations that are normally associated with focal prominence and NT become obscured under the influence of retroflex suffixation. Furthermore, NT is likely to be associated with lighter metrical weight compared to full tones, evidenced by the global duration reduction of the NT phrases, and lengthening of the NT syllable after suffixation. The results suggest that BM is highly flexible in its tonal realisation at both lexical and sentential levels which interacts with retroflex suffixation to express rich pragmatic functions, e.g. focus marking and diminutiveness.

Condition	Prompt	Answer
broad focus	What did you do just now?	I said <u>flower</u> three times.
on-focus	What did you say three times?	I said flower three times.
post-focus	Who said flower three times?	I said <u>flower</u> three times.
pre-focus	How many times did you say flower?	I said <u>flower</u> three times .

Table 1: Focus conditions and example prompts

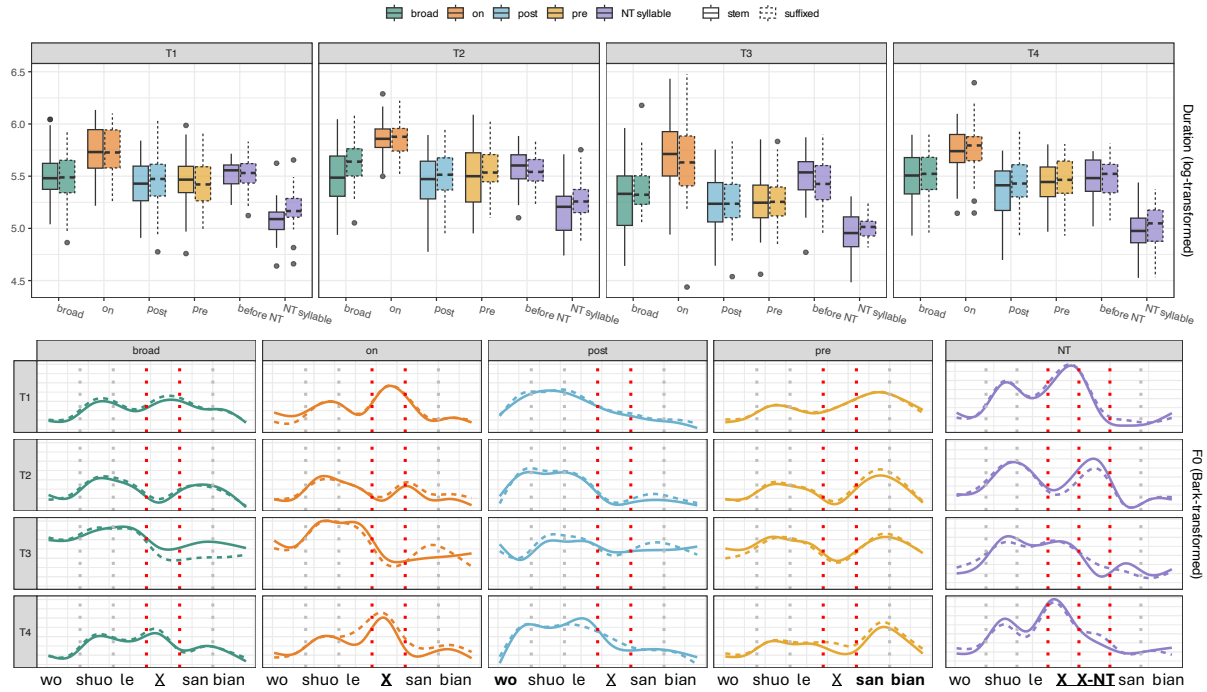


Figure 1: Top: Duration (log-transformed); Bottom: F0 (Bark-transformed) of the target words (X, between the red vertical lines) under various focus conditions; the focused parts are indicated in **bold**.

References

- [1] C. Xu, ‘Cross-dialectal perspectives on Mandarin neutral tone’, *Journal of Phonetics*, vol. 106, p. 101341, Sep. 2024, doi: 10.1016/j.wocn.2024.101341.
- [2] Q. Li and Y. Chen, ‘Prosodically conditioned neutral-tone realization in Tianjin Mandarin’, *Journal of East Asian Linguistics*, vol. 28, no. 3, pp. 211–242, Aug. 2019, doi: 10.1007/s10831-019-09194-4.
- [3] K. K. Li, F. Nolan, and B. Post, ‘Clustering lexical tones with intonation variation’, in *Proceedings of the Second International Conference on Tone and Intonation*, Singapore: Chinese and Oriental Languages Information Processing Society, 2023, pp. 87–88.
- [4] Y. Wang, M. Sabev, and B. Post, ‘Dynamic insights into incomplete neutralisation in Beijing Mandarin: retroflex suffixation and rime merger through GAMM analyses.’, presented at the Language Sciences Annual Symposium 2024, Cambridge: Cambridge Open Engage, 2024. doi: 10.33774/coe-2024-t94sh.
- [5] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, ‘Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi’, in *Interspeech*, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:12418404>
- [6] Y. Xu, ‘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, 2013, pp. 7–10.
- [7] Y. Chen, ‘Post-focus F0 compression—Now you see it, now you don’t’, *Journal of Phonetics - J PHONETICS*, vol. 38, pp. 517–525, Oct. 2010, doi: 10.1016/j.wocn.2010.06.004.
- [8] Y. Xu, ‘Effects of tone and focus on the formation and alignment of f0 contours’, *Journal of Phonetics*, vol. 27, no. 1, pp. 55–105, 1999, doi: <https://doi.org/10.1006/jpho.1999.0086>.

A Diachronic Perspective on Left-dominant and Right-dominant Tone Sandhi in Northern Wu Chinese: Evidence from Xiangshan

Yibing Shi, Francis Nolan, Brechtje Post

University of Cambridge

ys538@cam.ac.uk, fjn1@cam.ac.uk, bmbp2@cam.ac.uk

Keywords: lexical tone, tone sandhi, Northern Wu, diachronic analysis

Northern Wu Chinese is renowned for its two distinct tone sandhi systems: left-dominant sandhi (LDS) and right-dominant sandhi (RDS). For example, in Shanghainese [1], LDS manifests as left-to-right tone spreading, where the initial syllable's tone is spread across the whole domain. RDS, in contrast, preserves the tone of the final syllable while reducing the preceding tones. LDS typically occurs in highly lexicalised compounds, whereas RDS is more common in loosely bound syntactic phrases. These two systems—seemingly opposite in their tonal strategies—pose fundamental questions about their origin, function, and evolution. While existing theories attribute these behaviours to stress, the lack of consistent acoustic correlates in Chinese weakens this explanation. This study presents two alternative diachronic accounts for the co-existence of LDS and RDS in Northern Wu.

The Xiangshan dialect, which is spoken in Xiangshan, Zhejiang province in China, has six citation tones: HH, HL, LHL, LH, Hq, and LHq. The historical *ping* and *shang* categories have merged into one synchronic tone in each of the two registers (

Table 1). This study examines the disyllabic tone sandhi patterns in Adjective-Noun compounds and phrases in Xiangshan Chinese. A total of 287 tokens with non-checked tone combinations were collected from eight native Xiangshan speakers (mean age: 50, 4 female). F_0 patterns were identified using k-means clustering and auditory categorisation.

Despite mergers in the citation tone system, disyllabic tone sandhi retain historical distinctions of the initial tone. Adjective-Noun compounds and phrases beginning with a *yingping* tone mostly exhibit a falling (HHML) or a rising (MMMH) sandhi contour, while those starting with a *yingshang* tone favour a level (HHHH) or a rising-falling (MHHL) sandhi (Figure 1). The choice between the two possible outputs was found to be mostly speaker-specific. Similar patterns were found in Low-register-initial tokens: falling (LHML) or rising (LLLH) contours only exist in *yangping*-initial tokens, while *yangshang*-initial ones show level (LLHH) or rising-falling (LLHL) contours (Figure 2). Crucially, the second syllable also influences tonal outcomes in *yangping*-initial disyllables (Figure 3), where level sandhi occurs when the second syllable is HH, rising when LHL or LH, and rising-falling sandhi when HL. This interaction suggests a complex interplay between the left-dominant and right-dominant mechanisms with alternative historical accounts possible.

Diachronic account 1: Left-dominance-originated tone sandhi. Lexical tone sandhi patterns were initially derived via left-to-right tone spreading, creating two distinctive sets of sandhi patterns for *ping*- and *shang*-initial tokens. The developmental paths for disyllabic sandhi and monosyllabic tones have diverged ever since, with monosyllabic tones undergoing tone mergers, and disyllabic sandhi fossilised. However, this account cannot explain the right-dominant elements in the *yangping*-initial disyllabic lexical sandhi, as the non-initial tones should have been completely neutralised under this account.

Diachronic account 2: Change from Right-dominance to Left-dominance. Following Li's hypothesis [2], RDS historically preceded LDS. RDS, as a phonetically-driven process motivated by final lengthening, enables full tonal preservation in phrase-final position, and develops hand-in-hand with the monosyllabic system, thus always reflecting the up-to-date citation tones. LDS emerged later to mark lexical words as a result of a disyllabification trend which gained ground in Middle Chinese. Initially forming in structures like reduplicated items, where semantic dominance of the initial syllable influenced the phonetic representations, LDS later spread to other lexical constructions to contrast with more loosely connected phrases. The Xiangshan data support this account: the right-dominant elements in the *yangping*-initial sandhi patterns might be historical residuals of RDS. This account also aligns with descriptions of Wu tone sandhi in larger areas, where tone sandhi is rarely purely LDS, but usually

a mixture of RDS and LDS. Specifically, Qian’s proposed four stages of LDS—partial connection, dfferentiation, simplification, and spreading [3]—illustrates a transition from RDS to LDS.

Table 1 Tone system of the Xiangshan dialect.

category \ register	<i>ping</i> ‘level’	<i>shang</i> ‘rising’	<i>qu</i> ‘departing’	<i>ru</i> ‘checked’
<i>yin</i> ‘high’	HH		HL	Hq
<i>yang</i> ‘low’	LHL		LH	LHq

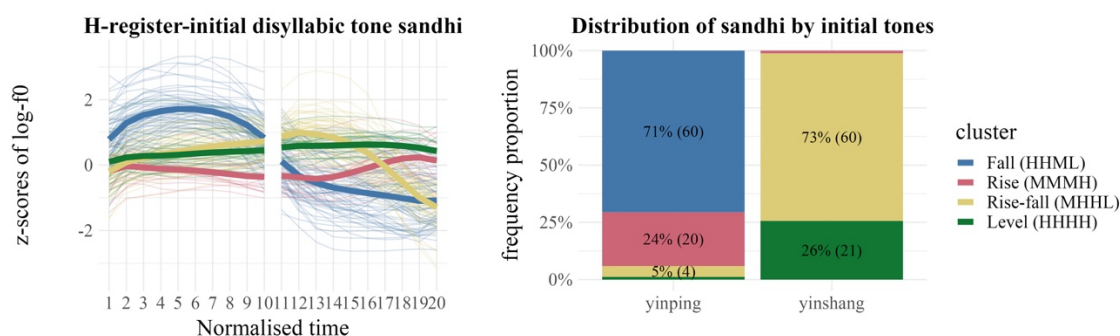


Figure 1 H-register-initial disyllabic tone sandhi patterns (left) and distributions of the patterns by initial tones (right).

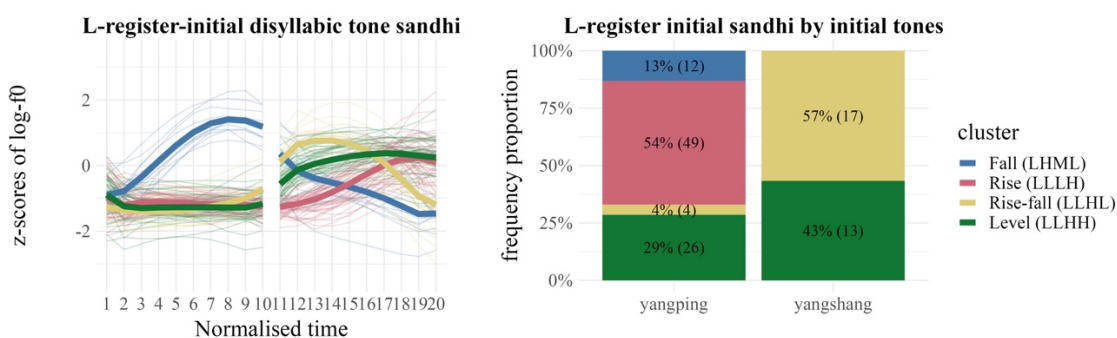


Figure 2 L-register-initial disyllabic tone sandhi patterns (left) and distributions of the patterns by initial tones (right).

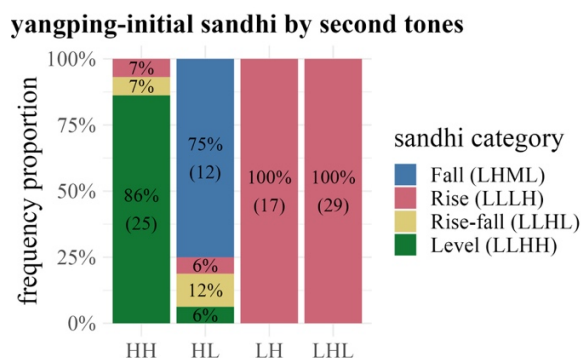


Figure 3 Distribution of yangping-initial sandhi by second tones.

References:

- [1] B. Xu, Z. Tang, and N. Qian, ‘Xinpai Shanghai fangyan de liandu biandiao [Tone sandhi in New Shanghai]’, *Fangyan Dialects*, no. 2, pp. 145–155, 1981.
- [2] X. Li, ‘Tone Sandhi In The Suzhou Dialect: Synchronic And Diachronic Perspectives’, *J. Chin. Linguist. Monogr. Ser.*, vol. 24, pp. 21–32, 2010.

- [3] N. Qian, *Dangdai Wuyu yanjiu [Studies in the contemporary Wu dialect]*. Shanghai: Shanghai educational publishing house, 1992. Accessed: Jul. 30, 2024.

Vowel-intrinsic F0 interacts with tones: Evidence from a Thai speech corpus

Sireemas Maspong^{1,2}, Teerawee Sukanchanon², Francesco Burroni^{1,2}

¹*Institute for Phonetics and Speech Processing, Spoken Language Processing Group, LMU Munich, Germany*

²*Center of Excellence in Southeast Asian Linguistics, Faculty of Arts, Chulalongkorn University, Thailand*

One of the most universal phenomena observed in human languages is vowel intrinsic F0 (IF0). This refers to the observation that, across both tonal and non-tonal languages, high vowels consistently exhibit higher F0 than low vowels. Crucially, the strength of the effect is argued to vary in a gradient manner with tone: it is most pronounced in high-F0 tones, moderate in mid-F0 tones, and diminished or neutralized in low-F0 tones [1]. However, previous studies on African tonal languages have yielded mixed results regarding the gradient interaction of tone and IF0. Some languages display a gradient interaction of tone and IF0 [2], while others show a categorical pattern, where IF0 is present for non-low tones with the similar effect sizes and absent for low tones [3]. In this paper, we investigate the interactions of IF0 and tone using (Central) Thai as a case study. Thai is an ideal language for testing the gradience of IF0, as it has five tonal categories, with F0 height at the vowel midpoint stratified into three levels in the following order: Tone 3 (High-Falling) > Tone 4 (Mid-Rising), Tone 1 (Mid) > Tone 5 (Low-Rising), Tone 2 (Low) (Figure 1). We examined the gradience of IF0 through two approaches. First, we analyzed whether vowel height, as a categorical variable (high vs. low vowels), correlates with vowel F0 across all tones. Second, we investigated whether F1, as a continuous measure of vowel height, correlates with F0 and whether this correlation varies by tone. Our findings support the hypothesis that IF0 exhibits a gradient effect by tone, both in categorical and continuous terms. However, unexpected patterns suggest that the gradient nature of IF0 is influenced by factors beyond raw F0 values of tones.

Methods. Acoustic data were obtained from the ThaiMIT speech corpus [4]. A total of 11,198 tokens were extracted, consisting of open-syllable monosyllabic words with long vowels from all tonal categories. Mean F1 and F0 values were measured at 10 ms intervals at the midpoint of the vowel intervals. For the categorical analysis, we subsetting the data to include tokens with high vowels (/i:/, u:/, ɯ:/) and low vowels (/ɛ:/, a:/, ɔ:/). A linear regression model was fitted with F0 as the dependent variable and vowel categories, tones, and their interaction as fixed effects. To examine the continuous correlation between vowel height and F0, all extracted tokens were included, and an additional linear mixed-effects regression analysis was performed. In this model, F1 replaced vowel categories as a fixed effect variable, while the model structure remained unchanged. Random slopes for subjects were included in both models.

Results. Both categorical and continuous analyses reveal that IF0 is present in nearly all tonal categories, except for Tone 2. Tone 3, with the highest F0 values, exhibits the strongest IF0 effect, followed by Tone 1 and Tone 5 with moderate effects, and Tone 4 with a smaller effect size. In contrast, Tone 2 shows no significant correlation between vowel height or F1 and vowel F0. The observed order of effect sizes is: T3 > T1, T5 > T4 > T2 (Figures 2 and 3).

Discussion. Our findings support the hypothesis that IF0 varies gradiently with F0 levels rather than categorically (e.g., non-low vs. low tones). IF0 demonstrates the strongest effects for high-F0 tones (e.g., T3), moderate effects for mid-F0 tones (e.g., T1), and neutralization for low-F0 tones (e.g., T2). However, unexpected patterns also emerged. Tone 4, despite its mid-F0 values, aligned more closely with T2 (low F0), while T5, a low-F0 tone, patterned with T1 (mid F0). These results suggest that factors contributing to IF0 extend beyond raw F0 values. Specifically, laryngeal targets and the timing of laryngeal specification may play critical roles. For instance, T5's alignment with T3 could result from a specific laryngeal setting absent in T2, which facilitates the rising F0 trajectory of T5 toward the end of the tone, even though its midpoint F0 remains low. Future research could explore contour-based analyses of IF0 or investigate the interactions between articulation—particularly laryngeal specifications and timing—and IF0 effects.

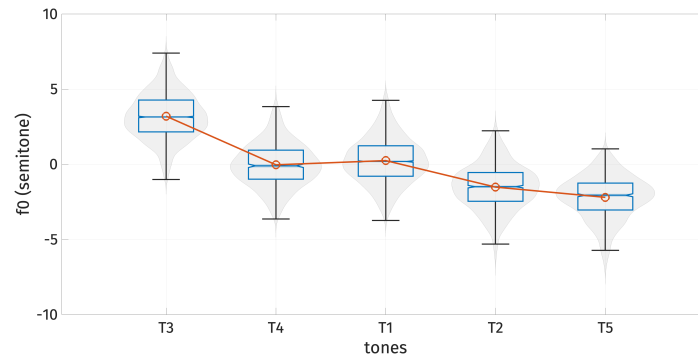


Figure 1 Mean F0 values for each tone.

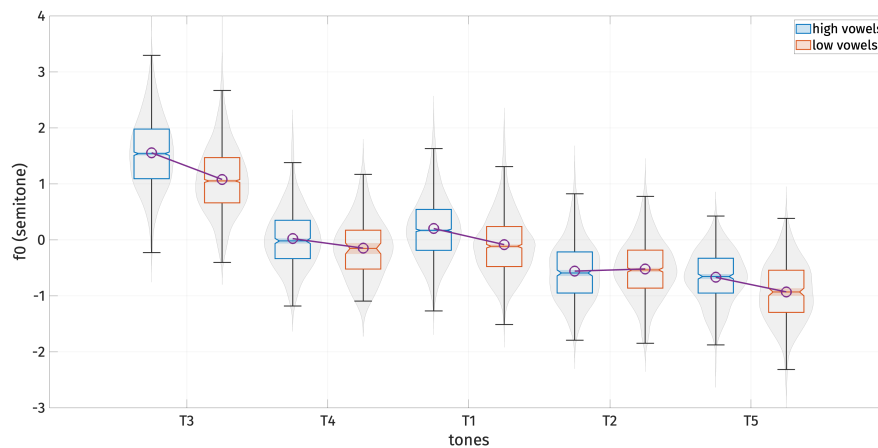


Figure 2 Mean F0 values for each vowel category by tone.

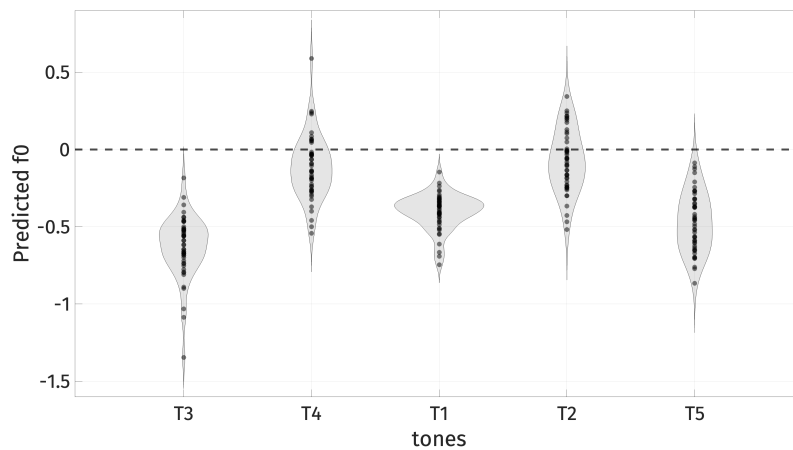


Figure 3 Predicted F0 change for each speaker, when F1 increases by one standard deviation (that is when vowels become lower).

Negative values represent negative correlation of F0 and F1 (Lower F0 correlates with lower vowels).

References

- [1] Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics*, 23(3), 349–366.
- [2] Hombert, J. M. (1977). Consonant types, vowel height and tone in Yoruba. *Studies in African Linguistics*, 8(2), 173–190.
- [3] Connell, B. (2002). Tone languages and the universality of intrinsic F0: Evidence from Africa. *Journal of Phonetics*, 30(1), 101–129.
- [4] Chanchaochai, N., Cieri, C., Debrah, J., Ding, H., Jiang, Y., Liao, S., Liberman, M., Wright, J., Yuan, J., Zhan, J., & Zhan, Y. (2018). GlobalTIMIT: Acoustic-Phonetic Datasets for the World's Languages. *Interspeech 2018*, 192–196.

The Role of Native Phonotactics in Tone Adaptation: Evidence from Nuosu Yi

Yao Zhang

Cornell University, Department of Linguistics

Keywords: Nuosu Yi, loanwords, tone adaptation, native markedness restriction

Introduction: While previous work on segmental adaptation supports the strata-indexed faithfulness model where “the less nativized an item is, the more it disobeys lexical constraints” [1], suprasegmental adaptation in tonal languages can ignore the input prominence and assign tone based on a default mechanism motivated by the native phonology [2]. This study investigates the tonal adaptation of Mandarin disyllabic loanwords into Nuosu Yi, a Sino-Tibetan language spoken in Sichuan, China. Overall, the study shows that tone adaptation in Yi is 1) based on the borrowing of Sichuan Mandarin rather than Standard Mandarin, and 2) driven by both faithfulness to the source tone and the markedness constraints of the native Yi lexicon, the latter notably contributing more. The finding not only offers a counter example to violability in lexical stratification with the fact that tone adaptation can amplify the markedness constraints inherent to the recipient language, but also reveal the perceptual complexity and uniqueness of tonal system of East Asian languages that “impedes the direct equation of input prominence with the native tonal categories [2]”.

Data and Method: My corpus study is based on 462 disyllabic loanwords [3]. Tab.1 shows the tonal adaptation patterns sorted by frequency from high to low. The tonal adaptation patterns can be better explained when the source dialect is considered to be Sichuan Mandarin rather than Standard Mandarin [4]. A multinomial logistic regression model is built to examine the source dialect as well as other predictors. A MaxEnt model, containing i) constraints that enforce faithfulness to Sichuan Mandarin tones and ii) markedness constraints on each of native Yi tones (e.g. *44 penalizes all surface occurrences of 44), is built to examine the contribution of source factors and their interaction with native markedness.

Results: AIC and BIC scores of the regression models show that Sichuan Mandarin (1561.382, 1604.821) is more likely to be the source dialect than Standard Mandarin (1607.651, 1604.821). However, faithfulness to Sichuan Mandarin does not explain the avoidance of 44 and 55, as Sichuan 45 and 52 also tend to be loaned as Yi 33 and 21 rather than the phonetically closer 44 or 55. The MaxEnt model demonstrates that the underrepresentation of 44 and 55 is likely an effect of native Yi tonotactics. Even though the high weights of IDENT(falling/mid/low) suggest Nuosu Yi prefers to preserve the falling contour and non-high pitch level of source tone, the high weight on *44 and *55 crucially suggests that avoidance of high tones is the result of markedness rather than faithfulness constraints (Tab.2). In other words, the low frequency of 44 and 55 in Yi loanwords (Fig.1a) results from their surface markedness in native Yi tonology. The markedness of 44 and 55 is supported by corpus results [4]. Fig.1b shows the relative frequency of each tone in native Yi lexicon, likely explaining the tendency for Mandarin high tones to adapt to Yi 33, as it is the less marked tone while 44 and 55 are more marked [5][6]. While avoidance of 44 and 55 can be explained by Yi tonotactic constraints, loanwords still avoid 44 and 55 more than would be expected given the native Yi lexicon. The gap between the type frequency of 44, 55 and 21, 33 in loanwords is much larger than that of native Yi, which differs from the segmental loan stratum where more lexical constraints can be disobeyed.

Conclusion: This study shows that tonal adaptation of Sichuan Mandarin loanwords into Nuosu Yi is driven by a balance of faithfulness to source tones and native tonal constraints. While some preservation of source tones occurs, the dominant influence of Nuosu Yi’s tonal markedness, avoidance of 44 and 55, underscores the primacy of native phonological structure. These results challenge the strata-indexed faithfulness model, demonstrating that tonal adaptation amplifies recipient-language markedness constraints rather than simply reflecting input prominence. Furthermore, the study highlights the unique perceptual and structural complexities of East Asian tonal systems which resist direct mappings between source and target tonal categories, thus contributing to a deeper understanding of how native phonological markedness shapes suprasegmental adaptation in loanwords.

Standard Mand.	Sichuan Mand.	Nuosu Yi	Count	Example
55	33	33	124	安培 an ⁵⁵ p ^h ei ³⁵ → ŋa ³³ p ^h i ²¹
		21	10	微分 wei ⁵⁵ fən ⁵⁵ → wo ²¹ fi ³³
		55	2	番茄 fan ⁵⁵ tɛ ^h iɛ ³⁵ → fa ⁵⁵ tɛ ^h i ²¹
		44	1	包谷 pao ⁵⁵ ku ²¹⁴ → pu ⁴⁴ kɿ ²¹
35	52	21	132	蚕豆 ts ^h an ³⁵ tou ⁵¹ → ts ^h a ²¹ tɿ ⁵⁵
		33	6	摩尔 muo ³⁵ ʌ ²¹⁴ → mo ³³ lɿ ²¹
214	45	33	40	海鸥 hai ²¹⁴ ou ⁵⁵ → he ³³ o ³³
		21	31	板凳 pan ²¹⁴ tən ⁵¹ → po ²¹ ti ³³
		44	2	党员 taŋ ²¹⁴ yan ³⁵ → ta ⁴⁴ yuo ²¹
		55	1	比丘 pi ²¹⁴ tɛ ^h iəu ⁵⁵ → pi ⁵⁵ tɛ ^h o ³³
51	11	21	80	大卡 ta ⁵¹ k ^h ɑ ²¹⁴ → ta ²¹ k ^h ɑ ³³
		33	20	道观 tao ⁵¹ kuan ⁵¹ → tɔ ³³ kɔ ³³
		55	13	大门 ta ⁵¹ mən ³⁵ → ta ⁵⁵ mo ³³

Faithfulness C	Weight	Δ log-likelihood
IDENT(falling)	3.12	54.36
IDENT(level)	0.47	0.78
IDENT(rising)	0	0
IDENT(mid)	1.63	15.17
IDENT(low)	1.60	12.13
IDENT(high)	0	0
Markedness C	Weight	Δ log-likelihood
*44	4.05	67.22
*55	1.48	25.06
*21	0.13	0.16
*33	0.00	0.00

Full model log-likelihood: -358.38

Table 2 Result of Maxent model. The weights of faithfulness constraints suggest that Nuosu Yi cares most about the preservation of the falling contour and least of the rising contour and high pitch; the weights of markedness constraints suggest that 44 is the most marked tone in Yi while 33 is the least.

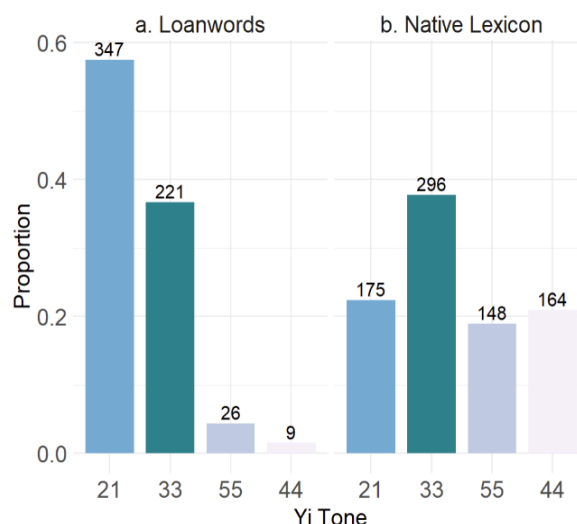


Figure 1 Type frequency of Yi tones in (a) loanwords and (b) native lexicon. In native Yi, 33 accounts for approximately 37.8% of the 783 attested syllables while the rest three tones are roughly evenly distributed. In loanwords, 21 and 33 are overwhelmingly frequent than 44 and 55.

Selected References: [1] Itô, J., & Mester, A. 2017. The phonological lexicon. [2] Kang, Y. 2010. Tutorial overview: Suprasegmental adaptation in loanwords. [3] Azimo, xiaoying. 2019. [A study on Chinese Loanwords in Liangshan Yi Language]. [4] Pan, Zhengyun.1990. [Preliminary Exploration of the Loanword Norms for Modern Liangshan Yi Language]. [5] Schilken, D., Walters, S. G., & Walters, D. 2012. *A Preliminary Study of Nuosu Yi Syllable Frequency in Text*. [6] Lama, ziwo. 1991. [A Preliminary Discussion on the Causes of the Emergence of the Second-High Tone in Yi Language].

Xin Li¹, Lesya Ganushchak², Xiaoqing Li^{3,4}, Yiya Chen^{1,5}

¹Leiden University Centre for Linguistics, ²Department of Psychology, Education, and Child Studies, Erasmus School of Social and Behavioural Sciences, Erasmus University Rotterdam, ³State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences,

⁴Department of Psychology, University of Chinese Academy of Sciences, ⁵Leiden Institute for Brain and Cognition, Leiden University

In connected speech, both segments and suprasegmental features, such as lexical tones, can undergo systematic changes depending on phonological context. Such contextual pronunciation variation requires speakers to retrieve and produce the appropriate word forms, a process not fully accounted for by current speech production models [1]. This study examines a specific type of tonal contextual variation in Standard Chinese, namely, low Tone (T3) sandhi, where T3 before another T3 is realized with a rising pitch contour (hereafter T3V in contrast to T3 in T3TX). Broadly speaking, two mechanisms have been proposed to explain the production of T3V: the storage vs. the computation mechanism. The storage mechanism posits that T3V and T3 are co-stored in the mental lexicon under the general T3 category, allowing direct retrieval or access of T3V [2][3], which would imply similar processing steps and time for T3 in disyllabic T3T3 vs. T3TX (X refers to non-T3 tones in Standard Chinese) words. In contrast, the computation mechanism proposes a single representation of canonical T3 with T3V (or T2, as often described in the literature) derived through an extra ‘rule-based’ operation that transforms T3 into T3V, thereby requiring an additional processing step and consequently, longer processing time for disyllabic T3 sandhi words than for non-sandhi words. Evidence supporting the computation account comes from specific studies that report a lack of word frequency effects in T3 sandhi production [4][5] as well as increased processing effort compared to non-sandhi words [4]. However, these findings also seem to diverge in terms of the specific level (or time course) at which T3 sandhi is presumed to occur —phonological, phonetic, or articulatory [4][5][6].

We utilized three picture-naming tasks to investigate the locus of T3 sandhi change during speech planning, aiming to gain further insights into the production stages underlying T3 contextual variation. In the **immediate naming (IN) task**, participants named pictures as quickly as possible upon presentation. In the **delayed naming (DN) task**, participants named pictures upon a naming cue, which occurred with a delay after picture presentation. The **delayed naming task with articulatory suppression (SN)** is a variant of the task that requires participants to repeatedly articulate an unrelated syllable (to the target picture name) during the delay. IN reflects the natural speech production process, encompassing phonological encoding, phonetic encoding, and articulation [7]. Latencies in DN primarily reflect the processing needed during the articulation stage, as the target’s information in the pre-motor phase is buffered and readily available for motor execution upon the cue [7]. Naming latencies in SN shed light on processing at the phonetic and articulatory levels, as pre-phonetic information is buffered and ready for subsequent stages when the naming cue appears [7]. Furthermore, we considered the frequency with which a particular T3 morpheme appears with its T3 variant, referred to as the T3 sandhi rate, defined as the proportion of their occurrences in a sandhi context preceding another T3 as [8] has reported its relevance for a lexical decision task. The results of all tasks, along with the possible frequency-modulating effect of T3 sandhi rate, collectively help identify the specific stage at which T3 sandhi may occur, and when frequency effects may matter, assuming that the sandhi production involves three distinct stages of planning.

We recruited 35 native Beijing Mandarin speakers who named 21 pairs of targets: disyllabic T3 sandhi words (e.g., T3T3, 脚掌, *jiao3zhang3*, ‘sole of feet’) and T3 non-sandhi words (e.g., T3TX, 绞架, *jiao3jia4*, ‘gallows’). The T3 sandhi and non-sandhi conditions were matched in whole word frequency, initial and final morpheme frequency, and syllable frequency.

Results from linear mixed effect modelling revealed a main effect of task: naming responses were significantly slower in the IN task than in both the SN and DN tasks. However, no significant differences in naming latencies were observed between T3T3 and T3TX words across the three tasks, nor was there any significant interaction (Figure 1). The lack of their difference may be taken as evidence that they were processed similarly throughout the planning stages, which, however, does not allow us to determine whether both T3T3 and T3TX undergo computation, or both are stored and directly retrieved without computation. Interestingly, we observed a very subtle but significant main effect of the sandhi rate for both T3T3 and T3TX target words across tasks, with slower naming latencies as the sandhi rate increased ($\beta = 0.0196$, $SE = 0.0091$, $t = 2.145$, $p = 0.0379$). Such an inhibitory effect suggests possible competition between the T3 and T3V forms in processing disyllabic T3-initial words. As the sandhi rate increases, the strength of the T3V representation may intensify, competing with the T3 representation during production. Furthermore, we also did an exploratory analysis and fitted separate models for each of the three tasks. Results revealed a significant sandhi rate effect only in the SN task ($\beta = 0.03372$, $SE = 0.01261$, $t = 2.675$, $p = 0.0106$), potentially reflecting competition tied to the retrieval of T3 or T3V during phonetic planning. While we emphasize the importance of replicating the reported results, these findings suggest T3 and T3V selection likely occurs during phonetic encoding, without invoking transformation-based computation, but rather reflecting direct competition between stored variants.

Keywords: Speech production; T3 sandhi words; T3 non-sandhi words; picture naming

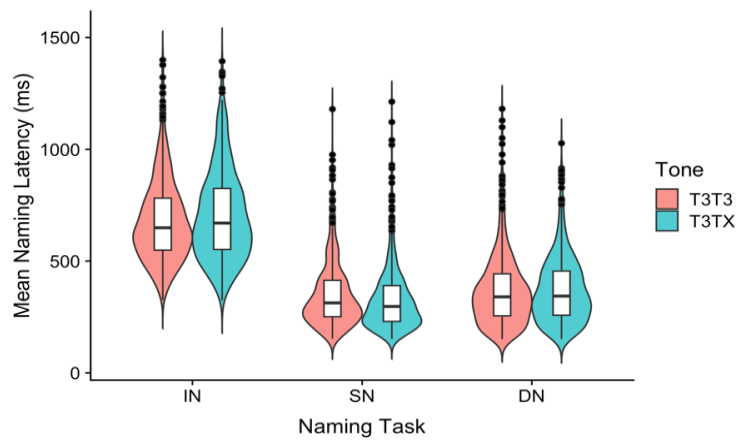


Figure 1: Violin plot of naming latencies (in ms) for each combination of tone and task.

Reference

- [1] Bürki, A. 2018. Variation in the speech signal as a window into the cognitive architecture of language production. *Psychonomic Bulletin and Review* 25 (6), 1973–2004
- [2] Li, X., & Chen, Y. 2015. Representation and Processing of Lexical Tone and Tonal Variants: Evidence from the Mismatch Negativity. *PLOS ONE*, 10(12), e0143097.
- [3] Nixon, J. S., Chen, Y., & Schiller, N. O. 2015. Multi-level processing of phonetic variants in speech production and visual word processing: evidence from Mandarin lexical tones. *Language, Cognition and Neuroscience* 30(5), 491–505.
- [4] Zhang, C., Xia, Q., & Peng, G. 2015. Mandarin third tone sandhi requires more effortful phonological encoding in speech production: Evidence from an ERP study. *Journal of Neurolinguistics* 33, 149–162.
- [5] Zhang, J., Zhang, C., Politzer-Ahles, S., Pan, Z., Huang, X., Wang, C., Peng, G., & Zeng, Y. 2022. The neural encoding of productive phonological alternation in speech production: Evidence from Mandarin Tone 3 sandhi. *Journal of Neurolinguistics*, 62, 101060.
- [6] Chen, X., Zhang, C., Chen, Y., Politzer-Ahles, S., Zeng, Y., & Zhang, J. 2022. Encoding category-level and context-specific phonological information at different stages: An EEG study of Mandarin third-tone sandhi word production. *Neuropsychologia*, 175, 108367.
- [7] Laganaro, M., & Alario, F.-X. 2006. On the locus of the syllable frequency effect in speech production. *Journal of Memory and Language*, 55(2), 178–196.
- [8] Gao, F., & Lin, C.-J. C. 2024. Incorporating Frequency Effects in the Lexical Access of Mandarin Tone 3 Sandhi. *Language and Speech*, 00238309241260062.

Variability in caregiver productions and turn-timing in 1-to-2-year-olds

Bettina Braun, Marieke Einfeldt, Sarah Warchhold
Department of Linguistics, University of Konstanz, Germany

Infant-directed speech has been shown to have shorter utterances, higher and more variable pitch (Fernald et al., 1989) which affects the child's language development (Golinkoff et al., 2015). In this project, we analyse whether caregiver's productions (sentence types, prosody) impact turn-timing during joint picture book viewing, a frequent activity between caregivers and young children. Previous research shows longer gap durations in transitions to a child (median of 900-1200ms at 9-12-months (mo) and 500-600ms at 3-6mo and at 18mo) than to the mother (median of 400-600ms, independent of child's age), cf. Hilbrink et al., (2015). A case study showed that non-repeat verbal responses to where- and which-questions took around 1s between 1;4 to 3;5 (Clark & Lindsey, 2015). In an eye-tracking study, 2.5-year-olds anticipated turn-transitions earlier following questions than declaratives, but prosody played no role (Lammertink et al., 2015); in contrast, a flattened intonation reduced anticipation of turn-transitions in 3-year-olds (Keitel & Daum, 2015). Given this discrepancy, we investigate factors that affect the silent gaps during natural interactions, focusing on age, sentence type (interrogative, declarative, imperative, but also labels and tags) and prosody (boundary tones, nuclear tunes, variability in tonal events). We predict that children react more frequently and quickly after interrogatives than other sentence types and after final rises than after falls.

We collected audiovisual recordings of 50 caregiver-child dyads with children between 12 and 24 months. So far, we analyzed 21 dyads (mean age_{child}=19.5 months, SD=3.5). The data was collected during remote testing in the families' homes via iPads. The child sat in front of the tablet and the caregiver described a colorful picture (containing 3 persons and 6 animals) to the child for one minute. Recordings were segmented and categorized into children's utterances (separated by pauses), caregivers' utterances (complete nuclear tunes) as well as silences and (breathing) noises (1554 events). Overlaps were ignored. The caregiver's productions were coded for sentence type (e.g., declarative, label, interrogative, question tag (e.g., *hm?*)) and annotated using GToBI (Grice et al., 2005). Diacritics (* or +) were removed to operationalize tonal variability (Carcassi et al., 2019) for the entire utterance and the final two tones. Caregivers often (n>10) used high-rising *wh*-questions and falling declaratives to children aged ≤18mo and high-rising interrogatives, falling or rising labels, falling affirmations, and low-rising polar questions to children aged >18mo. Since age, sentence type, and prosody are not independent, we modeled responder (caregiver or child) and gap duration separately for syntactic and prosodic factors. We only included cases in which there were more than 4 instances per sentence type or boundary tone in the caregivers' turn. Increasing age positively influenced child (vs. caregiver) response (p=0.04), but not gap durations (p=0.09). Children responded more often after interrogatives than declaratives (p=0.001, Tukey-corrected) or labels (p=0.05). Gap durations were also shorter after interrogatives (580ms) than labels (1050ms, p=0.02). Children responded more often after high rises (H-^H%) than low tunes (L*L-%, p<0.01) and falls (H*L-%, p<0.01). Gap durations were however shorter following low rises (L*L-%/L-H%, av. 490ms) than low tunes (L*L-%, av. 1230ms, p=0.05), see Fig. 1. GToBI boundary tones and global tonal entropy (Carcassi et al., 2019) had no effects.

This suggests that in natural interactions, children react more often and more quickly after interrogatives than labels and more often to high-rises than low tunes/falls. Furthermore, latency was shorter after mid rising tunes. We plan to broaden the sample of caregiver-child dyads and sentence types, and include gestures to determine non-verbal reactions.

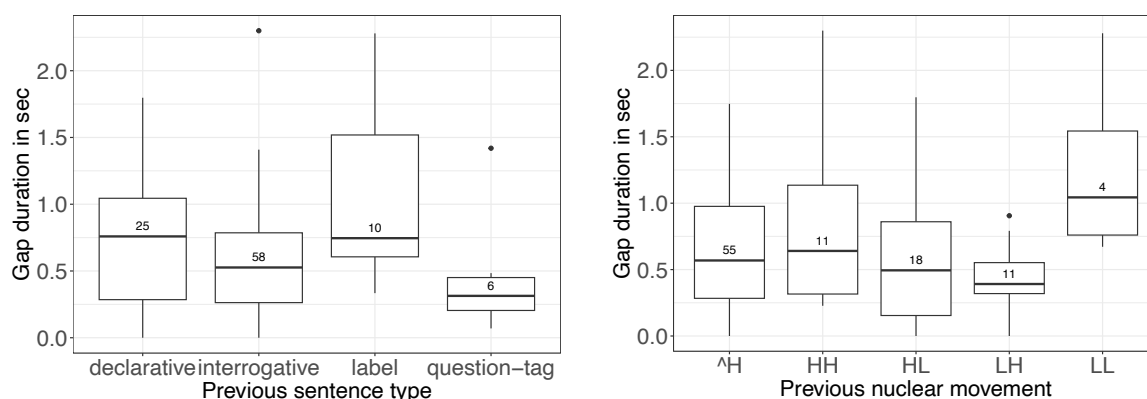


Fig. 1. Gap duration before child response as a function of sentence type and variability in nuclear tune of caregiver's preceding turn (^H indicates a high rise, HH a high-ending nuclear accent followed by a high plateau, HL a high-ending nuclear accent followed by a low edge tone, LH a low-ending accent followed by a high plateau or a low rise, LL a low-ending nuclear accent with a low edge tone). Number of child responses are added above the median.

References

- Carcassi, G., Aidala, C. A., & Barbour, J. (2019). *Variability as a better characterization of Shannon entropy*. <https://doi.org/10.48550/ARXIV.1912.02012>
- Clark, E. V., & Lindsey, K. L. (2015). Turn-taking: A case study of early gesture and word use in answering WHERE and WHICH questions. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00890>
- Fernald, A., Taeschner, T., Dunn, J., Papousek, M., & De Boysson-Bardies, B. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16, 477–501.
- Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)Talk to Me: The Social Context of Infant-Directed Speech and Its Effects on Early Language Acquisition. *Current Directions in Psychological Science*, 24(5), 339–344. <https://doi.org/10.1177/0963721415595345>
- Grice, M., Baumann, S., & Benz Müller, R. (2005). German Intonation in Autosegmental-Metrical Phonology. In J. Sun-Ah (Ed.), *Prosodic Typology. The Phonology of Intonation and Phrasing* (pp. 55–83). Oxford University Press.
- Hilbrink, E. E., Gattis, M., & Levinson, S. C. (2015). Early developmental changes in the timing of turn-taking: A longitudinal study of mother–infant interaction. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01492>
- Keitel, A., & Daum, M. M. (2015). The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00108>
- Lammertink, I., Casillas, M., Benders, T., Post, B., & Fikkert, P. (2015). Dutch and English toddlers' use of linguistic cues in predicting upcoming turn transitions. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00495>

Parallels with Emotional Prosody: Affective Iconicity of Lexical Tones in Standard Chinese

Tingting Zheng^{1,2}, Clara C. Levelt^{1,2}, Yiya Chen^{1,2}

¹Leiden University Centre for Linguistics, Leiden University;

²Leiden Institute for Brain and Cognition, Leiden University.

t.zheng@hum.leidenuniv.nl, c.c.levelt@hum.leidenuniv.nl, yiya.chen@hum.leidenuniv.nl

Emotional prosody is the primary prosodic channel for conveying affect in speech, but not the only one. Lexical tone, primarily used to distinguish word meanings in tone languages like Standard Chinese (SC), also carries affective meaning. Recent studies show that pitch variations in SC tones can signal emotional valence and arousal of general lexicon (Yao et al., 2013; Yap et al., 2014; Zheng et al., 2024, 2025), similar to affective phoneme cues in Indo-European languages—a phenomenon known as affective iconicity or emotional sound symbolism (Adelman et al., 2018; Aryani et al., 2018; Louwerse & Qu, 2017).

Despite the proposed link between tonemes and affective meaning in tonal languages like SC, it is important to note that existing studies mainly rely on written text. As a result, it remains to be empirically verified whether and how tonemes influence emotional responses in auditory contexts when semantic content is absent. To address this gap, the present study investigated how lexical tones in spoken nonce words affect raters' perception of their emotional valence and arousal.

The stimuli included four syllables ([til], [tul], [nil], [nul]) with the four SC lexical tones: high-level tone (T1, H), rising tone (T2, R), low(-dipping) tone (T3, L), and falling tone (T4, F), produced without emphasis by a male native SC speaker who was unfamiliar with the research purpose. As syllable-final [l] does not occur in SC, all syllables were therefore perceived as nonce words. Figure 1 displays the pitch contours of all sixteen tokens. Native SC listeners were asked to rate the arousal (high vs. low arousal) and valence (negative vs. positive valence) in a two-alternative forced-choice task. Each of the sixteen tokens was presented pseudo randomly in two separate blocks—one for arousal rating and one for valence rating—on the Gorilla platform (Anwyl-Irvine et al., 2020). The final analysis included 121 participants for the arousal task (age 21.53 ± 2.11 years; 72 females) and 135 participants for the valence task (age 21.52 ± 2.13 years; 84 females).

Generalized linear mixed models (logistic regression) were utilized for analyses using the *lme4* package (Bates et al., 2014). Results showed that lexical tone significantly predicted participants' choices of emotional arousal and valence (arousal: $\chi^2 = 91.33$, $p < .001$, $R^2 = 3.46\%$; valence: $\chi^2 = 235.78$, $p < .001$, $R^2 = 7.93\%$). Pairwise comparisons with Bonferroni corrections indicated that the T4 syllables were 2.35 times more likely than the T1 syllables, and 3.10 times more likely than the T3 syllables to be associated with high arousal choice. Similarly, the T2 syllables were 1.85 more likely than the T1 syllables and 2.44 times more likely than the T3 syllables to be associated with high arousal. However, there was no significant difference between T2 and T4 syllables. Regarding valence, syllables with T4 were 3.99 and 4.08 times more likely to be rated as negative than T1 and T2, respectively. Similarly, T3 syllables were 4.00 and 4.08 times more likely to receive negative ratings than T1 and T2 syllables, respectively. But there was no difference between T4 and T3 syllables. The frequency distribution of emotional arousal and valence choices of the raters for each tone was illustrated in Figure 2.

Our findings thus confirm that lexical tones significantly influence emotional responses in the auditory context. The relatively high arousal rating of T4- and T2-syllables is likely due to their dynamic pitch variations (reflected in wider pitch range and steeper contour). The association of T1 and T3 with lower arousal rating reflects that they have static high and low tone targets (despite the fact that T3 is often produced with a rising tail in isolation). Moreover, both the rising contour of T2 and the high pitch of T1 increase the likelihood of positive-valence perception. These results are compatible with findings on the characteristic pitch patterns reported for emotional prosody where higher pitch, wider pitch range, and steeper pitch slope are associated with high arousal; and higher pitch level and a rising contour is linked to positive valence (e.g., Bänziger & Scherer, 2005; Frick, 1985; Laukka et al., 2005; Yap et al., 2014). These parallel associations between pitch variation and emotional valence/arousal in tonemes and emotional prosody suggest the inherent pitch iconic quality in lexical tones and their potential to subtly shape speakers' emotional experiences, further highlighting the universal use of pitch in emotional perception.

Keywords: lexical tone, emotional arousal, emotional valence, Standard Chinese, affective iconicity

Figure 1

The pitch contours of auditory stimuli.

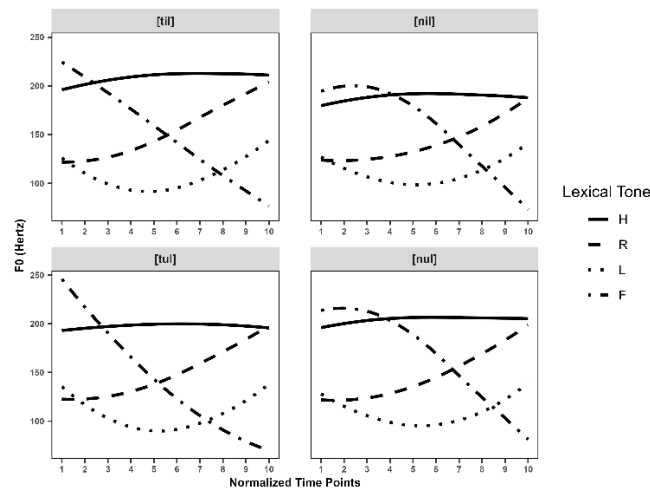
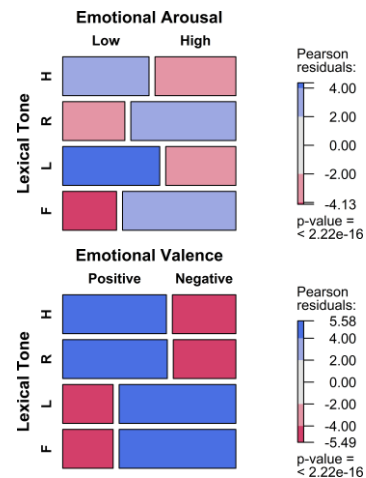


Figure 2

Frequency variation of arousal and valence.



References

- Adelman, J. S., Estes, Z., & Cossu, M. (2018). Emotional sound symbolism: Languages rapidly signal valence via phonemes. *Cognition*, 175, 122–130. <https://doi.org/10.1016/j.cognition.2018.02.007>
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1), 388–407. <https://doi.org/10.3758/s13428-019-01237-x>
- Aryani, A., Conrad, M., Schmidtke, D., & Jacobs, A. (2018). Why ‘piss’ is ruder than ‘pee’? The role of sound in affective meaning making. *PloS One*, 13(6), e0198430. <https://doi.org/10.1371/journal.pone.0198430>
- Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46(3–4), 252–267. <https://doi.org/10.1016/j.specom.2005.02.016>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *arXiv:1406.5823 [Stat]*. <http://arxiv.org/abs/1406.5823>
- Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97(3), 412–429. <https://doi.org/10.1037/0033-2909.97.3.412>
- Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, 19(5), 633–653. <https://doi.org/10.1080/02699930441000445>
- Louwerse, M., & Qu, Z. (2017). Estimating valence from the sound of a word: Computational, experimental, and cross-linguistic evidence. *Psychonomic Bulletin & Review*, 24(3), 849–855. <https://doi.org/10.3758/s13423-016-1142-2>
- Yao, Y., Lin, J., & Huang, C.-R. (2013). Lexicalized emotion? Tonal patterns of emotion words in Mandarin Chinese. *PolyU Institutional Repository*.
- Yap, D., Casasanto, L. S., & Casasanto, D. (2014). *Metaphoric Iconicity in Signed and Spoken Languages*.
- Zheng, T., Levelt, C. C., & Chen, Y. (2024). The Adaptive Value of Mandarin Tones for Affective Iconicity. *Proc. Speech Prosody 2024*, 562–566. <https://doi.org/10.21437/SpeechProsody.2024-114>
- Zheng, T., Levelt, C. C., & Chen, Y. (2025). The affective iconicity of lexical tone: Evidence from standard Chinese. *The Journal of the Acoustical Society of America*, 157(1), 396–408. <https://doi.org/10.1121/10.0034863>

Variability and incomplete sandhi neutralizations in Taiwanese checked tones

Sheng-Fu Wang (Institute of Linguistics, Academia Sinica, Taiwan)

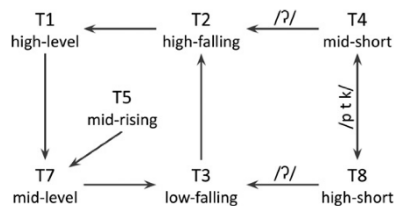
This study examines the two checked tones in Taiwanese (Taiwan Southern Min) as a case of how phonological processes result in dynamic patterns of change in progress. These tones occur in syllables with a plosive coda and are high (T8) and mid (T4) tones in base forms. In a sandhi process (tonal alternation in a nonfinal position of a “sandhi” domain), they change values following a “cycle” (Fig. 1): with /-p, -t, -k/ codas, T8 and T4 flip their values; with /-ʔ/, they change to T3 (low-falling) and T2 (high-falling) tones respectively.

Previous research has noted various aspects of variability: (1) base T8 varies across dialects, regions, and age groups, with younger speakers often realizing it as a mid-tone (Fon & Khoo, 2024). (2) T4 and T8 in base tones nearly merge in central Taiwan (e.g., Chen, 2010). (3) The loss of plosive codas due to Mandarin influence may further alter tonal values (Weng & Lee-Kim, 2023; Pan, 2017). Building on these observations, this study asks three questions. First, how widespread is the mid-base T8? Second, is the near merger of base T4 still a feature of Central Taiwan? Third, is there acoustic evidence of sandhi T4/T8’s mapping into base T3/T2?

Speech data from 33 speakers were extracted from the Taiwanese Across Taiwan corpus (Liao, 2021). Pitch measurements were extracted using ProsodyPro and transformed into semitones. The analysis included base/sandhi T4 and T8 syllables, with base T2 and T3 syllables added as references (See token counts in Tab. 1). Speakers are categorized into three major regions based on places of birth and two age groups with 30 as the threshold.

Generative Additive Models (Fig. 2) show that base T8 is mostly realized as a mid-tone in the North and Central regions, as evidenced by its contour's significant initial difference from T2. The near merger of base T4/T8 is observed in older speakers not only in the Central region, while younger speakers in the Central region are showing a clear divergence between base T4 and T8. The results also show that sandhi T4 has significantly different tonal registers from T2 especially in the later proportions of the contour, suggesting the neutralization suggested by the sandhi cycle is incomplete. The mapping of sandhi T8 into T3 is more consistently observed, as there are even cases of contours without significant differences (e.g., North older speakers). The durational analyses (Fig. 3) show that /-ʔ/ checked sandhi syllables are longer than /-p, -t, -k/ counterparts ($p < .01$ for sandhi T4, $p < .001$ for sandhi T8 except for n.s. in Central speakers and older Southern speakers) and resemble the nonchecked counterparts (e.g., T2 & T3), which potentially motivates the description of merger into those tones in the sandhi circle.

These results confirm a wider spread of some features previously noted for certain age groups (mid-base T8 for younger speakers) and regions (T4/T8 near merger) while recording some potential new developments (e.g., base T4/T8 split for Central younger speakers). Another notable finding is how checked tone sandhi is not complete neutralization as the circle suggests. The splitting behavior in sound change regarding base and sandhi tones (e.g., unmerging base forms with no apparent change in sandhi forms) has implications on lexical and phonological representations. (Keywords: Taiwanese, Southern Min, tone sandhi, checked tone, sound change)



	T4	T8	T2	T3
Base	1292	451	1972	842
Sandhi	5310	1757		

Table 1: Token distributions

Figure 1: The tone sandhi circle of mainstream Taiwanese (Fig. 4, Fon & Khoo, 2024)

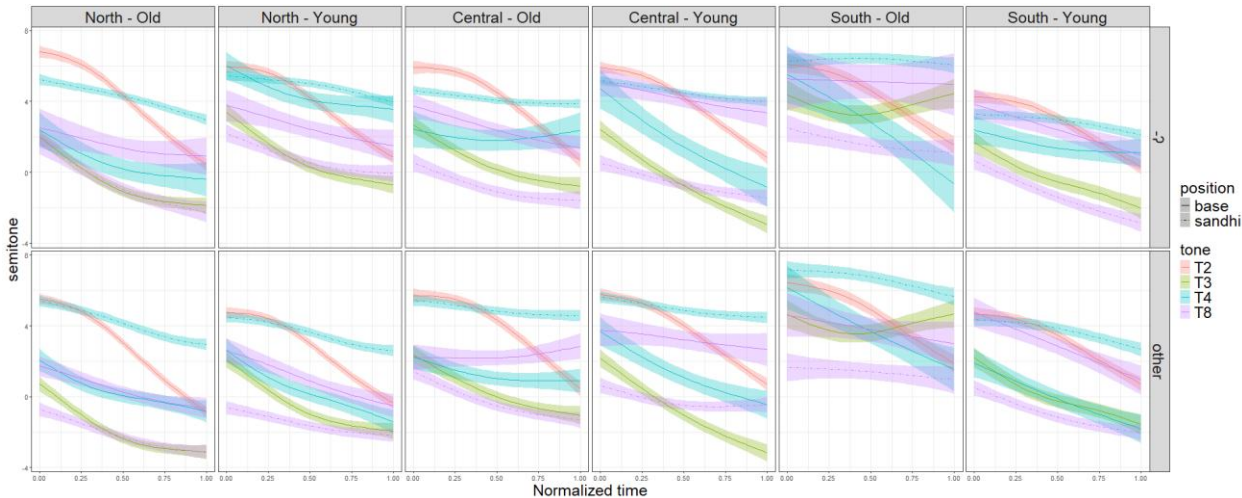


Figure 2: pitch contour as a function of tone type and region-age.

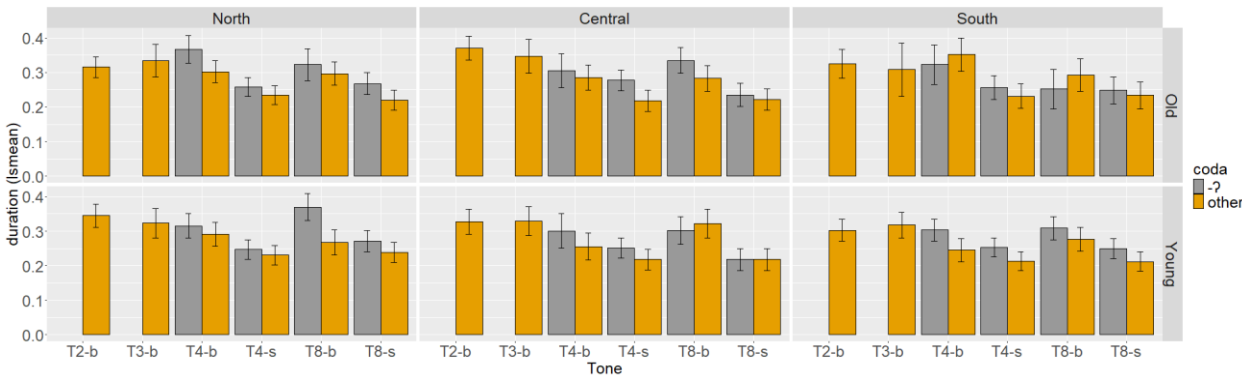


Figure 3: syllable duration as a function of tone type, age, and region

References:

Chen, S. C. (2010). New Sound Variation in Taiwan Southern Min: Vowel Systems and the Lower Register Entering Tone in Taipei, Changhua, and Tainan. *Language and Linguistics*, 11(2), 425-468.

Fon, J., & Khoo, H. L. (2025). *The Phonetics of Taiwanese*. Cambridge University Press.

Y.-F. Liao. (2022). Taiwanese Across Taiwan (TAT) Corpus [online] Available: http://www.aclclp.org.tw/use_mat_c.php#tat.

Pan, H. H. (2017). Glottalization of Taiwan Min checked tones. *Journal of the International Phonetic Association*, 47(1), 37–63.

Weng, W. C., & Lee-Kim, S. I. (2023). Loss of unreleased final stops among Mandarin-Min bilinguals: Structural convergence of languages in contact. *Journal of Phonetics*, 101, 101279.

Voicing, Vowel Height and Register in Triang, an Early-Stage Register Language

Ryan Gehrman
Payap University
ryan_g@payap.ac.th

Keywords: *voicing, vowel height, register, registrogenesis, Austroasiatic*

This study presents a detailed acoustic analysis of contrasts with registrogenetic potential in Triang, a conservatively non-registral Austroasiatic language of southern Laos. We examine the effects of onset stop voicing (voiced, voiceless) and vowel height (close, open) on acoustic properties that typically co-vary with register (f0, F1, F2, H1*-H2*, H1*-A1*, H1*-A3* and CPP).

Register contrast is a prominent feature in many Austroasiatic and Austronesian languages of Mainland Southeast Asia, typically manifesting as binary contrasts cued by combinations of differential pitch, voice quality, and vowel quality [1, 2]. The diachronic origins of these contrasts lie in the transphonologization of historical onset voicing contrasts, vowel height distinctions, or complex patterns involving both factors [3, 4]. Despite their prevalence, few phonetic studies have investigated conservatively non-registral languages (cf. [5, 6]), and no studies have specifically examined the role of vowel height in early-stage registrogenesis. It remains unclear whether vowel height conditioning constitutes an entirely separate process necessitating prior voicing-conditioned registrogenesis, as has been suggested [7], or whether both factors interact from the earliest stages.

To investigate these questions, we recorded thirteen Triang speakers (seven female) producing twenty-four words, each repeated four times in a sentence frame. The target words contained voiced and voiceless stops at two places of articulation (labial, coronal), preceding seven different vowel qualities. Acoustic measurements were extracted at 1ms intervals via *praatSauce* [8] and normalized across speakers. To assess the effects of onset voicing and vowel height on register cues, we calculated mean values for each measurement across three phases of the rime (0-20, 50-70, and 100-120 ms after stop release) and constructed linear mixed effects models with onset voicing, onset place of articulation, and vowel as fixed effects and speaker as a random effect. For vowel height analyses, we removed mid vowels and substituted vowel height (close vs. open) for vowel as a fixed effect. Two-way interactions were included when they improved model fit.

Our results, based on data from eleven speakers (six female), show that negative VOT for voiced stops was remarkably stable, with only occasional partial devoicing before release (Fig. 1). Onset voicing significantly predicted acoustic correlates of register across multiple phases of the rime in the expected directions for nearly all measures (except H1*-A1*), with particularly strong effects on pitch and vowel height (Figs. 2, 3; Table 1). Vowel height was also a significant predictor of register-associated cues, consistent with the intrinsic f0 effect and the relatively breathy/lax phonation associated with closer vowels [9]. Importantly, significant interactions between onset voicing and vowel height were observed for f0 (phases 1 and 2) and F1 (phase 1) (Figs. 4, 5), suggesting that these factors may interact synergistically in early registrogenesis.

These findings indicate that acoustic cues associated with register covary with both onset voicing and vowel height in Triang, and that distinctive patterns of variation affecting f0 and F1 immediately after stop release demonstrate phonetic differentiation of the four voicing-by-vowel height categories emerging simultaneously at an early stage of registrogenesis. This evidence challenges the view that vowel height conditioning necessarily operates as a separate, secondary process in register development and instead suggests that the interplay between onset voicing and vowel height can pertain from the initial stages, potentially explaining the complex conditioning patterns observed in certain more innovatively registral languages of the region.

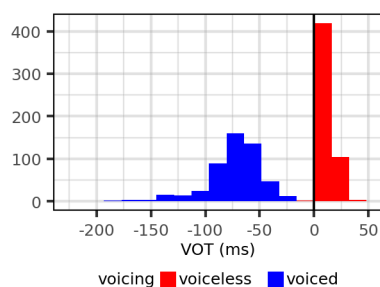


Figure 1: Voicing & VOT

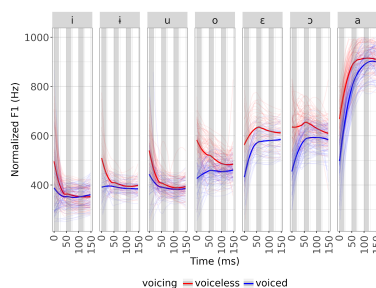


Figure 2: Voicing, vowel & F1

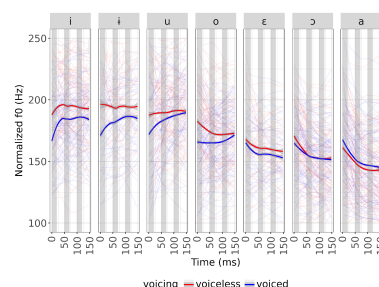


Figure 3: Voicing, vowel & f0

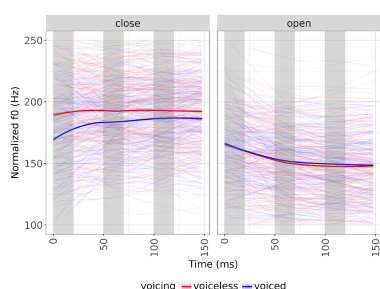


Figure 4: Voicing, height & f0

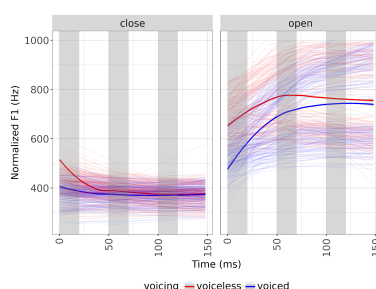


Figure 5: Voicing, height & F1

***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, n.s. = not significant

measure	predictors	1	2	3
f0	onset voicing	***	***	***
f0	vowel height	***	***	***
f0	interaction	***	n.s.	n.s.
F1	onset voicing	***	***	***
F1	vowel height	***	***	***
F1	interaction	***	*	n.s.
F2	onset voicing	***	***	n.s.
F2	vowel height	***	***	***
H1*-H2*	onset voicing	n.s.	***	n.s.
H1*-H2*	vowel height	***	***	***
H1*-A1*	onset voicing	n.s.	n.s.	n.s.
H1*-A1*	vowel height	***	***	***
H1*-A3*	onset voicing	***	***	n.s.
H1*-A3*	vowel height	n.s.	*	*
CPP	onset voicing	***	***	***
CPP	vowel height	***	n.s.	***

Table 1: significance measures

References

- [1] Brunelle, Marc & James Kirby. 2016. Tone and phonation in Southeast Asian languages. *Language & Linguistics Compass* 10:4. 157-207.
- [2] Brunelle, Marc & Thành Tấn Tạ. 2021. Register in languages of Mainland Southeast Asia: The state of the art. In Paul Sidwell & Mathias Jenny (eds.), *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*. (World of Linguistics 8). Berlin/Boston: de Gruyter Mouton.
- [3] Huffman, Franklin. 1976. The register problem in fifteen Mon-Khmer languages. In Philip N. Jenner, Laurence C. Thompson & Stanley Starosta (eds.), *Austroasiatic Studies*, 575-590. Honolulu: The University Press of Hawaii.
- [4] Gehrman, Ryan. 2022. *Desegmentalization: Towards a Common Framework for the Modeling of Tonogenesis and Registrogenesis in Mainland Southeast Asia with Case Studies from Austroasiatic*. Edinburgh: University of Edinburgh PhD thesis.
- [5] Brunelle, Marc, Jeanne Brown & Phạm Thị Thu Hà. 2022. Northern Raglai voicing and its relation to Southern Raglai register: Evidence for early stages of registrogenesis. *Phonetica* 79:2. 151-188.
- [6] Kirby, James, Pittayaporn, Pittayawat & Brunelle, Marc. 2022. Transphonologization of onset voicing: Revisiting Northern and Eastern Kmhmu'. *Phonetica* 79:6. 591-629.
- [7] Gehrman, Ryan. 2015. Vowel height and register assignment in Katuic. *Journal of the Southeast Asian Linguistic Society* 8. 56-70.
- [8] Kirby, James. 2018. praatsauce: Praat-based tools for spectral analysis. online: <https://github.com/kirbyj/praatsauce>.
- [9] Lotto, A.J., L.L. Holt & K.R. Kluender. 1997. Effect of voice quality on perceived height of English vowels. *Phonetica* 54, 76-93.

From Lexical Tones to Pitch Accents: How Taiwan Mandarin Speakers Realize Focus in L2 English

Sherry Chien¹, Argyro Katsika²

^{1,2}*University of California, Santa Barbara*

sherrychien@ucsb.edu , argyro@ucsb.edu

This study examines how Taiwan Mandarin speakers realize different types of focus in their L2 English. In Taiwan Mandarin, focus is marked by raising F0, increasing intensity, and lengthening the duration of the entire focused word, which can span more than one syllable (e.g., S. Chen et al., 2009). Lexical tones in the more well-studied Standard Chinese also exhibit enhanced F0 contours under focus compared to post-focus conditions, making the tones in the focused constituent more distinctive (e.g., Y. Chen, 2010), a pattern that may also hold for Taiwan Mandarin. In contrast, English marks focus by associating a nuclear pitch accent with the stressed syllable of the focused word (Beckman & Pierrehumbert, 1986), along with increased duration, intensity, and F0 (Cooper et al., 1985). While both systems involve pitch modulation, Mandarin tones are tied to every syllable for lexical meaning, whereas English pitch accents are linked to the stressed syllable for intonational meaning. This distinction makes English and Taiwan Mandarin an interesting L1-L2 pair, as they differ in how they use the same phonetic dimensions to mark lexical and phrasal prosody, as well as the stretch of speech (e.g., syllable, word) affected by focus-related F0 manipulation. Twelve native Taiwan Mandarin speakers participated in a question-answering experiment. Two sets of initial-stressed English words with varying word lengths were elicited phrase-medially under three focus conditions (Table 1). Vowel duration and F0 values of each syllable were extracted using VoiceSauce (Shue et al., 2011). Vowels were labeled by their position within the word: V1, V2, and V3 for the vowel in the word's first, second, and third syllable. Linear mixed-effects models were conducted to examine the effects of Word Length (monosyllabic, disyllabic, trisyllabic), Vowel Number (V1, V2, V3), and Focus (contrastive, narrow, unfocused) on four dependent variables normalized by the test words: duration, mean F0, max F0, and intensity. Results showed significant main effects of focus across all acoustic measures. Both contrastive and narrow focus led to longer duration (Figure 1), higher mean F0 (Figure 2) and max F0 (Figure 3), and higher intensity (Figure 4) compared to the unfocused condition, regardless of word length or vowel position. An interaction effect between Word Length and Focus was observed for duration: contrastive focus had more lengthening effect in shorter words than in longer ones (Figure 1). No significant differences were found between narrow and contrastive focus for any acoustic measures. Word Length and Vowel Position showed no significant main effects in any models, indicating that focus-induced effects (longer duration, higher F0 values, higher intensity) applied to the entire focused word, not just the stressed syllable (V1), which typically bears the greatest focus-related prominence in English (Beckman & Pierrehumbert, 1986). The findings suggest that Taiwan Mandarin speakers used duration, F0, and intensity to mark focus in their L2 English, with the effects extending across the entire focused word, reflecting their L1 focus prosody. Lengthening in contrastive focus was more pronounced in monosyllabic words and less so in polysyllabic words, suggesting a reduced effectiveness in applying this cue to longer words. Both F0 measures were higher in focused words compared to unfocused ones, likely reflecting the magnified F0 contours of L1 lexical tones when under focus, which may have been transferred to their L2 English. While the speakers clearly distinguished focused from unfocused conditions, no significant differences were found between contrastive and narrow focus. Future research should explore a wider range of focus types (e.g., broad focus) to examine whether Taiwan Mandarin speakers systematically differentiate focus types. Additionally, non-focus-related pitch accents (e.g., prenuclear accents) should be examined to assess whether the observed focus-marking patterns hold across different prosodic contexts in L2 English.

	monosyllabic	disyllabic	trisyllabic
Vowel /æ/	man	manor	manager
Vowel /eɪ/	mane	mainland	mania

Table 1: Test words used in the experiment.

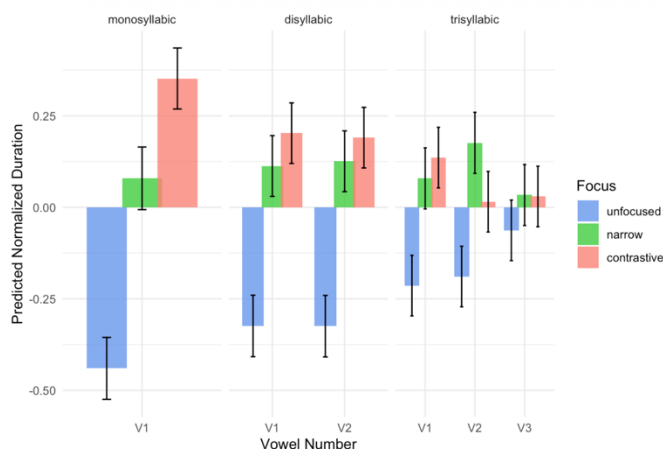


Figure 1: Predicted duration (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused) by Word Length (monosyllabic, disyllabic, trisyllabic) and Vowel Number (V1, V2, V3).

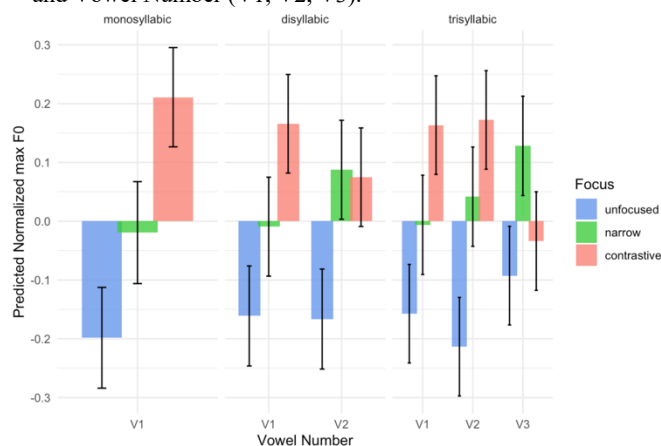


Figure 3: Predicted max F0 (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused) by Word Length (monosyllabic, disyllabic, trisyllabic) and Vowel Number (V1, V2, V3).

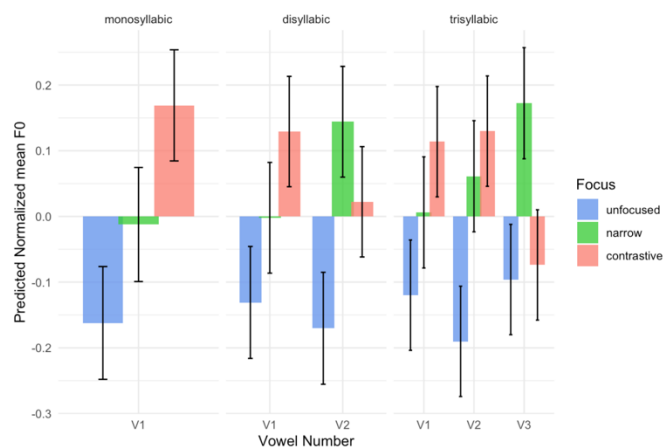


Figure 2: Predicted mean F0 (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused) by Word Length (monosyllabic, disyllabic, trisyllabic) and Vowel Number (V1, V2, V3).

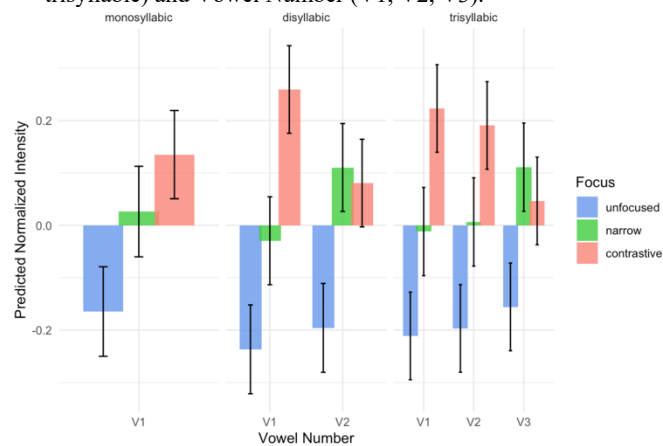


Figure 4: Predicted intensity (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused) by Word Length (monosyllabic, disyllabic, trisyllabic) and Vowel Number (V1, V2, V3).

References

- Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255–310. <https://doi.org/10.1017/S095267570000066X>
- Chen, S., Wang, B., & Xu, Y. (2009). Closely related languages, different ways of realizing focus. *Interspeech 2009*, 1007–1010. <https://doi.org/10.21437/Interspeech.2009-298>
- Chen, Y. (2010). Post-focus F0 compression—Now you see it, now you don't. *Journal of Phonetics*, 38(4), 517–525. <https://doi.org/10.1016/j.wocn.2010.06.004>
- Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, 77(6), 2142–2156. <https://doi.org/10.1121/1.392372>
- Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846–1849.

Establishing underlying tonal contrast for two surface Mid tones: A case for Western Naxi

QIÚ Ào

Naxi is a Tibeto-Burman language spoken in southeast China. It is a typical tonal language. While tonal complexity of Eastern Naxi is well discussed in previous studies [1, 2], Western Naxi is commonly reckoned as operating with rather straightforward syllable-tone systems, lacking processes of tonal change [3, 4]. However, in the Sānbà dialect, a Western Naxi variety scarcely described by previous scholars, we discovered that tonal patterns of verbs are far from being straightforward.

Although in Sānbà Naxi only four contrasting surface tones are observed, namely H(igh), M(id), L(ow), and R(ising), which are typical of Western Naxi dialects, Sānbà Naxi verbs actually operate with 5 underlying tones, among which two distinct M tones (termed M_a and M_b) neutralize when appearing in monosyllabic verb forms. When verb forms become multi-syllabic after processes such as negation, reduplication, and nominalization, the two M tones surface differently.

Figure 1 shows that the F₀ curves of the M_a tone differs little (< 5 Hz) from that of the M_b tone when pronounced in isolation. However, in negated forms, the F₀ of the negation prefix /mɿ-/ before an M_a verb root is approximately 20 Hz lower than before an M_b verb root, as is shown in Figure 2a; the negator carries a surface L tone and a surface M tone in these two respective circumstances. Similarly, in Figure 2b, we can see that the nominalizer surfaces as carrying the H tone when appended after the M_a verb root, and as carrying the M tone when attached to the M_b verb root.

Category	Isolation	Negation	Reduplication	Nominalization	Example: /ka/
H	H	H. M	H.M	H .M	‘to scoop’
M _a	M	L. M	M.H	M .H	‘to close (books)’
M _b	M	M. M	M.M	M .M	‘to swallow’
L	L	M.L	L.M	L .M	‘to be tired’
LH	LH	L. R	R.R	L .H	‘to be drunk’

Table 1 (Tones in **boldface** indicate verb roots)

In addition to the patterns demonstrated in Table 1, we also explored the tonal behaviors of the 5 categories in ‘V₁ V₂’ and ‘V₁ NEG-V₂’ constructions. Examples such as (a) and (b) below support the analysis of the M_b category as underlyingly toneless, as well as the postulation of a floating H tone (#H) for M_a (see Figure 3 for derivations). Underlying forms of the other tonal categories were also proposed, along with phonological tone rules deriving the surface forms therefrom.

- (a) R + M_b → L.H ly^R ‘to look at’ ɳɖɕ^{M_b} ‘should’ ly^Lɳɖɕ^H ‘should look at’
R + M_a → L.M ly^R ‘to look at’ t^hɕ^{M_a} ‘can’ ly^Lt^hɕ^M ‘can look at’
- (b) M_a + M_b → M.H sy^{M_a} ‘to kill’ ɳɖɕ^{M_b} ‘should’ sy^Mɳɖɕ^H ‘should kill’
M_a + M_a → M.M sy^{M_a} ‘to kill’ t^hɕ^{M_a} ‘can’ sy^Mt^hɕ^M ‘can kill’

Comparing the Sānbà tonal system with that of the well-documented Lijiāng dialect of Western Naxi, we found a clearcut correspondence of M_a:H and M_b:M (Sānbà:Lijiāng). The fact that the floating High tone (M_a) corresponds to an overt H tone offers supporting evidence for the floating H tone postulated. This also goes to suggest that M_a and M_b are of different historical origins, which provides basis for the underlying tonal contrast being used in comparative studies.

Keywords: Naxi, tone, generative phonology, underlying tone

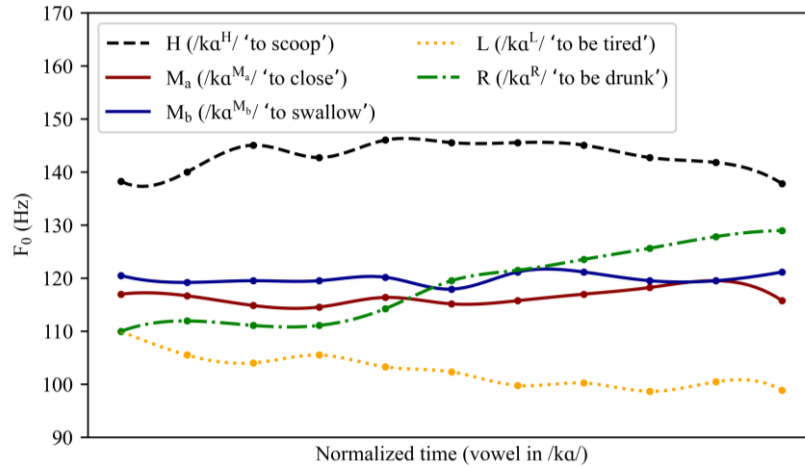


Figure 1: F₀ curves of the five tones in isolation

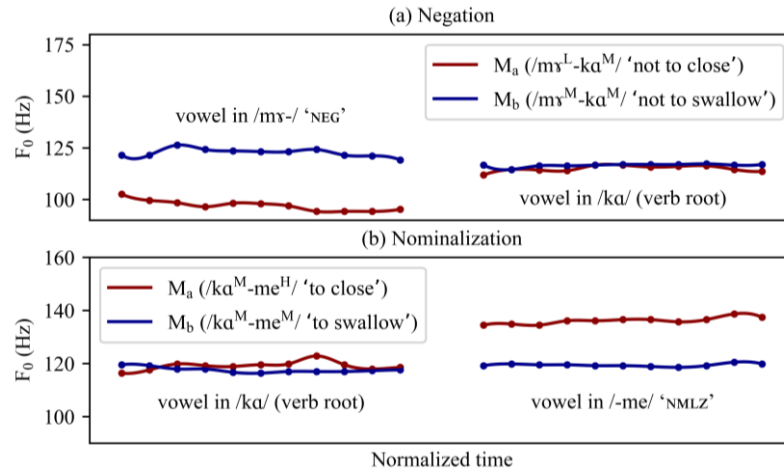


Figure 2: F₀ curves of disyllabic forms of M_a and M_b verbs

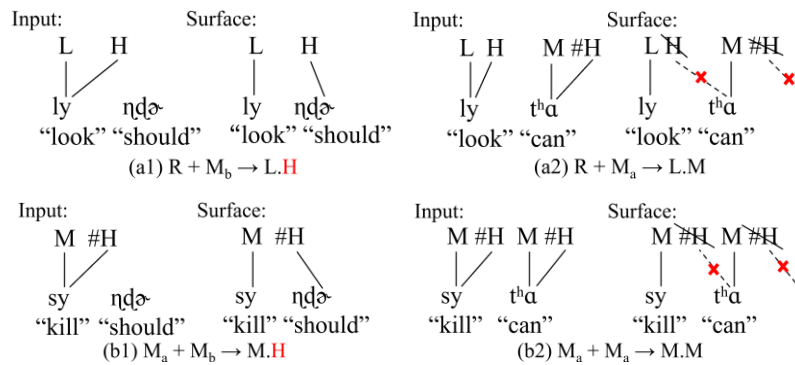


Figure 3: Derivations for examples (a) and (b)

- [1] Michaud, A. (2017). *Tone in Yongning Na*. Language Science Press. <https://doi.org/10.5281/zenodo.439004>
- [2] Dobbs, R., & Lă, M. (2016). The two-level tonal system of Lataddi Narua. *Linguistics of the Tibeto-Burman Area*, 39(1), 67–104. <https://doi.org/10.1075/ltba.39.1.04dob>
- [3] Hé, J., & Jiāng, Z. (1985). *Nàxīyǔ jiǎnzhi* [A brief documentation of Naxi]. Mǐnzú Chūbǎnshè.
- [4] Michaud, A. (2008). Phonemic and tonal analysis of Yongning Na. *Cahiers de Linguistique–Asie Orientale*, 37(2), 159–196. <https://doi.org/10.1163/1960602808X00064>

Roles of Duration in Cueing the Flat-Falling Tonal Contrast in Mandarin

Wei Zhang¹ and Wentao Gu²

¹School of Foreign Studies, Nanjing University, China

²School of Chinese Language and Literature, Nanjing Normal University, China

Introduction: Mandarin has four lexical tones: high-flat (H), rising (R), low-dipping (L) and falling (F). Although duration is not a primary cue for tonal contrast, durational differences exist among these tones in the citation form, e.g., L is the longest while F is the shortest (Xu, 1999). Mixed results have been reported for the role of duration in cueing tonal contrast between H and F. By comparing two or three steps of duration, Wang & Peng (2012) and Feng & Peng (2018) found there is no significant effect of duration on perceptual identification of the H-F tonal contrast. Using two-dimensional F0 continua, Zhang & Gu (2023, 2024) revealed that F0 range instead of F0 slope is the primary cue for distinguishing F from H, while duration plays only a secondary role as F is inherently shorter than H.

The patterns revealed in Zhang & Gu (2023, 2024) suggested that the perceptual identification rate of F may not vary linearly with duration; also, there could be interactions between F0 range and duration. This motivated us to explore whether the secondary role of duration in tonal identification varies across distinct intervals of duration. Additionally, we noted different durational settings in the literature, e.g., 250/375/500 ms in Feng & Peng (2018), and 7 steps from 100 ms to 400 ms in Zhang & Gu (2023, 2024). To make a comprehensive investigation into this issue, this study examined perceptual identification of the H-F tonal contrast in a wider duration domain with a finer step size.

Methods: To better elicit phonetic mode of perception, we adopted connected speech rather than isolated syllables as materials. On the basis of an utterance [pa⁵⁵ tɤ ʂəŋ⁵⁵] (“A sound like *bā*”), a set of speech stimuli were resynthesized by manipulating F0 and duration of the H-tone vowel [a⁵⁵]. A two-dimensional F0 continuum was set up between the H and F tones, with F0 fall range varying on 7 steps (0–76 Hz, with a step size of 12.7 Hz) and duration varying on 19 steps (50–500 ms, with a step size of 25 ms). Thirty native speakers of Mandarin (15F) participated in the perceptual experiment, making a two-alternative forced choice for each stimulus to determine whether the initial syllable was [pa⁵⁵] (H tone) or [pa⁵¹] (F tone).

Results: Figure 1 shows the rates of identification as the F tone. Visual inspection suggests that identification curves vary considerably with duration, and their relationship is nonlinear—with differences across duration steps being uneven and generally more pronounced for shorter syllables. We fitted the identification rates with a mixed-effects logistic regression model, in which F0 fall range, duration and their interaction were fixed effects, while participant was a random effect with random intercept and all random slopes included. The results of the model are shown in Table 1. While duration exhibits a weaker effect than F0 range—thereby confirming F0 range as the primary cue—its impact exceeds that reported in Zhang & Gu (2023, 2024). Moreover, there is a significant interaction effect between F0 range and duration.

To examine the effects in more detail, we divided the duration domain into four intervals (i.e., 50–75 ms, 100–200 ms, 225–375 ms, and 400–500 ms) based on the patterns shown in Figure 1, and fitted the identification rates in each interval with a Bayesian logistic regression model using *brms* (Bürkner, 2018). Table 2 shows the results of the models in four intervals. Across all four intervals, F0 range shows credible effects that are greater than those of duration. In Interval 1, there is a credible interaction between F0 range and duration. In other three intervals, duration exhibits *negative* effects, with a weaker effect in Interval 4.

Discussion: Results demonstrate that the role of duration in cueing the H-F tonal contrast in Mandarin is not uniform but varies across distinct duration intervals. For very short syllables (< 100 ms), the observed interaction effect suggests complexity in tone identification, and for wide F0 ranges a slightly extended temporal window facilitates identifying dynamic tones that are characterized by time-varying pitch movements. When syllables exceed 100 ms, duration shows negative effects on F tone identification, likely reflecting the inherently shorter duration of F compared to H. However, as syllable duration increases further (≥ 400 ms), the negative effect weakens, possibly approaching a ceiling effect. Finally, it should be noted that identification rates vary with duration more conspicuously than reported in Zhang & Gu (2023, 2024). This discrepancy is likely attributable to the use of duration-manipulated syllables embedded in a fixed carrier phrase, wherein tone identification may be subject to perceptual normalization processes that calibrate duration cues relative to the speech context.

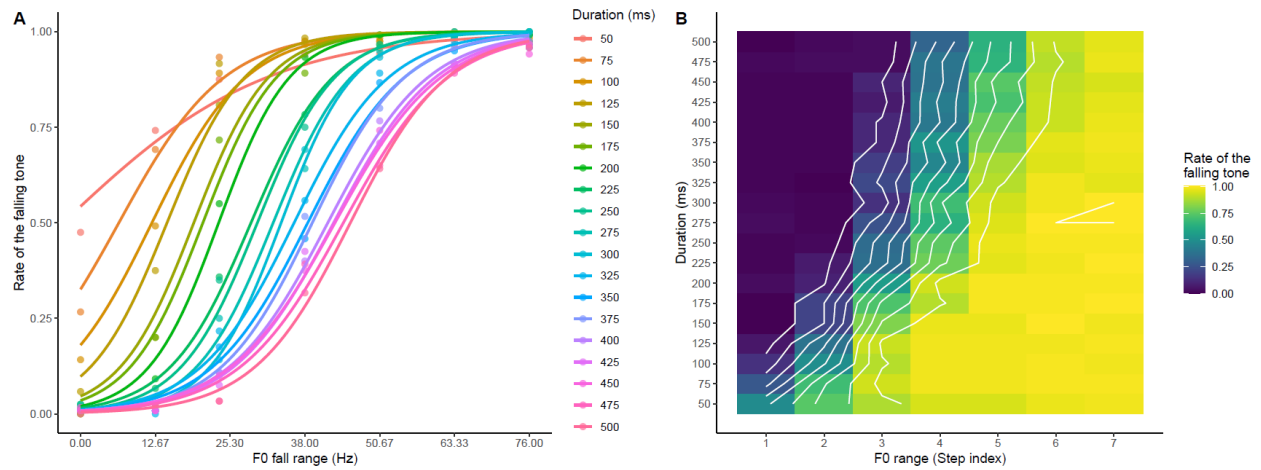


Figure 1: (A) Identification rates of the F tone as a function of F0 fall range at varying durations (the colors), smoothed using a generalized linear model fitting the empirical data. (B) Heat map of the identification rates (the colors) as a function of F0 fall range (the abscissa) and duration (the ordinate), together with a set of constant-rate contours with a 0.1 rate difference between adjacent contours.

Table 1: Results of the mixed-effects logistic regression model on the rates of identification as the F tone. (All effects here are significant, $ps < .001$)

	(Intercept)	F0Range	Duration	F0Range:Duration
β (SD)	1.30	8.60	-3.65	2.22

Table 2: Results of the Bayesian logistic regression models for the four duration intervals. The posterior estimates of effect sizes and associated 95% credible intervals are presented. Bolded values indicate statistically credible effects, with a posterior probability exceeding 95% for directional consistency (positive or negative).

	Duration Interval 1 (50–75 ms)		Duration Interval 2 (100–200 ms)		Duration Interval 3 (225–375 ms)		Duration Interval 4 (400–500 ms)	
	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
(Intercept)	4.09	[2.56, 5.38]	3.81	[2.44, 5.02]	0.75	[0.28, 1.22]	-0.66	[-1.19, -0.12]
F0Range	7.23	[4.90, 9.24]	11.13	[7.49, 14.27]	9.55	[8.67, 10.48]	8.75	[7.62, 9.96]
Duration	0.08	[0.02, 0.14]	-1.29	[-2.41, -0.08]	-1.29	[-1.73, -0.86]	-0.53	[-0.87, -0.18]
F0Range:Duration	0.21	[0.10, 0.33]	0.23	[-2.09, 2.65]	-0.19	[-1.49, 1.12]	-0.16	[-1.17, 0.84]

References

- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411.
- Feng, Y., & Peng, G. (2018). The effect of duration on categorical perception of Mandarin tone and voice onset time. *Proceedings of the 6th International Symposium on Tonal Aspects of Languages*, pp. 164–168, Berlin, Germany. <https://doi.org/10.21437/TAL.2018-33>
- Wang, D., & Peng, G. (2012). Effects of pitch range and duration on tone categorical perception. *Proceedings of the 3rd International Symposium on Tonal Aspects of Languages*, Nanjing, China. https://www.isca-speech.org/archive/tal_2012/tl12_O2-03.html
- Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, 27(1), 55–105.
- Zhang, W., & Gu, W. (2023). F0 range instead of F0 slope is the primary cue for the falling tone of Mandarin. *The Journal of the Acoustical Society of America*, 153(6), 3439–3446. <https://doi.org/10.1121/10.0019712>
- Zhang, W., & Gu, W. (2024). The roles of F0 range/slope and duration in cueing the Mandarin rising or falling tones. In: *East Meets West: Explorations in Tone and Intonation—Proceedings of TAI 2023*, Singapore: COLIPS Publications, pp. 115–116.

Acknowledgments: This work is supported by the Major Program of the National Social Science Fund of China (13&ZD189).

Tonal effects on articulation in Mandarin: An ultrasound study

Albert Lee, Penghao Cai, Zhiqiang Zhu, Peggy Mok

albertlee@eduhk.hk, caip@eduhk.hk, zhiqiangzhu@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

Keywords: Individual differences; Tone-vowel interaction; Tongue-surface; Ultrasound

Background: The link between fundamental frequency (f_0) and vowel height has been well documented. Generally, lowering f_0 may increase vowel aperture due to the thyrohyo-mandibular chain (Erickson et al., 2017). Studies such as Svensson Lundmark et al. (2021) have reported findings consistent with this tendency. However, Shaw et al.'s (2016) EMA data from Mandarin speakers appeared to observe counter-evidence. Specifically, for the vowel /i/ condition *alone*, tones with a low f_0 onset conditioned a *higher* tongue body position. In this paper, we revisited this anomalous finding by testing a larger sample.

Methods: We recruited 23 native Beijing Mandarin speakers (6 male, mean age 24.00 ± 2.80) from Hong Kong universities. We compared two vowels (/a/, /i/), four tones (T1, T2, T3, T4), two speech rates (normal, slow), and 3 CV structures (CV, CVn, CVŋ). Participants read target syllables framed in the carrier sentence ‘我 读 X 字 (I read the character X)’, presented in Simplified Chinese characters. Ultrasound (depth = 90 mm, Frame rates = 60 Hz) and audio recordings were collected. Tongue surface splines were tracked using AAA's built-in DeepLabCut (Wrench & Balch-Tomes, 2022) and analyzed using Smooth Spline ANOVA (Davidson, 2006). The tongue spline data were converted from Cartesian to polar coordinates by shifting the origin of the ultrasound image to about the center of the ultrasound probe ($X = \text{median } X$, $Y = 0$).

Results: Based on visual inspection of by-speaker SS ANOVA plots, significant differences in tongue positions between high T1 and low T3 were observed in 17 of 23 participants. They were then categorized based on whether these differences were congruent with physiological predictions (i.e. higher for T1, lower for T3), as illustrated as Table 1. There was substantial inter-speaker variability in terms of how tongue position interacted with lexical tone. Most speakers were in the congruent groups ($N = 10$, rows 1 to 3), fewer ($N = 7$, rows 4 to 9) were in the incongruent groups, of which only 2 of 23 speakers (row 4) replicated Shaw et al.'s (2016) findings. There were more speakers in the incongruent /i/ ($N = 6$) than in the incongruent /a/ groups ($N = 1$ or 2).

Discussion: We set out to verify Shaw et al.'s (2016) anomalous findings in his vowel /i/ condition, but our larger sample ($N = 23$) turned out to uncover an even more complex cross-speaker variation. This echoes Svensson Lundmark et al. (2021) who also anecdotally reported speaker variability. At this stage, two questions are worth asking: (i) Is Mandarin anomalous in terms of tone-vowel interaction? (ii) What are the underlying factors of the observed individual variability? To answer these questions, future researchers can revisit tone-vowel interaction through a language that has an even more complex tone inventory. To this end, currently collection of Cantonese ultrasound and audio data is underway.

References

- Davidson, L. S. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *Journal of the Acoustical Society of America*, 120(1), 407–415.
- Erickson, D. M., Honda, K., & Kawahara, S. (2017). Interaction of jaw displacement and F0 peak in syllables produced with contrastive emphasis. *Acoustical Science and Technology*, 38(3), 137–146.
- Shaw, J. A., Chen, W., Proctor, M. I., & Derrick, D. J. (2016). Influences of tone on vowel articulation in Mandarin Chinese. *Journal of Speech, Language, and Hearing Research*, 59, S1566–S1574.
- Svensson Lundmark, M., Frid, J., Ambrazaitis, G., & Schötz, S. (2021). Word-initial consonant-vowel coordination in a lexical pitch-accent language. *Phonetica*, 78(5–6), 515–569.
- Wrench, A. A., & Balch-Tomes, J. (2022). Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut. *Sensors*, 22, 1133.

Table 1. Classification of speakers by physiological congruence and vowel conditions.

		/a/	/i/	Number of participants
Physiological congruence: ("+" if consistent with physiological predictions, else "-"; "0" for N. S. difference between high vs. low tones)	1	+	+	4
	2	+	0	2
	3	0	+	4
	4	+	-	2 (cf. Shaw et al. 2016)
	5	-	+	1
	6	-	-	0
	7	-	0	0
	8	0	-	3
	9	Slow - Normal +	-	1
	10	0	0	6

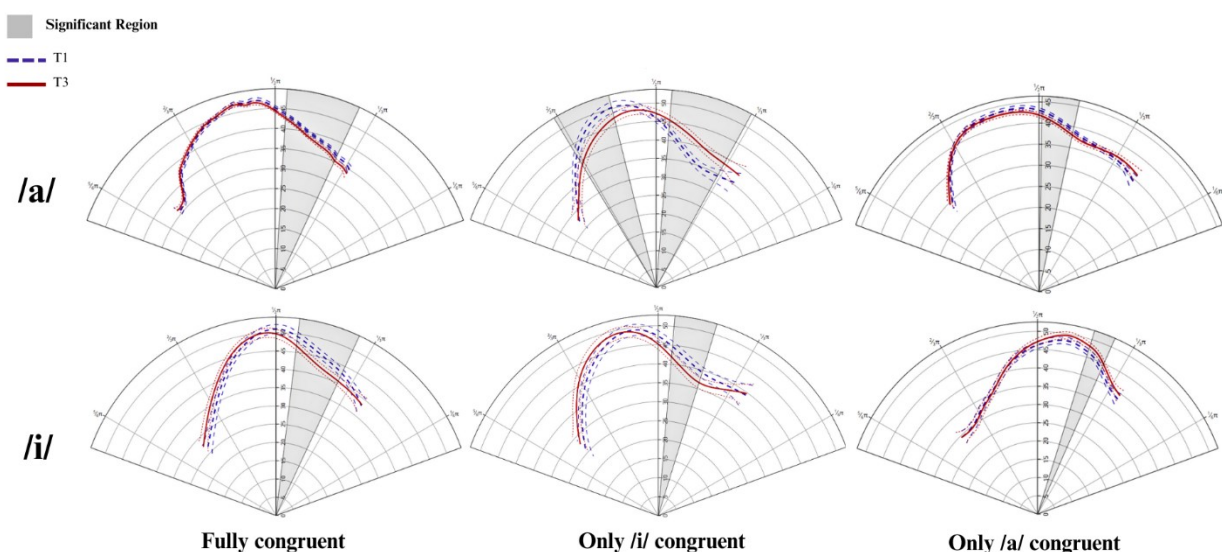


Figure 1. Smoothing spline ANOVA plots by tone (high T1 in blue, low T3 in red), vowel, and physiological congruence conditions. Grey areas indicate significantly different segments.

Examining the interaction between lexical stress and phrasal focus in autistic speech

Aya Awwad & Jeremy Steffman (The University of Edinburgh)

While autistic (henceforth ASD) people's ability to use lexical stress and contrastive focus has been reported in several studies, much remains to be understood about prominence-marking in the prosody of ASD speech (Grice et al., 2023). This is especially true as it pertains to the interplay of lexical stress and phrasal prominence, which we investigate in a task designed to elicit contrastive focus across different patterns of lexical stress in English.

In languages like English both lexical stress and pitch accents, which mark focus, impact the shape of phrasal prominence (Bolinger, 1965; Shattuck-Hufnagel, 1998), which is realized through acoustic correlates such as F0 and duration. Most fundamentally, focus-marking phrase-level prominence docks on lexically stressed syllables, and post-focal material is "de-accented" meaning lexically stressed syllables post-focus are predicted to lack F0 peaks. However, in autistic speech, there is paucity of research that investigates the interaction between lexical stress and focus, and how they jointly are exhibited in the expression of prominence.

Previous research findings report ASD people's successful placement of lexical stress (Van Santen et al., 2010; Arciuli & Bailey, 2019; and Arciuli et al., 2020). However, many studies report difficulties in producing contrastive focus at a phrasal level, by ASD people (Peppe et al., 2011; Diehl & Paul, 2012; Gargan & Andrianopoulos, 2022; and Patel et al., 2023). Moreover, even when lexical stress is reported to be placed on the correct syllable, it is often produced with different phonetic realizations than those of non-autistic people. These phonetic realisations include overall longer duration (Grossman et al., 2010; Paul et al., 2008), lower intensity (Arciuli & Bailey, 2019) or significantly different duration and intensity than non-autistic people (Ballarad et al., 2012). Finally, in studies of ASD speech, lexical stress and contrastive focus are often assessed separately. This stands at odds with a more linguistically informed understanding of lexical and phrasal prominence in which the acoustic representation of lexical stress and phrasal prominence are intertwined as described above.

We employ a picture-description task that will result in the production of a phrase which consists of a monosyllabic adjective + disyllabic noun, with the goal of assessing the production of contrastive vs broad focus in ASD speech, and the relationship between lexical stress, and phrase-level focus. Participants describe two pictures of toy animals that two individuals received in each trial ("Lana got a blue baboon, and Hana got a red rhino"). The description of the first picture provides the context, to which the description of the second picture provides the target region. If the context and target regions share information about the noun, contrastive focus is predicted to be placed on the adjective in the target region. If the context and target regions share information about the adjective, the contrastive focus is predicted to be placed on the noun in the target region. In cases where no similarities are shared between the context and target regions, broad focus is predicted, see Table 1.

The stimuli are built using four monosyllabic colours (i.e. red, blue, brown and green), and four disyllabic animals reflecting either lexical stress on the first syllable (i.e. rhino and tiger) or on the second syllable (i.e. baboon and racoon). There are 48 trials per participant counterbalancing nouns and adjectives, with 16 produced in each focus condition. We recruited 28 participants for this study (14 ASD and 14 non-autistic). The data represented currently is with 6 ASD speakers and 6 non-autistic speakers as a comparison group (all are speakers of British English). We will use k-means clustering for time-series data (Genolini & Falissard, 2011) to assess emergent distinctions in F0 across the three focus conditions and two stress conditions. FPCA (Ramsay and Silverman, 2005; Gubian et al., 2015) will also be used to analyse F0 trajectories over time series. Below are two plots that visualise the data we have so far. Plot 1 represents scaled F0 over time intervals of the adjective + noun target region which consists of three syllables in total. Plot 2 represents the duration of each syllable (indicated as syllable label on the x axis). Both plots are grouped by participant group (autistic -A- or non-autistic -T-) and lexical stress placement on the noun (1 for first syllable stress and 2 for second syllable stress).

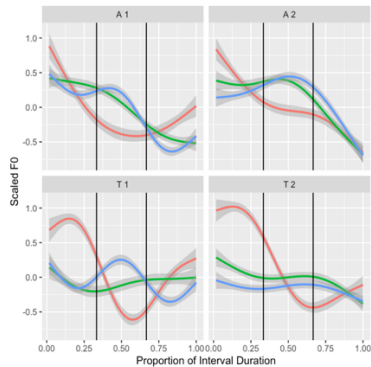
Methods: We fit a univariate FPCA to the data and model the first three scores (72.5% of variance accounted for) using Bayesian mixed effects regression, predicting the scores as a function of focus, stress and group with random intercepts for speaker and by speaker slopes for focus and stress, and their interaction. We report the probability of direction pd, which when greater than 97.5% is taken to be credible evidence for an effect.

Results: Score 1: main effects of focus (pds > 98%) and interaction between focus and lexical stress (pds = 99%) indicate that adjective focus entails a higher PC1 score (higher early peaks on PC1 curve and thus higher F0 on color words). These effects are diminished when lexical stress is on the first syllable. No credible effect of group or interactions with group were observed. For score 2: A credible two-way interaction between group and focus (pd = 99%) shows that non-autistic speakers have lower PC 2 scores for adjective focus, and comparable scores for broad and noun focus. Autistic speakers show a reversed pattern with higher PC2 scores for adjective focus which can be seen by the fall in PC2 curve and comparable scores for broad and narrow focus. For PC3, credible interactions between group and focus conditions (pds < 97.5%) show that non-autistic participants have lower PC3 scores (higher early peak in PC3 curve) for adjective focus. Autistic speakers have minimal differences in PC3 scores across focus conditions. Overall, the results of this study suggest some interaction between focus conditions and lexical stress in autistic speech. Further interpretations of the results will be discussed in the poster.

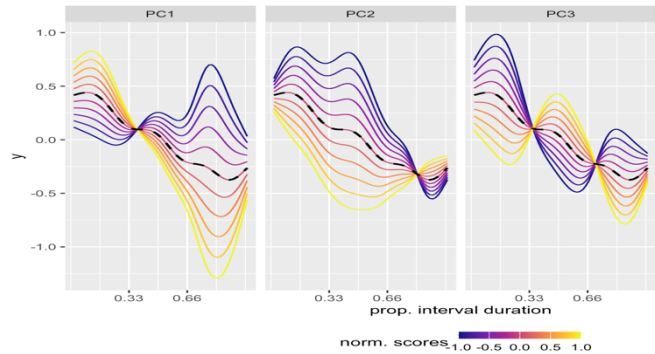
Keywords: Lexical stress, focus, prominence, autism, prosody

Noun Stress	Focus	Context phrase	Target phrase
First syllable	Broad focus	Blue baboon	Red rhino
First syllable	Adj focus	Blue rhino	Red rhino
First syllable	Noun focus	Red baboon	Red rhino
Second syllable	Broad focus	Blue rhino	Red baboon
Second syllable	Adj focus	Blue baboon	Red baboon
Second syllable	Noun focus	Red rhino	Red baboon

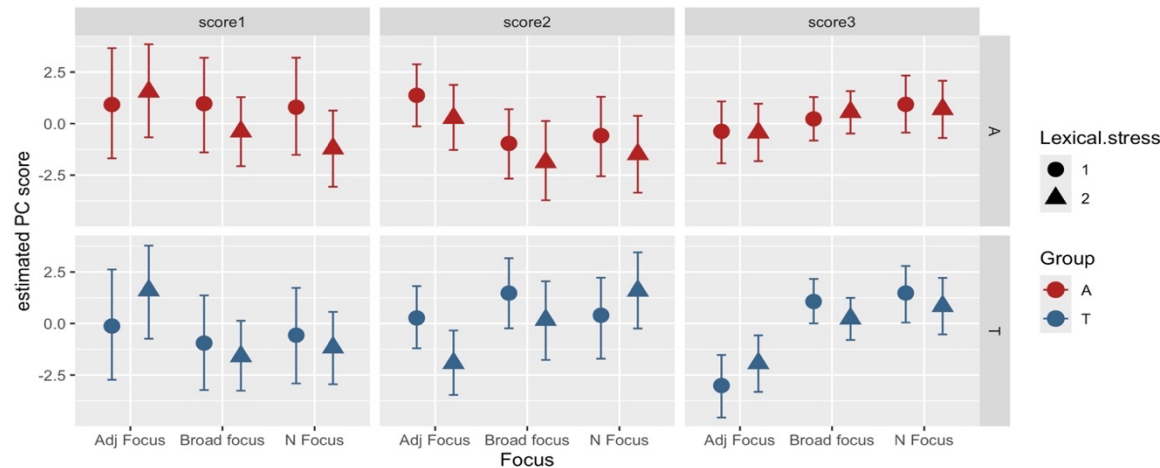
Table1: Stress and focus conditions in the picture description task, with examples.



Plot 1: Scaled F0 over time intervals of the target. Vertical lines separate each syllable.



Plot 2: PC curve for PC1, PC2, and PC3.



Plot 3: Estimated PC scores across different focus conditions, with color referring to participant groups and shape referring to lexical stress condition.

References:

Arciuli, J., & Bailey, B. (2019). An acoustic study of lexical stress contrastivity in children with and without autism spectrum disorders. *Journal of Child Language*, 46(1), 142-152.

Arciuli, J., Colombo, L., & Surian, L. (2020). Lexical stress contrastivity in Italian children with autism spectrum disorders: an exploratory acoustic study. *Journal of Child Language*, 47(4), 870-880.

Ballard, K. J., Djaja, D., Arciuli, J., James, D. G., & van Doorn, J. (2012). Developmental trajectory for production of prosody: Lexical stress contrastivity in children ages 3 to 7 years and in adults.

Bolinger, D. (1965). The atomization of meaning. *Language*, 41(4), 555-573.

Diehl, J. J., & Paul, R. (2012). Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders. *Research in autism spectrum disorders*, 6(1), 123-134.

Gargan, C. E., & Andrianopoulos, M. V. (2022). Receptive and expressive lexical stress in adolescents with autism. *International Journal of Speech-Language Pathology*, 24(6), 636-646.

Genolini, C., & Falissard, B. (2011). KmL: A package to cluster longitudinal data. *Computer methods and programs in biomedicine*, 104(3), e112-e121.

Grice, M., Wehrle, S., Krüger, M., Spaniol, M., Cangemi, F., & Vogeley, K. (2023). Linguistic prosody in autism spectrum disorder—An overview. *Language and Linguistics Compass*, 17(5), e12498.

Grossman, R. B., Bemis, R. H., Skwerer, D. P., & Tager-Flusberg, H. (2010). Lexical and affective prosody in children with high-functioning autism.

Gubian, M., Torreira, F., & Boves, L. (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49, 16-40.

Patel, R., & Campellone, P. (2009). Acoustic and perceptual cues to contrastive stress in dysarthria.

Paul, R., Bianchi, N., Augustyn, A., Klin, A., & Volkmar, F. R. (2008). Production of syllable stress in speakers with autism spectrum disorders. *Research in autism spectrum disorders*, 2(1), 110-124.

Peppé, S., Cleland, J., Gibbon, F., O'Hare, A., & Castilla, P. M. (2011). Expressive prosody in children with autism spectrum conditions. *Journal of Neurolinguistics*, 24(1), 41-53.

Ramsay, J. O., & Silverman, B. W. (2005). Principal components analysis for functional data. *Functional data analysis*, 147-172.

Shattuck-Hufnagel, S., & Turk, A. (1998). The domain of phrase-final lengthening in English. *Journal of the Acoustical Society of America*, 103(5), 1235-1236.

Van Santen, J. P., Prud'Hommeaux, E. T., Black, L. M., & Mitchell, M. (2010). Computational prosodic markers for autism. *Autism*, 14(3), 215-236.

Analysis of F0 Contour Alignment in Expressions of Attitude

Hansjörg Mixdorff

Berliner Hochschule für Technik, Germany

Keywords — attitudes, F0 contours, German, Hindi, Fujisaki model, intonemes

Following the framework established by Riilliard et al. [1], we collected audio-visual corpora of attitudinal expressions in German, Cantonese and Hindi, recognizing attitudes as fundamental components of speech communication. For instance, we react to our collocutor's assertions with approval or doubt, while shaping our reactions based on the social hierarchy. In this respect, attitudes exhibit emotional components but can be seen much more as communicative strategies than simple displays of affect. A crucial feature of the speech signal is the fundamental frequency contour *F0*, which we modulate during speech. Previous studies have explored how different attitudes influence mean and s.d. of *F0*[2]. However, there have been, to our knowledge, no studies examining whether there are consistent melodic contours or "tunes" associated with attitudes. The most notable exception is KIM by Kohler [3] which attributes to different *F0* peak alignments different pragmatic functions. Hence the current study aims to investigate whether specific tonal patterns appear in certain attitudes. Since attitudes are paralinguistic phenomena, they interact with underlying linguistic elements depending on the particular language such as stressed syllables and phrase boundaries. For the current study, we concentrate on the German and Hindi equivalents of "a banana", in German "eine Banane" [**a**In@ bana:n@], Hindi एक केला [**ek** kela], word stress syllables in bold. These data are by-products of the audio-visual corpora of attitudinal expression mentioned above. We are aware, that German and Hindi differ as to their prosodic phonologies. While German is an accent-timed language with *F0* reflecting word prominence and Hindi is said to employ *F0* mostly for phrasal marking, as we will see, in the context of our approach the prosodic labels attributed invariably reflect prominence as well as phrase type information.

Isačenko and Schädlich [4] and Stock and Zacharias [5] describe *F0* contours in German as characterized by a series of communicatively motivated tone switches, major transitions in the *F0* contour aligned with accented syllables, the phonetic realization of phonologically distinct intonational elements, referred to as "intonemes." Stock defined three classes of intonemes (see Table 1). These intonemes primarily differentiate sentence modality, although there exists a variant of the I↓ intoneme, I↓(E), which denotes emphatic accentuation and appears in contrastive, narrowly focused contexts. Intonemes for reading style speech can be predicted by applying a set of phonological rules to assess word accentability and accent group formation. Building upon this conceptual framework, Mixdorff and Jokisch developed a model of German prosody that links prosodic features such as *F0*, duration, and intensity to the syllable as the fundamental unit of speech rhythm [6]. To quantify the interval and timing of the tone switches relative to the syllabic grid, this framework adopts the Fujisaki model [7] (see Figure 1) which generates a given *F0* contour by combining three main components: a speaker-specific base frequency *Fb*, a phrase component, and an accent component. The latter arises from step-wise accent commands linked to accented syllables or boundary tones and are defined by their onset and offset times *T1* and *T2* and amplitude *Aa*.

We only included the 12 presenters rated best for German and Hindi in perception studies [8] and obtained a total of 12 speakers × 16 attitudes × 2 repetitions = 384 utterances per language. All target utterances were force-aligned at the phone level to establish syllable boundaries. *F0* values were extracted at intervals of 10 ms using the default method in PRAAT [9] and Fujisaki model parameter estimation [10].

Due to space limitations, we can only discuss very few results regarding the intonemes registered for German. Table 1 presents the types of attitudes considered, their abbreviations, and the frequency of observed patterns. Additionally, it provides the mean accent command amplitude *Aa* as a measure of *F0* range. In general, the underlying text "eine Banane" receives phrase stress on "na:" with a possible secondary stress on "al". Therefore, tone switches are primarily associated with these syllables. Various patterns are identified, each assigned an abbreviation: **IE**: (emphatic) information intoneme I ↓, characterized by a single accent command closely aligned with "na:" (see Fig. 1, top). **N+Iearly**: a single accent command starting in "al" and ending before "na:", indicating a non-terminal intoneme N ↑ on "al" and I ↓ intoneme before "na:". **N+Ilate**: a single accent command starting in "al" and ending after "al", suggesting a non-terminal intoneme N ↑ on "al" and I ↓ intoneme after "na:". **C**: Contact intoneme (*F0* rise) C ↑ associated with "na:". For IE and C, some variation is observed in combinations with a weak non-terminal N ↑ intoneme on the syllable "al": N_IE, N_C (instances provided in the "N_" column, see also Fig. 1, top). In a few cases of C ↑ intonemes, an additional high boundary tone B is associated with a second accent command on the last syllable "n@". Chi-square test shows a significant dependence of the intoneme distributions on attitude ($p < 0.01$), though obviously that does not hold for each pair of attitudes. Generally, there is a distinct division between declarative attitudes/I↓ intonemes and interrogative attitudes/C↑ intonemes. Neutral statements (DECL) are predominantly linked with IE, followed by N+Iearly and N+Ilate, typically displaying relatively low accent command amplitudes. The ADMI statement is nearly exclusively associated with IE, mirrored in SEDU and OBVI. QUES, SURP, and DOUB are overwhelmingly connected with C ↑ intonemes, some of which are further emphasized by boundary tones B. Only for UNCE did we note a significant number of C ↑ intonemes alongside the majority of IE which seems plausible, as uncertainty often goes along with a plea for support. N+Ilate occurrences are relatively infrequent except for IRRI and CONT, attitudes with strong negative connotations, wherein the syllable "na:" is pronounced in a stretched, flat fashion. Perceptually, these exhibit high pitch on "na:" similar to IE, highlighting the word "banana" whereas N+Iearly results in low *F0* on "na:", sounding more toned down. Interestingly, this pattern is most commonly observed for ARRO, but also for AUTH and POLI. N+Ilate occurrences are relatively infrequent except for IRRI and CONT, attitudes with strong negative connotations, wherein the syllable "na:" is pronounced in a stretched, flat fashion. Perceptually, these exhibit high pitch on "na:" similar to IE, highlighting the word "banana," whereas N+Iearly sounds more toned down. Interestingly, this pattern is most commonly observed for ARRO, but also for AUTH and POLI.

In conclusion, in our small corpus, besides important differences in *F0* range there are significant differences in the frequencies of underlying *F0* patterns or "tunes" in terms of tone switches. Within the declarative attitudes, most display a majority of (emphatic) IE intonemes, only ARRO has a majority of N+Ie. Other attitudes like CONT, AUTH and POLI at least have a substantial combined number of N+Ie and N+Il patterns. These cases are also connected with narrower *F0* range. Obviously, our "microscopic" analysis needs to be verified on a larger dataset with longer utterances.

Table 1: Types of intonemes for German, defined in Stock and Zacharias [5]. The boundary tone B was added by Author1.

type	function
N↑ intoneme	indicating incompleteness and continuation with rising tone switch on accented syllable
I↓ intoneme	occurring at declarative-final accents with a falling tone switch, conveying a message
C↑ intoneme	associated with question-final accents, rising tone switch, establishing contact
B	boundary tone, rising tone switch on question-final syllables (not necessarily accents)

Table 2: frequency of tonal patterns observed (total of 24) for German, as well as mean accent command amplitude Aa. “N_” denotes frequency of separate small accent command on syllable “aI” and frequency of tonal patterns observed.

attitude	abbrev.	IE	N+Ie	N+II	C	N_	B	Aa mean
admiration	ADMI	22	1	1	0	10	0	0.44
arrogance	ARRO	7	13	3	0	5	0	0.28
authority	AUTH	9	9	5	0	6	0	0.23
contempt	CONT	9	6	6	0	6	0	0.18
neutral statement	DECL	13	7	3	0	7	0	0.28
doubt	DOUB	0	0	1	22	6	1	0.78
irony	IRON	17	2	3	0	13	0	0.35
irritation	IRRI	14	0	8	1	8	0	0.31
obviousness	OBVI	19	0	4	0	10	0	0.34
politeness	POLI	9	8	5	0	6	0	0.36
neutral question	QUES	0	0	0	24	7	2	0.80
seductiveness	SEDU	20	0	0	2	13	1	0.41
sincerity	SINC	13	7	2	0	6	0	0.34
surprise	SURP	2	0	0	22	18	5	0.73
uncertainty	UNCE	15	0	1	7	15	0	0.36
walking-on-eggs	WOEG	14	5	0	1	10	0	0.23

Figure 2: Examples of prevalent F0 patterns in the German data (speaker number and attitude). (1) S04 ADMI (2) S05 POLI (3) S05 CONT (4) S04 QUES (5) S11 POLI. Each panel displays from the top to the bottom: The speech wave form, the F0 contour (extracted +++, modelled ---) the underlying step-wise accents commands of the Fujisaki model. Syllable boundaries are indicated by dotted vertical lines.

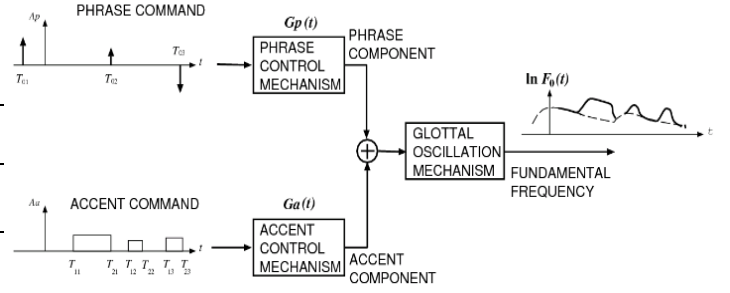
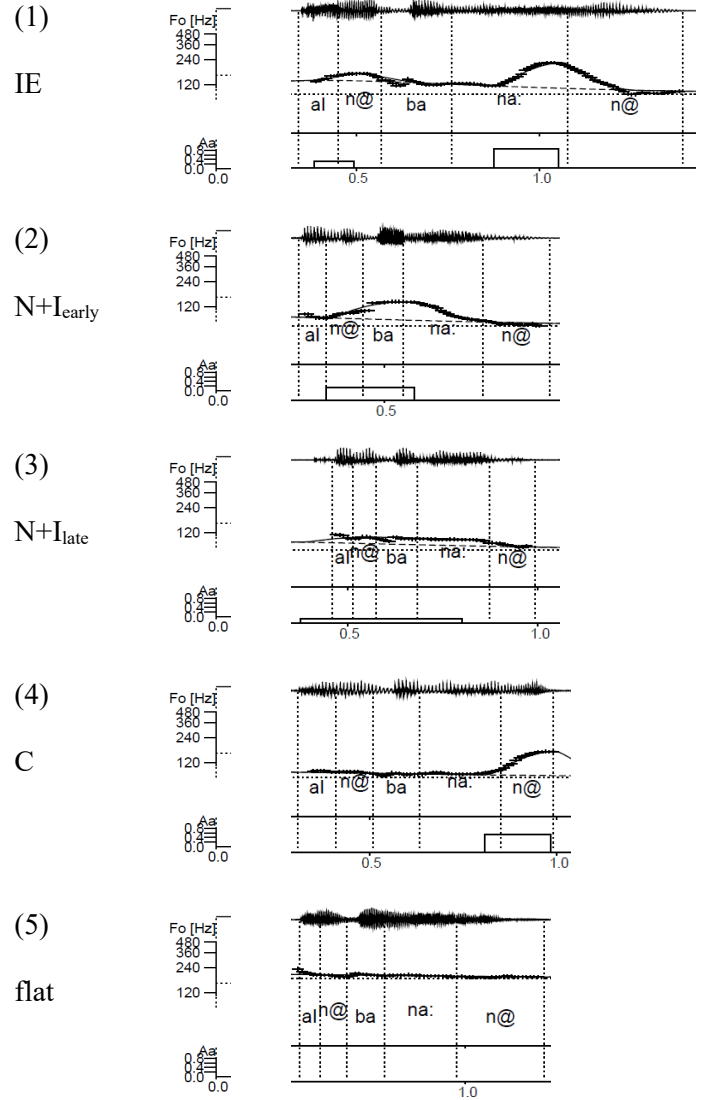


Figure 1: Schematic diagram of the Fujisaki Model



References:

- [1] Rilliard, A., Erickson, D., Shochi, T., and de Moraes, J. 2013. Social face to face communication - American English attitudinal prosody. INTERSPEECH 2013. 1648-1652.
- [2] Mixdorff, H., Hönemann, A., and Rilliard, A. 2015. Acoustic-prosodic Analysis of Attitudinal Expressions in German. Proceedings of Interspeech 2015, Dresden, Germany.
- [3] Kohler, K.J. 1991. A model of German intonation. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)* 25, 295-360.
- [4] Isačenko, A.V., Schädlich, H.J. 1982. *Untersuchungen über die deutsche Satzintonation*, Akademie-Verlag, Berlin, 1964.
- [5] E. Stock, C. Zacharias, *Deutsche Satzintonation*, VEB Verlag Enzyklopädie, Leipzig.
- [6] Mixdorff, H. and Jokisch, O. 2001. Building An Integrated Prosodic Model of German. In Proceedings of Eurospeech 2001, vol. 2, pp. 947-950, Aalborg, Denmark.
- [7] Fujisaki, H., Hirose, K. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” *Journal of the Acoustical Society of Japan* 5, 233-241.
- [8] Mixdorff, H., Nayan, N., Rilliard, A., Rao, P., Ghosh, D. 2023. Developing A Corpus Of Audio-Visual Attitudinal Expressions In Hindi. ICPhS 2023, Prague, Czech Republic.
- [9] Boersma, P. 2001. Praat, a system for doing phonetics by computer,” *Glott International* 5, 341-345.
- [10] Mixdorff, H. 2000. A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. Proceedings of ICASSP 2000, vol. 3, pages 1281-1284, Istanbul, Turkey.

A Tenth Century ToBI Transcription of Tiberian Hebrew

Sophia L. Pitcher

University of the Free State/SIL

sophia.pitcher@sil.org

The Tiberian scribal scholars who produced the Aleppo Codex (ca. 930 CE) devised a set of prosodic notations to preserve their oral reading tradition of the Hebrew Bible, effectively leaving a ToBI transcription of the entire corpus for modern phonologists to investigate. These graphemes, known as Masoretic accents, iconically indicate contrastive pitch and four levels of breaks in the flow of speech: clitics, prosodic words, minor prosodic phrases, and major prosodic phrases [see 4]. The system comprises two classes of graphemes traditionally described as conjunctives and disjunctives [5]. Conjunctive and disjunctive graphemes represent pitch accents, while disjunctives additionally distinguish tones at the boundaries of minor and major prosodic phrases [4]. This paper provides examples of the system’s iconic representation of the intonation and prosodic phrase structure of Tiberian Hebrew and reports findings regarding the system’s interface with the following: preboundary lengthening; syntax; the semantics of relative clauses; the identification of pronominal copulas; discourse and verbal uses of the copula; predicate focus (PF) and sentence focus (SF) constructions.

Scholars of the Masoretic accents broadly agree regarding the following: 1) the graphemes indicate the ‘melody’, stress, and phrase structure of the recited text; 2) the segmentations marked by disjunctive graphemes represent prosodic, not syntactic, phrasing; 3) the phenomena that the accents represent contribute to the meaning of the text and are an inextricable aspect of the recitation; and 4) the innovators of the graphemes and their contemporaries regarded the phenomena of the accents as an integral part of the Hebrew language, on par with the phonology of the consonants and the vowels.

While Dresner [2] was the first to demonstrate the prosodic basis of the Masoretic accents, he accepted the 19th century algorithm traditionally thought to determine the sequences of disjunctive graphemes in the text. Pitcher [4], however, provides a wholly linguistic and integrated analysis of the accentual graphemes based on the correspondence of their features and functions to those of prosodic systems. The phonological structure represented by the accents is motivated by the graphemes' alignment with syntactic constituents in accordance with Match Theory and by their alignment with prosodic words that exhibit preboundary lengthening. Furthermore, the prosodic graphemes are organized iconically to indicate two levels of embedded phrasing that are recognized by the graphemes' patterns of distribution.

(1) Discourse and Verbal Uses of the Copula

Tiberian Hebrew distinguishes discourse uses of the copula from verbal uses by the insertion of a prosodic phrase boundary after the copula. Numbers 16:7 provides an example of the discourse use, where a prosodic phrase boundary inserted directly after the copula (*wəħə'jə*) signals irrealis mood/future-tense time frame. Genesis 4:2 provides an example of the verbal use, where the prosodic boundary is absent at the juncture of the copula (*wəjhi*). Here the copula is cliticized to the subject 'Abel' (*hevel*), forming a single prosodic phrase with this constituent.

Discourse Use of the Copula – Insertion of a Prosodic Phrase Boundary (Numbers 16:7)

1(φ(הָיָא הַקֹּהֵן־וְשֵׁי) φ(אֲשֶׁר־יִבְחַר יְהוָה) φ(הָאֵשׁ)) 1(φ(וְהָיָא))
 ((wəħə'jə)φ) ((ħə'ʔij)φ (ʔaʕərjiv'ħar YHWH)φ (huʔ haqqə'də)φ)1
 ([[future]])φ)1 ((the.man)φ (whom-3ms.will.choose LORD)φ (he the.holy.one)φ)1
 [future] The man whom the LORD chooses, he will be the holy one.

How the Listener's Own Pitch Production Influences their Tone Perception: The Influence of Motor System in Speech Perception

Han Jiang¹, Xiaocong Chen¹, Martin, Pickering², Jiwei Ma¹, Caicai Zhang¹

¹The Hong Kong Polytechnic University, ²The University of Edinburgh

han-04.jiang@polyu.edu.hk, xiaocong.chen@polyu.edu.hk, martin.pickering@ed.ac.uk,
jiwei.ma@polyu.edu.hk, caicai.zhang@polyu.edu.hk

According to the motor theory [1] and the prediction by production theory [2], the speech motor system contributes to speech perception, especially under adverse listening conditions [3]. While previous studies have found evidence for the involvement of the motor system in the perception of segments (e.g., consonants), less explored is the perception of lexical tones, which are systematic differences in fundamental frequency (F0) controlled by the larynx [4]. There is considerable variability in F0 across speakers of a tonal language such as Cantonese; even within the same gender, one speaker's F0 characteristics could differ from those of another speaker [5]. Such variability makes lexical tone an ideal case to segregate the influences of the listener's F0 characteristics and the speaker's F0 characteristics on speech perception. To address these issues, we used articulatory suppression [6], a method to engage the speech motor system (sound articulation) and limit its role in simultaneous auditory perception, in Cantonese tone perception using a within-subjects design. We hypothesized that articulatory suppression would particularly hinder the use of the listener's own F0 characteristics.

Twenty-eight native Cantonese speakers were recruited. The experiment involved two tasks: (1) Production task, where participants produced four key Cantonese syllables to record their own pitch; and (2) Identification task, split into two sessions (with and without articulatory suppression) separated by at least two weeks. In the identification task, each participant listened to Cantonese syllables with white noise containing three synthesized level tones (high-level tone – T1, mid-level tone – T3, and low-level tone – T6) from 34 speakers with varying F0 ranges and were instructed to determine the tone of each syllable. In the articulatory suppression session, participants produced the sound [m2], a syllable with a high-rising tone (Tone 2) in Cantonese, while listening to the stimuli.

Following prior research indicating that speakers of tonal languages use relative pitch rather than absolute pitch in tone perception [7], we computed the relative F0 of listeners and the relative F0 of speakers (stimuli) as the semitone difference from the gender-specific population mean F0. These two predictors, together with the articulatory suppression condition (with vs. without), were included into a mixed-effect multinomial logistic regression. The result showed a significant main effect of the speaker's relative pitch on participants' tone perception (T1/T3 decision ratio: $\beta=0.77$, $p<0.001$; T6/T3 decision ratio: $\beta=-0.38$, $p<0.001$), replicating the prior finding. Tonal syllables produced by speakers with higher relative pitch tend to be perceived as the high-level tone (T1), while those from speakers with lower relative pitch tended to be perceived as the low-level tone (T6). There was a significant interaction between the speaker's and listener's relative pitch in the perception of tonal materials (T6/T3 decision ratio: $\beta=0.02$, $p<0.001$). Specifically, listeners with lower relative pitch tended to perceive tonal materials ranged from -5 to 5 semitones (relative to the population mean F0) as T3 (a mid-level tone) (highlighted, see Figure 1), possibly due to their lower threshold of T3. Furthermore, there is a significant interaction between the condition (with vs. without articulatory suppression) and listener's relative pitch (T1/T3 decision ratio: $\beta=0.10$, $p<0.005$; T6/T3 decision ratio: $\beta=-0.04$, $p<0.05$). The modulation of the listeners' relative pitch on T3 perception was reduced in the condition with articulatory suppression (see Figure 2). Taken together, the study provides some evidence that the motor system is engaged in tone perception, as the influence of listeners' relative pitch in tone identification differed between the two sessions with and without articulatory suppression. These findings supported our hypothesis that articulatory suppression would impede the listener's reliance on their own F0 information in speech perception.

Keywords: tonal perception, motor system, articulatory suppression, own voice, pitch

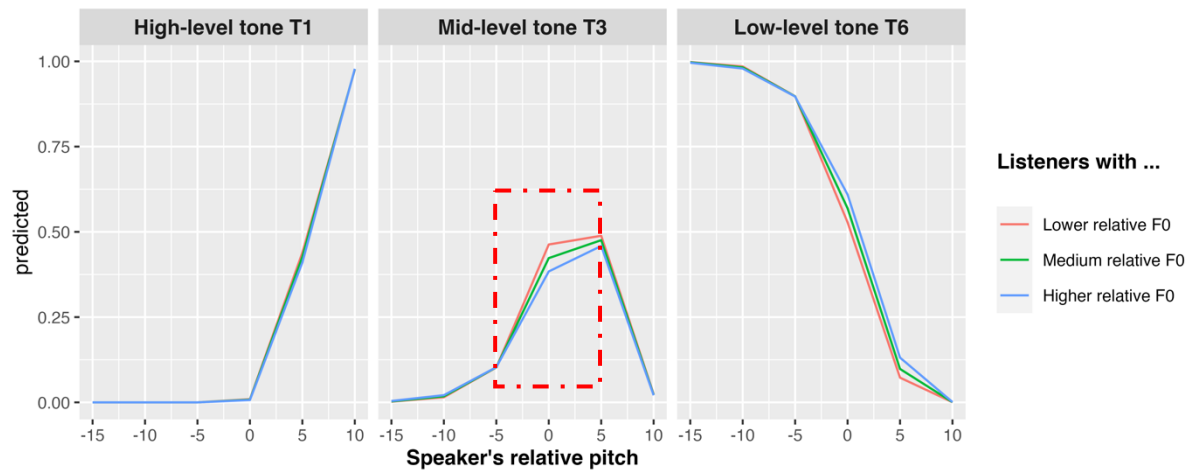


Figure 1. General Predictions on Interactions between Stimuli's Pitch and Listeners' Pitch

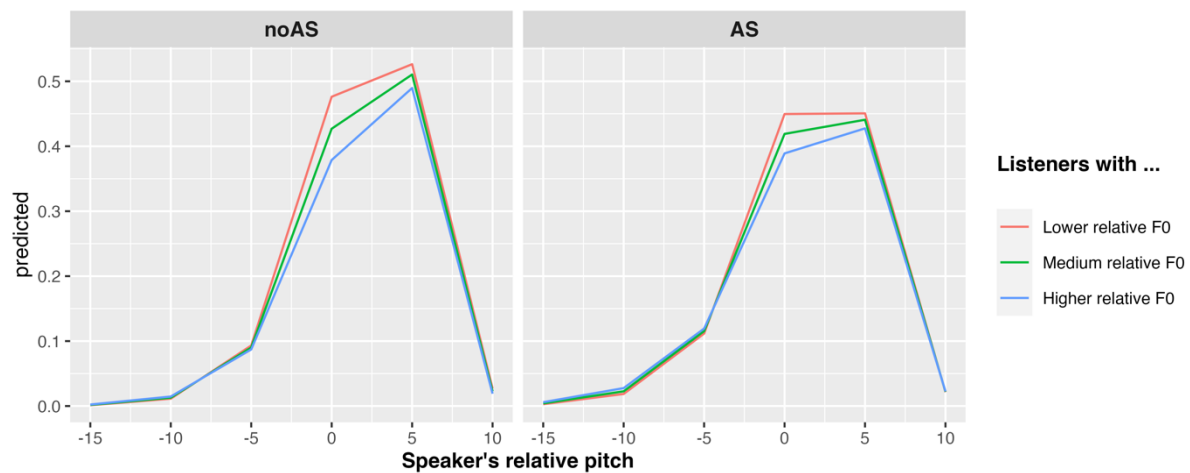


Figure 2. Articulatory Suppression Condition's Influence on the Identification of Mid-level tone T3

References

- [1] Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychological review*, 74(6), 431.
- [2] Pickering, M.J., & Garrod, S. 2007. Do people use language production to make predictions during comprehension?. *Trends in cognitive sciences*, 11(3), 105- 110.
- [3] Wu, Z.M., Chen, M.L., Wu, X.H., & Li, L. 2014. Interaction between auditory and motor systems in speech perception. *Neuroscience bulletin*, 30, 490-496.
- [4] Liang, B., Li, Y., Zhao, W., & Du, Y. 2023. Bilateral human laryngeal motor cortex in perceptual decision of lexical tone and voicing of consonant. *Nature Communications*, 14(1), 4710.
- [5] Zhang, C., & Chen, S. 2016. Toward an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance*, 42(8), 1252.
- [6] Yang, N. 2021. *Inner Speech in College ESL Reading: A Mixed Methods Study* (Doctoral dissertation, University of Cincinnati).
- [7] Chen, S., Zhang, C., Lau, P., Yang, Y., & Li, B. 2022. Modelling representations in speech normalization of prosodic cues. *Scientific Reports*, 12(1), 14635.

Vowel mergers and register restructuring in Chanthaburi Khmer

Pittayawat Pittayaporn¹, Sireemas Maspong^{1,2}, Sothornin Mam¹

¹ Center of Excellence in Southeast Asian Linguistics, Faculty of Arts, Chulalongkorn University, Thailand

² Institute for Phonetics and Speech Processing, Spoken Language Processing Group, LMU Munich, Germany

Register, a type of tonal contrast, emerged in many Southeast Asian languages following the loss of onset voicing distinctions [1,2]. While most register languages maintain multiple acoustic correlates, some have evolved to retain only vowel quality differences [3]. Standard Khmer represents a unique endpoint in this evolution, having completely replaced pitch and phonation contrasts with vowel quality distinctions, resulting in an extensive inventory of monophthongs and diphthongs. In contrast, Chanthaburi Khmer (CK) maintains both voice quality and vowel quality distinctions, presenting an opportunity to examine an intermediate stage in register evolution.

While previous studies examine the acoustics of historical reflexes of Middle Khmer (MK) vowels in CK [4,5], they focus on historical development rather than analyzing the synchronic vowel system and its interaction with register. This study presents new acoustic evidence demonstrating that vowels and registers operate independently in CK, as shown by the mergers of high vowels in the high register with mid vowels in the low register with respect to quality while maintaining phonation distinctions.

Methods. To investigate vowel quality mergers acoustically, we recorded five native CK speakers producing words containing MK vowels *i:, *e:, *ɛ:, *u:, *ɤ:, *u:, *o:, *ɔ:, in both high and low registers. We then extracted vowel quality measures (F1, F2) and voice quality measures (H1-A3, CPP) during the vowel portions of each word.

Results. Preliminary results show that the reflexes of MK high vowels *i:, *u:, *u: with a high register are realized as closing diphthongs [ei], [ɤu], and [ou] with F1 raising towards the end of the vowel. The reflexes of MK high-mid vowels *e:, *ɤ:, *o: with a low register are realized as [ei], [ɤu], and [ou]. Consequently, high vowels with a high register and high-mid vowels with a low register converge in F1-F2 space, indicating merger. Voice quality measurements confirm that registers maintain a robust distinction. Similarly, *o: with a high register and *ɔ: with a low register are realized as a closing diphthong [ɔo], indicating vowel merger, while maintaining distinct phonation type. These mergers are illustrated in Figure 1 and Figure 2.

Discussion. To understand the restructuring revealed by our acoustic results, we identify two crucial sound changes from MK to CK: (1) lowering of high register vowels, and (2) diphthongization of low register vowels driven by phonation-induced lengthening. Through the first process, MK high vowels *i:, *u:, *u: and high-mid vowel *o: with high register lowered to [ei], [ɤu], [ou], and [ɔo]. Through the second process, MK high-mid vowels *e:, *o:, and low-mid vowel *ɔ: with low register developed offglides, becoming [ɛi], [ɤu], [ou], and [ɔo]. These changes led to cross-register vowel mergers, creating a system in which vowel quality and phonation type function independently. The ordering, where vowel restructuring precedes voice quality loss, aligns with previous research [5] and supports established models of transphonologization [6]. Further perception study is needed to confirm whether vowel quality and phonation type function independently in native speakers' cognition.

Keywords: Register, transphonologization, vowel quality, phonation type, merger

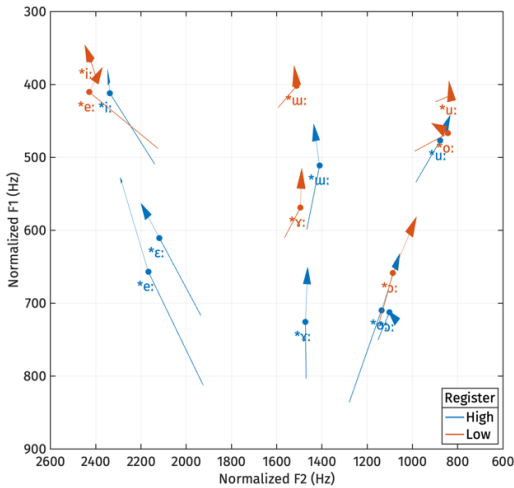


Figure 1 F1 and F2 space of Middle Khmer vowel reflexes in Chanthaburi Khmer. Beginning of the arrows represent F1 and F2 values at the first 20% of the vowel, the dot represents the values at the middle of the vowel, and the tip of the arrows represents the value at the 80% of the vowel

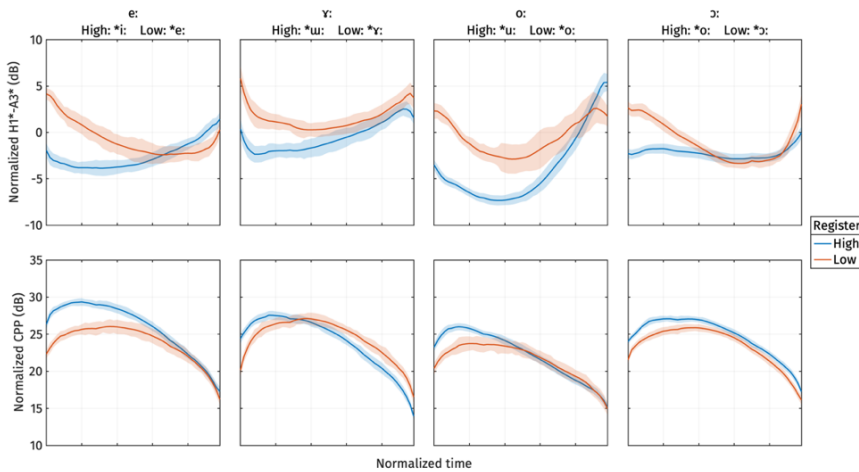


Figure 2 Voice quality measures of each merged vowel pair

References

- [1] Henderson, E. J. A. (1952). The Main Features of Cambodian Pronunciation. *Bulletin of the School of Oriental and African Studies*, 14(1), 149–174.
- [2] Brunelle, M., & Kirby, J. (2016). Tone and Phonation in Southeast Asian Languages. *Language and Linguistics Compass*, 10(4), 191–207.
- [3] Huffman, F. E. (1976). The register problem in fifteen Mon-Khmer languages. In P. N. Jenner & et.al (Eds.), *Austroasiatic Studies 1 (Oceanic Linguistics Special Publication 13)* (pp. 575–589). University of Hawaii Press.
- [4] Wayland, R., & Jongman, A. (2003). Acoustic correlates of breathy and clear vowels: The case of Khmer. *Journal of Phonetics*, 31(2), 181–201.
- [5] Maspong, S. (2023). *Synchrony and diachrony of vowel quality difference across registers in Mon-Khmer languages* [Ph.D. dissertation]. Cornell University.
- [6] Hyman, L. M. (2013). Enlarging the scope of phonologization. In *Origins of Sound Change: Approaches to Phonologization*. Oxford University Press.

Realization of Mandarin Neutral Tone: Adjacent Tones and Duration Effects across Styles

Jingyi Sun*, Yaru Wu, Nicolas Audibert, Martine Adda-Decker

**Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), Paris, France*

jingyi.sun@sorbonne-nouvelle.fr

In Standard Mandarin, neutral tone (NT) is traditionally considered phonologically underspecified (Chao, 1968), yet its pitch realization shows systematic variation across contexts. Two competing models offer different accounts: the PENTA model (Xu, 2005; Chen & Xu, 2006) posits that NT pitch is primarily shaped by carryover effects from preceding tones with minimal anticipatory influence, while the Stem-ML framework (Kochanski & Shih, 2000; 2003) allows for both carryover and anticipatory adjustments based on dynamic target interpolation. However, empirical evidence on how NT pitch contours are shaped by adjacent tones and temporal dynamics remains limited, particularly under naturalistic speech conditions where syllable duration and prosodic planning exhibit substantial variability. This study re-examines NT realization in spontaneous and read Mandarin speech, focusing on whether anticipatory adjustments systematically shape NT contours alongside carryover effects, and how these patterns are modulated by adjacent tones and duration.

The data come from a Mandarin corpus comprising 20 hours of spontaneous speech and 6 hours of read speech by 17 male and 13 female speakers (Sun et al., 2024). Forced alignment (McAuliffe et al., 2017) provided word, syllable, and phoneme annotations. We focused on two frequent NT sequences around Tone 4: "TX-NT-Tone4" (spontaneous: 4,216; read: 1,701) and "Tone4-NT-TX" (spontaneous: 4,285; read: 1,563), which may span across word boundaries.. NT tokens with creaky voice, durations over 0.3 seconds, or at sentence boundaries were excluded, and misclassified cases were corrected. f_0 trajectories were extracted using the FCN-F0 convolutional neural network (Ardillon & Roebel, 2019), and normalized via linear interpolation with 10 sampling points per token. The Generalized Additive Mixed Models (GAMM; Wood, 2017) analyzed f_0 contours, with fixed effects for normalized time and duration, random effects for speaker variability, and tensor smooth interactions to model the relationship between normalised time and duration across tonal contexts.

Figure 1 shows the frequency distribution of NT durations across styles. NTs in spontaneous speech are generally shorter than those in read speech, with durations between 0.03s and 0.06s accounting for over 50% of tokens in both styles. This variation enabled examination of NT contours under significant time constraints.

Figure 2 presents GAMM-estimated f_0 contours across subgroups and styles with three different total durations, 0.04s (left column), 0.08s (middle column), and 0.14s (right column). Panels (a) and (b) show NT contours when followed by Tone 4, while panels (c) and (d) display contours when NT preceded by Tone 4. Across all panels, NTs maintained a compressed pitch range, generally not exceeding two semitones, yet clear contouring patterns emerged depending on tonal context, duration, and style.

In panels (a) and (b) of the TX-NT-T4 subgroup, NTs preceded by Tone 1/2/4 showed declining contours at 0.08s and 0.14s, while those preceded by Tone 3 consistently exhibited rising trajectories across almost all durations. This upward trend, evident as early as 0.04s in read style, illustrates the combined influence of Tone 3's low offset and the high onset of Tone 4. Despite physiological limits on pitch movement speed at short durations (Ohala & Ewan, 1973), this early divergence suggests a stable anticipatory effect beyond passive carryover. At 0.08s, NTs preceded by Tone 2 also showed rising contours but with overall higher f_0 (1–2 semitones above T3), while at 0.14s, T3- and T2-context NTs diverged more clearly: the former remained rising, while the latter reversed into a falling contour. These patterns reveal that duration shapes the extent and direction of tonal integration between preceding and following tones.

Panels (c) and (d) of the T4-NT-TX subgroup highlight that when the NT followed Tone 4, carryover effects dominated, with contours generally falling across all durations. However, NTs followed by Tone 3 consistently exhibited slightly elevated f_0 trajectories compared to other contexts, particularly at short and mid durations, indicating residual anticipatory influence even in a carryover-dominated configuration.

Across all conditions, read speech exhibited wider pitch excursions than spontaneous speech, yet the overall contouring patterns remained consistent, suggesting that speech style primarily modulates the degree of tonal realization rather than the underlying integration mechanisms.

These findings demonstrate that while carryover effects from preceding tones govern NT contours, systematic anticipatory adjustments—especially triggered by following Tone 3—emerge early and are modulated by syllable duration. Crucially, anticipatory effects are not confined to the final portion of NT, challenging the strictly sequential target realization proposed by the PENTA model. Instead, NT realization reflects a dynamic tonal integration process, consistent with the predictions of the Stem-ML framework, operating even under significant temporal and articulatory constraints.

Keywords: neutral tone, Chinese Mandarin, spontaneous speech, carryover effect, pitch target

Acknowledgements

This work was supported by the Laboratoire d'Excellence "Empirical Foundations of Linguistics" (LabEx EFL, ANR-10-LABX-0083), as part of the ANR project DIPVAR (ANR-21-CE38-0019). Jingyi Sun was funded by the China Scholarship Council (Grant No. 202208410095).

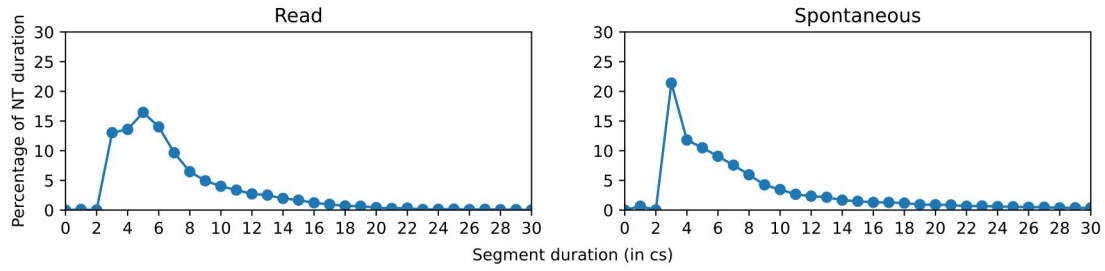


Figure 1 Distribution of NT durations in read and spontaneous speech (%)

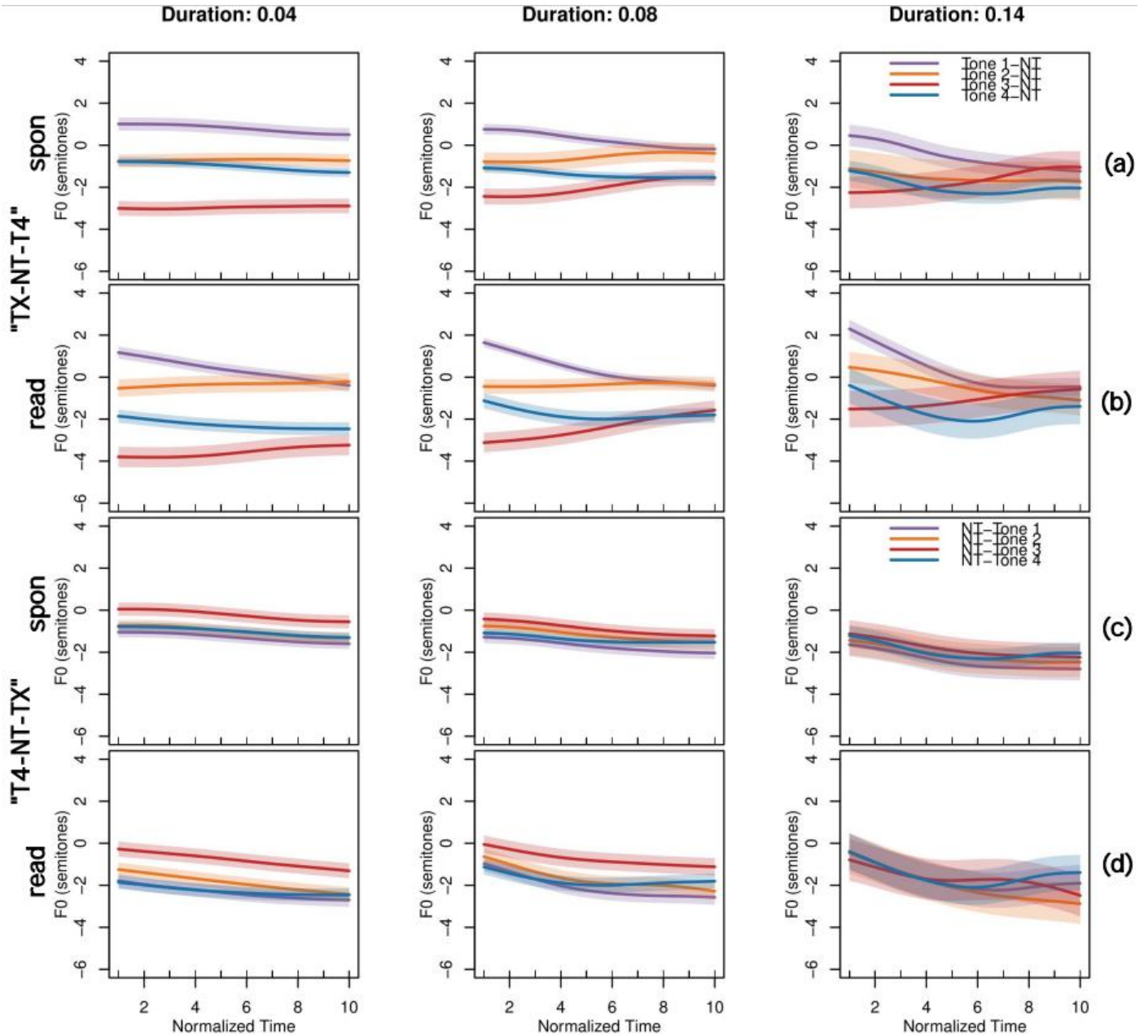


Figure 2 Fitted f_0 contours of NT in varied tonal contexts and durations (0.04s, 0.08s, 0.14s), with 95% CI

References

1. Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.
2. Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech communication*, 46(3-4), 220-251.
3. Chen, Y., & Xu, Y. (2006). Production of weak elements in speech: Evidence from f_0 patterns of neutral tone in Standard Chinese. *Phonetica*, 63(1), 47 - 75.
4. Kochanski, G., & Shih, C. (2000). Stem-ML: Language-independent prosody description. In *Proceedings of Interspeech 2000* (pp. 239 - 242).
5. Kochanski, G., & Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, 39(3 - 4), 311 - 352.
6. Sun, J., Wu, Y., Audibert, N., & Adda-Decker, M. (2024). Création d'un corpus parallèle de styles de parole en mandarin via l'auto-transcription et l'alignement forcé. In *Actes des 35èmes Journées d'Études sur la Parole* (pp. 291 - 300).
7. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech 2017* (pp. 498 - 502).
8. Ardailon, L., & Roebel, A. (2019). Fully-convolutional network for pitch estimation of speech signals. In *Proceedings of Interspeech 2019* (pp. 2005 - 2009).
9. Wood, S. N. (2017). *Generalized additive models: An introduction with R (2nd ed.)*. Chapman and Hall/CRC.
10. Ohala, J. J., & Ewan, W. G. (1973). Speed of pitch change. *The Journal of the Acoustical Society of America*, 53, 345 - 345.

An EEG Investigation of the Processing Mechanisms of Phonologically-conditioned and Semantically-conditioned Tonal Variations in Mandarin Word Production

Fumo Huang¹, Xiaocong Chen¹, Jiayi Lu¹, Caicai Zhang¹

¹*The Hong Kong Polytechnic University*

Keywords: Mandarin Tone Sandhi, Polyphonic Characters, Speech Production, ERP, Phonological Processing

In Mandarin Chinese, third-tone (T3) sandhi exemplifies phonologically conditioned tonal variations, where the tone is realized as different tonal variants depending on the following tonal contexts: either as a rising tone when followed by another third tone (the T3 sandhi) or as a low-falling tone when followed by other tones (the half-third sandhi) [1]. On the other hand, some polyphonic characters, which are characters with multiple pronunciations, exhibit semantically conditioned tonal variations. For example, the character 数 has two tonal variants (shu3 ‘to count’ and shu4 ‘number’), and its tonal realization is determined by lexical or semantic context. Although both tone sandhi and polyphonic characters lead to phonetic variants, the underlying mechanisms may differ, as the tone sandhi involves the same morpheme without altering the meaning, while polyphonic characters involve distinct morphemes with unique pronunciations. However, few empirical studies have compared the processing mechanisms between these two phenomena. Thus, this study investigates whether tone sandhi and polyphonic characters engage the same or different neural processing mechanisms during speech production using the event-related potential (ERP) method.

Thirty-nine native Mandarin speakers (20 females) participated in a phonologically primed picture naming task. Participants were asked to produce disyllabic target words, preceded by a visual monosyllabic prime together with the corresponding auditory form [2]. Target stimuli consisted of 24 T3 sandhi (e.g., 雨伞, yu3san3) and 24 disyllabic words containing polyphonic characters (with two tonal variants) in the initial position (e.g., 相机, xiang4ji1). The tonal overlap between the primes and the initial morpheme of the target was manipulated. For the sandhi words, three types of primes were included: T3 prime (sharing underlying tonal category, e.g., 与, yu3), T2 prime (sharing the surface tonal contour, e.g., 鱼, yu2), and T1/T4 prime (tonally-unrelated, e.g., 预, yu4). Likewise, for polyphonic stimuli, three types of primes were included: homophonous (e.g., 项, xiang4), variant (e.g., 乡, xiang1, alternative tonal variant of the polyphonic characters), or non-variant for polyphonic words (e.g., 祥 xiang2, tonally-unrelated). Both naming latencies and ERP signals were recorded.

For sandhi words, both T3 and T2 primes significantly reduced naming latencies (Fig.1A). ERP results revealed distinct effects for T3 and T2 primes across different time-locked stages. T3 primes elicited larger positive amplitudes during stimulus-locked processing stages (320–500 ms time windows) in the midline posterior region (Fig. 2A). In contrast, T2 primes induced larger negative amplitudes during response-locked processing stages (–400 to –250 ms) in the left anterior, midline anterior, left middle, midline middle, and right middle regions (Fig. 3A). These findings support the activation of both tonal variants of T3, but the processing of the underlying tonal category and the context-specific tonal variant showed distinct temporal dynamics [2]. For polyphonic stimuli, homophonous and variant primes also shortened naming latencies (Fig.1B). Moreover, both prime types elicited larger positive amplitudes during stimulus-locked processing stages (320–500 ms time windows) in the midline posterior and right posterior regions (Fig. 2B), suggesting simultaneous activation of different phonetic variants of polyphonic characters in the early stage of production, consistent with previous research on the processing of polyphonic wordforms [3][4]. Homophonous primes also exhibited significant effects in the early and late phases of response-locked ERPs in the left anterior, right anterior, and left middle regions (Fig. 3B), characterized by larger negative amplitudes, possibly suggesting the facilitation of articulatory planning. In general, this study underscores distinct neurocognitive mechanisms for the two types of tonal variations.

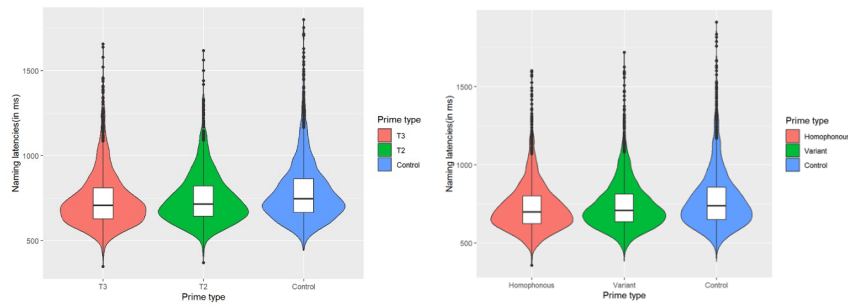


Figure 1A. Violin plots for naming latencies of sandhi words in different prime type conditions (T3, T2, and control) (left)

Figure 1B. Violin plots for naming latencies of polyphonic words in different prime type conditions (homophonous, variant, and control) (right).

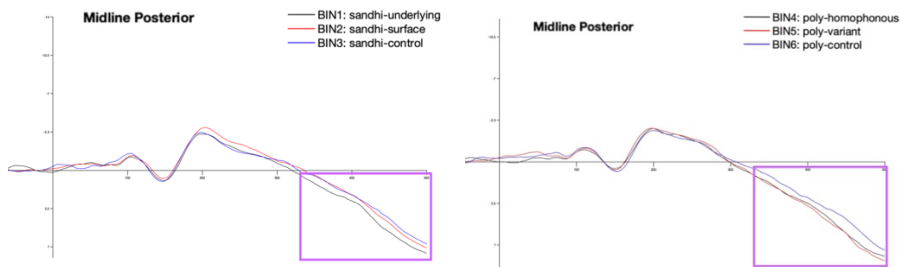


Figure 2A. Grand average stimulus-locked ERPs at the midline posterior region for sandhi words in different prime type conditions (T3, T2, and control) (left)

Figure 2B. Grand average stimulus-locked ERPs at the midline posterior region for polyphonic words in different prime type conditions (homophonous, variant, and control) (right).

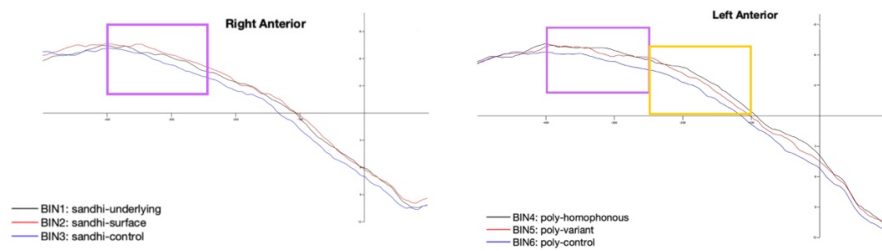


Figure 3A. Grand average response-locked ERPs at the right anterior region for sandhi words in different prime type conditions (T3, T2, and control) (left).

Figure 3B. Grand average response-locked ERPs at the left anterior region for polyphonic words in different prime type conditions (homophonous, variant, and control) (right).

References

- [1] Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, 27(1), 153-201.
- [2] Chen, X., Zhang, C., Chen, Y., Politzer-Ahles, S., Zeng, Y., & Zhang, J. (2022). Encoding category-level and context-specific phonological information at different stages: An EEG study of Mandarin third-tone sandhi word production. *Neuropsychologia*, 175, 108367.
- [3] Tan, L., & Peng, D. (1991). Visual recognition processes of Chinese characters: A study of the effect of grapheme and phoneme. *Acta Psychologica Sinica*.
- [4] Verdonschot, R. G., La Heij, W., Tamaoka, K., Kiyama, S., You, W. P., & Schiller, N. O. (2013). The multiple pronunciations of Japanese kanji: A masked priming investigation. *Quarterly Journal of Experimental Psychology*, 66(10), 2023-2038.

This study aims to understand the relationships between auditory perception, music perception, and literacy skills in Cantonese adults with congenital amusia (amusia hereafter). Amusia is a neurogenetic musical disorder characterized by difficulties in recognizing familiar melodies without lyrics and detecting out of tune melodies (Peretz et al., 2002). While amusia primarily affects musical pitch perception, it also extends to the speech domain, negatively affecting native and non-native lexical tone perception and intonation perception (Liu et al., 2010; Nan et al., 2010; Tillmann et al., 2011). In recent studies, English adults with amusia showed impaired segmental phonological awareness, and pitch discrimination threshold significantly predicted phonological awareness performance across amusia and control groups (Jones et al., 2009; Sun et al., 2017). However, it is unknown whether phonological awareness is impaired in tonal language speakers with amusia. Since suprasegmental information (i.e., pitch variations) systematically distinguishes word meaning in tonal languages, we aimed to investigate whether both segmental and suprasegmental phonological awareness are impaired in tonal language speakers with amusia. Furthermore, recent studies revealed a comorbidity between amusia and developmental dyslexia whose core deficit lies in phonological processing (Couvignou et al., 2019; Couvignou & Kolinsky, 2021). As informed by the Temporal Sampling Framework (TFS) hypothesis that the phonological processing deficit in dyslexia originates from more basic auditory rise time processing difficulties across language which can also impact music rhythm processing (Goswami et al., 2011), we also assessed rise time processing in amusical and control participants. In short, we examined (1) whether tonal language speakers with amusia underperformed controls on auditory perception at both pitch and temporal domains, phonological awareness at both segmental and suprasegmental levels, and literacy, and (2) whether auditory perception at both pitch and temporal domains predicted segmental and suprasegmental phonological awareness and literacy in tonal language speakers.

Thirteen Hong Kong Cantonese speakers with amusia and 11 matched controls were tested on three auditory discrimination tasks (pitch discrete, pitch glide and rise time), three phonological awareness tasks (onset phoneme deletion, spoonerism and tone oddity), and three literacy tasks (word reading, one minute reading and dictation). The results showed that the amusia group performed significantly worse than the control group in all three auditory measures (pitch discrete: $W = 131, p < .001$; pitch glide: $W = 126, p < .001$; rise time: $t(12.59) = 2.7, p < .05$), the spoonerism ($W = 24.5, p < .01$) and the tone oddity ($t(22) = -3.29, p < .01$) tasks, as well as the word reading ($W = 18.5, p < .01$) task. Fixed order stepwise regression (collapsing the two groups) showed that individual differences in the pitch discrete task predicted the participants' performance on all the three phonological awareness measures, while rise time only predicted tone oddity. Besides, individual differences in both pitch discrete and rise time tasks predicted word reading, while individual differences in the rise time task also predicted one minute reading, but their contribution became nonsignificant if the three phonological awareness measures were controlled for.

These findings suggest that Cantonese adults with amusia have poor segmental and suprasegmental phonological awareness and word reading, relative to controls, which are associated with their auditory processing at both pitch and temporal domains. This study has extended previous studies in two aspects. First, our results suggest that poor segmental phonological awareness is not only evident in non-tonal but also tonal language speakers with amusia. Second, our study further showed that tonal language speakers with amusia have poor suprasegmental phonological awareness as well. From a theoretical perspective, this study supports the TFS in the context of amusia, by showing that adults with amusia also exhibit rise time processing deficits which may explain their poor phonological awareness and reading difficulties. Taken together, this study highlights the shared role of auditory processing for both music perception and reading.

Keywords: congenital amusia, auditory processing, phonological awareness, reading difficulty

Table 1 Stepwise Regressions Showing the Unique Variance in Phonological Awareness Accounted for by the Auditory Measures

Step	Predictors	Onset phoneme deletion		Spoonerism		Tone oddity	
		β	ΔR^2	β	ΔR^2	β	ΔR^2
1	Age	-0.06	0.06	0.03	0.05	0.24	0.12
	Gender	0.24		0.22		0.29	
2	Typing method	0.23	0.04	0.22	0.04	0.19	0.03
3	Auditory processing						
	Pitch discrete	-0.6**	0.34**	-0.62**	0.36**	-0.66**	0.40***
	Pitch glide	-0.12	0.01	-0.22	0.04	-0.32	0.07
	Rise time	-0.24	0.05	-0.31	0.09	-0.55**	0.28**

Table 2 Stepwise Regressions Showing the Unique Variance in Literacy Accounted for by the Auditory Measures

Step	Predictors	Word reading		One minute reading		Dictation	
		β	ΔR^2	β	ΔR^2	β	ΔR^2
1	Age	0.13	0.03	0.25	0.18	0.19	0.11
	Gender	0.14		0.38		-0.24	
2	Typing method	0.03	0.00	0.01	0.00	-0.03	0.00
3	Auditory processing						
	Pitch discrete	-0.63**	0.36**	-0.33	0.10	-0.3	0.08
	Pitch glide	0.1	0.01	-0.18	0.02	-0.19	0.03
	Rise time	-0.6**	0.35**	-0.41*	0.16*	0.28	0.07

Table 3 Stepwise Regressions Showing the Unique Variance in reading Accounted for by the Auditory Measures after Controlling for Phonological Measures

Step	Predictors	Word reading		one minute reading	
		β	ΔR^2	β	ΔR^2
1	Age	0.13	0.03	0.25	0.18
	Gender	0.14		0.38	
2	Typing method	0.03	0	0.01	0
3	Onset phoneme deletion	0.33	0.41*	0.01	0.08
	Spoonerism	-0.31		0.25	
	Tone oddity	0.74*		0.06	
4	Auditory processing				
	Pitch discrete	-0.33	0.04	-0.31	0.04
	Pitch glide	0.33	0.07	-0.11	0.01
	Rise time	-0.34	0.07	-0.47	0.14

For all tables: * $p \leq .05$, ** $p < .01$, *** $p < .001$

References

- Couvignou, M., & Kolinsky, R. (2021). Comorbidity and cognitive overlap between developmental dyslexia and congenital amusia in children. *Neuropsychologia*, 155, 107811. <https://doi.org/10.1016/j.neuropsychologia.2021.107811>
- Couvignou, M., Peretz, I., & Ramus, F. (2019). Comorbidity and cognitive overlap between developmental dyslexia and congenital amusia. *Cognitive Neuropsychology*, 36(1–2), 1–17. <https://doi.org/10.1080/02643294.2019.1578205>
- Goswami, U., Wang, H.-L. S., Cruz, A., Fosker, T., Mead, N., & Huss, M. (2011). Language-universal Sensory Deficits in Developmental Dyslexia: English, Spanish, and Chinese. *Journal of Cognitive Neuroscience*, 23(2), 325–337. <https://doi.org/10.1162/jocn.2010.21453>
- Jones, J. L., Lucker, J., Zalewski, C., Brewer, C., & Drayna, D. (2009). Phonological processing in adults with deficits in musical pitch recognition. *Journal of Communication Disorders*, 42(3), 226–234. <https://doi.org/10.1016/j.jcomdis.2009.01.001>
- Liu, F., Patel, A. D., Fourcin, A., & Stewart, L. (2010). Intonation processing in congenital amusia: Discrimination, identification and imitation. *Brain*, 133(6), 1682–1693. <https://doi.org/10.1093/brain/awq089>
- Nan, Y., Sun, Y., & Peretz, I. (2010). Congenital amusia in speakers of a tone language: Association with lexical tone agnosia. *Brain*, 133(9), 2635–2642. <https://doi.org/10.1093/brain/awq178>
- Peretz, I., Ayotte, J., Zatorre, R. J., Mehler, J., Ahad, P., Penhune, V. B., & Jutras, B. (2002). Congenital amusia: A disorder of fine-grained pitch discrimination. *Neuron*, 33(2), 185–191. [https://doi.org/10.1016/s0896-6273\(01\)00580-3](https://doi.org/10.1016/s0896-6273(01)00580-3)
- Sun, Y., Lu, X., Ho, H. T., & Thompson, W. F. (2017). Pitch discrimination associated with phonological awareness: Evidence from congenital amusia. *Scientific Reports*, 7(1), Article 1. <https://doi.org/10.1038/srep44285>
- Tillmann, B., Rusconi, E., Traube, C., Butterworth, B., Umiltà, C., & Peretz, I. (2011). Fine-grained pitch processing of music and speech in congenital amusia. *The Journal of the Acoustical Society of America*, 130(6), 4089–4096. <https://doi.org/10.1121/1.3658447>

Gender Differences in the Neural Processing of Prosodic Cues in Information-Seeking and Rhetorical Questions

Mariya Kharaman¹, Bettina Braun², Carsten Eulitz³
^{1,2,3}University of Konstanz

Successful communication requires not only an understanding of the literal meaning of an utterance but also, often, the interpretation of implied meanings, which can be particularly challenging. Rhetorical questions, for instance, suggest a particular answer without explicitly stating it, e.g., *Who likes paying taxes?* implied answer: *Nobody does!* (Bayer and Obenauer, 2011; Biezma and Rawlins, 2017). The ability to interpret such implied meanings can be influenced by various factors, including linguistic, prosodic, cultural, social, and biological aspects. The present EEG study provides neurobiological insights into how prosody and biological gender influence the temporal dynamics of prosodic processing in information-seeking (ISQs) and rhetorical questions (RQs) in German, using an expectancy violation paradigm (Paulmann et al., 2012) with visually cued contexts.

Participants were presented with a visual cue that triggered either an ISQ or RQ interpretation (a watch as a mnemonic for ISQs – *What time is it?*; a pile of coins as a mnemonic for RQs – *Who likes paying taxes?*). After that they heard a lexically ambiguous *wh*-question, intonationally compatible with ISQ or RQ interpretations and varying in voice quality. The questions were identical to the stimuli in Kharaman et al. (2019) and Neitsch et al. (2018): 32 sentences, each produced by a female speaker in both ISQ and RQ prosodic realizations, with breathy and modal voice quality on the *wh*-word. The “tonal ISQ” featured a rise on the *wh*-word, a high transition and an early-peak accent on the final noun, while the “tonal RQ” exhibited a rising-falling medial-peak accent on the noun only. Finally, participants had to decide whether the auditory stimulus matched the visual cue (Fig.1 for details). Stimuli were presented in 256 randomized trials across 32 blocks, with consistent context within each block. We expected a prosodic expectancy positivity (PEP) for mismatched conditions, where the prosodic realization conflicted with the visually cued context. Statistical differences between illocution types and contexts were assessed for groups of 17 females (mean age 23.06, SD = 2.46) and 16 males (mean age 25.38, SD = 3.59) using non-parametric Monte Carlo cluster-based permutation tests (2000 randomizations, $\alpha < 0.05$). Tests were conducted separately for tonal ISQs and RQs, with predefined time windows of 500–600 ms and 1400–1600 ms. In our stimuli, the early time window reflects the processing of the object noun, while the late time window represents the final integration of all available prosodic cues within the utterance. The results reveal similar processing of tonal ISQs and gender-dependent differences in the processing of tonal RQs. Both female and male listeners exhibited a prosodic expectancy positivity (PEP) for ISQs during the early time window (left part of Fig.2). However, for RQs, this positivity was observed only in female listeners during the late time window, but not in male listeners (right part of Fig.2).

This pattern highlights a distinct temporal and gender-dependent sensitivity to prosodic cues. A mismatch between a visually cued context and the prosodic realization of an utterance resulted in an earlier PEP for ISQs (in both genders) compared to RQs (only in females). One interpretation for why participants detected the mismatch earlier in ISQs may be that the rising accent on the *wh*-word and the high transition precludes a rising-falling nuclear accent on the final noun, typical of RQs. Hence the prenuclear part in tonal ISQs may have been sufficient for identifying ISQs and detecting mismatches with the cued context earlier. The absence of this accent in tonal RQs delayed the detection of expectancy violations until the nuclear rising-falling accent was heard. Interestingly, both male and female participants showed the PEP when listening to tonal ISQs, but only female participants when listening to RQs. These results

provide valuable insights into the temporal dynamics of prosodic processing and its gender-dependent differences. However, further research with larger and more diverse samples is necessary to confirm these findings and refine our understanding of the biological and cognitive factors influencing prosodic interpretation.

Figures

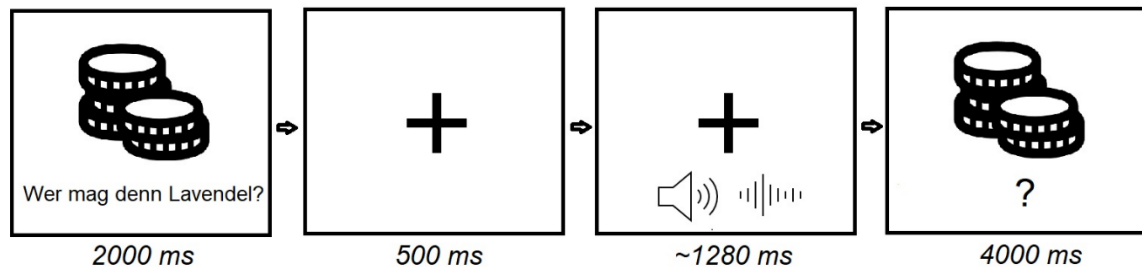


Fig. 1: A visual cue and a written sentence are presented for 2 sec, followed by a 500 ms fixation cross; a sentence is played while the fixation cross remains on screen; a visual cue and a question mark are displayed until a response is given, or for up to 4 sec if no response is provided.

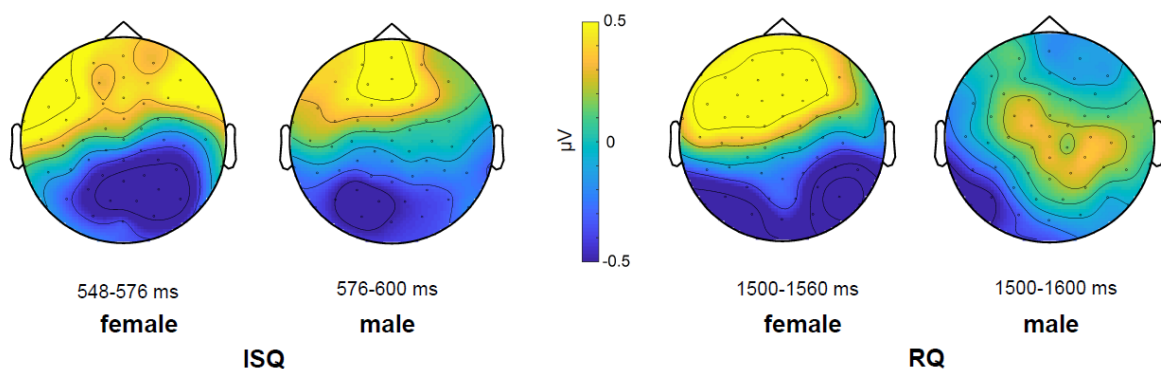


Fig.2: ERP topographies showing differences for ISQs and RQs (incoherently minus coherently cued contexts), for both female and male groups, in the time windows where significant effects were observed. Note that the differences for RQs were not significant in the male group. For the male group, the figure shows the topography for RQs difference in the window of interest, where the female group exhibited a coherence effect.

References

- Bayer, J. and Obenauer, H.-G. (2011). Discourse particles, clause structure, and question types. *The Linguistic Review*, 28:449–491.
- Biezma, M. and Rawlins, K. (2017). Rhetorical questions: Severing asking from questioning. In *Semantics and linguistic theory*, pages 302–322.
- Kharaman, M., Xu, M., Eulitz, C., and Braun, B. (2019). The processing of prosodic cues to rhetorical question interpretation: Psycholinguistic and neurolinguistics evidence. In *Proceedings of Interspeech*.
- Neitsch, J., Braun, B., and Dehé, N. (2018). The role of prosody for the interpretation of rhetorical questions in German. In *Proceedings of the International Conference on Speech Prosody*.
- Paulmann, S., Jessen, S., and Kotz, S. A. (2012). It's special the way you say it: An ERP investigation on the temporal dynamics of two types of prosody. *Neuropsychologia*, 50(7):1609–1620.

The Productivity of the Lexical Tone Sandhi in Wenzhou Wu Chinese

Weijun Zhang, Huangyang Xie, Peggy Pik Ki Mok

The Chinese University of Hong Kong, Hong Kong

weijunzhang@cuhk.edu.hk, huangyangxie@cuhk.edu.hk, peggymok@cuhk.edu.hk

The productivity of tone sandhi in different Chinese languages has received much attention. Some simple and transparent tone sandhi patterns like the T3 sandhi in Standard Mandarin and the rightward spreading in Shanghai Wu were shown to be fully productive [1, 2, 3]. In contrast, some complex and opaque tone sandhi patterns such as the circular sandhi chain in Taiwanese Southern Min were claimed to be largely unproductive [4, 5]. Differences in productivity were also found in variable tone patterns even within the same language (e.g., the extension and substitution sandhi rules in Wuxi Wu) [6]. Wenzhou Wu, a dialect of Wu Chinese, has complex word-level tone sandhi patterns, in addition to a simple phrase-level sandhi rule [7]. Its lexical tone sandhi is generally right-dominant. The patterns could be non-local (the sandhi tone does not necessarily occur within a syllable, i.e., spreading involved) and non-structure-preserving (the sandhi tone could be out of the citation tone inventory).

The current study tested the productivity of all disyllabic lexical tone sandhi patterns in Wenzhou Wu. Twelve middle-aged native speakers participated in picture-naming tasks without the words written, using four types of disyllabic stimuli: 1) real words presented in a single picture, 2) real words presented in two pictures (one for each monosyllabic morpheme), 3) pseudo words combining two existing morphemes also in pictures, and 4) words consisting of an accidental gap syllable (represented by a generated fictional image with the audio) and an existing monosyllabic morpheme in picture (henceforth ‘AG words’).

The results illustrated in Figures 1 and 2 indicated that Wenzhou Wu speakers could apply correct lexical tone sandhi to nearly all real words and most pseudo words, contrasting that the productivity significantly decreased with AG words ($p < .001$). Notably, for AG words, there was a clear distinction between the two types of sandhi patterns: the correct sandhi rates were significantly higher ($p < .001$) when the tone combination shares the same lexical and phrasal tone sandhi patterns (Figure 2) than that with a different lexical sandhi rule from the simple phrasal sandhi (Figure 1). In addition, the strategies adopted by the native speakers when encountering difficulties in lexical tone sandhi application were particularly noteworthy. Using citation tones for both syllables of disyllabic words was highly marked and disfavoured in Wenzhou Wu. When lexical tone sandhi was not correctly applied, native speakers broadly either produced a wrong but attested sandhi tone or treated the AG word as an intonational phrase, thus applying the simple phrasal tone sandhi instead.

Besides, given that the tone sandhi of Wenzhou Wu shows a pattern of convergence and (partial) neutralization of the tone distinctions on the sandhi position, the effect of the phonetic similarity between the base tones and the sandhi tone was also investigated. The results revealed that matching contours (both rising or both falling) between the base and the sandhi tones significantly facilitated the application of correct lexical sandhi, as shown in the patterns of T5 + T3 in Figure 3 (both the citation T5 and the first syllable of the sandhi pattern carrying a falling contour referring to Table 1) and T3 + T7 in Figure 4 (*mutatis mutandis*, for a rising contour). Opposite contours between the base and the sandhi tones would lead to higher rates of keeping the citation tones (e.g., T3 + T3 and T5 + T7).

Overall, this study showed that violations of phonotactic constraints (in the case of AG words) would intervene in the lexical sandhi application in Wenzhou Wu, whereas the lack of lexical representations (pseudo words) would not have a significant effect. Both the phonological complexity of the sandhi patterns and the phonetic similarity between the base tone and the sandhi form were crucial factors in tone sandhi application.

Keywords: tone sandhi, productivity, phonological complexity, Wu Chinese

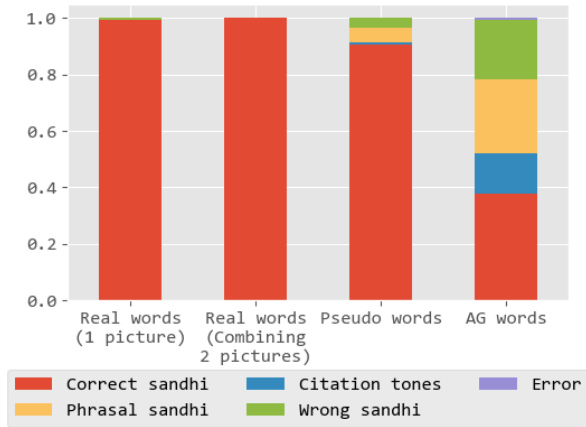


Figure 1: Tone patterns of the tone combinations with different lexical and phrasal tone sandhi.

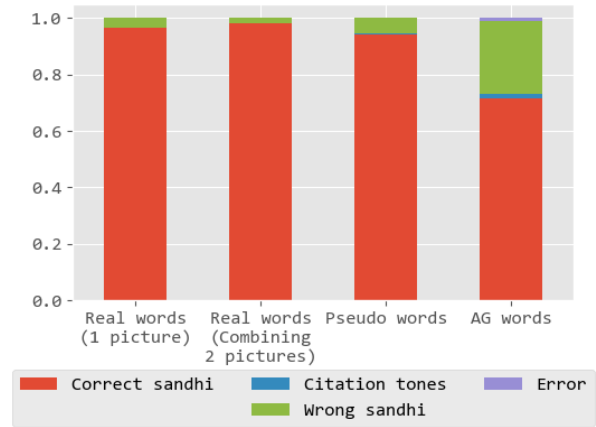


Figure 2: Tone patterns of the tone combinations with the same lexical and phrasal tone sandhi.

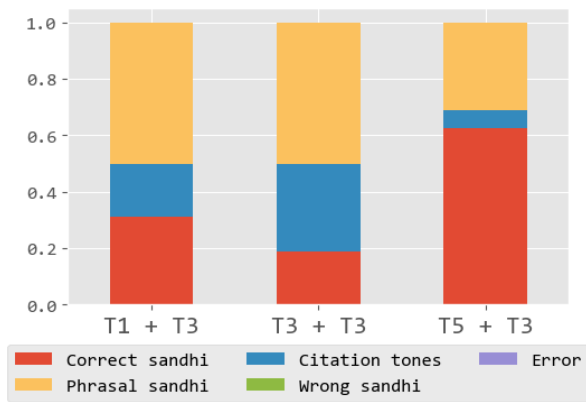


Figure 3: Tone patterns of the AG words with the [42 + 35] sandhi pattern.

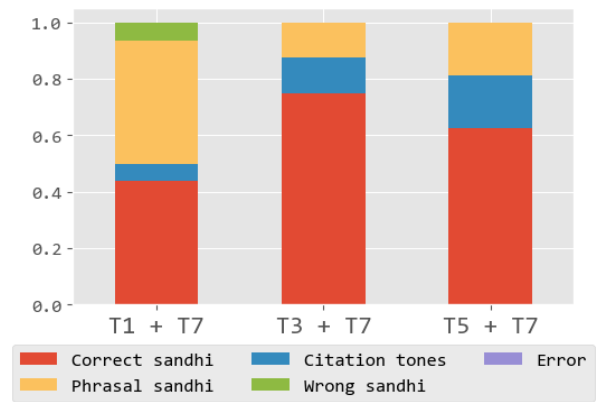


Figure 4: Tone patterns of the AG words with the [35 + 213] sandhi pattern.

$\sigma 1$ (sandhi position)	T1 [33]	T3 [34]	T5 [52]	T1 [33]	T3 [34]	T5 [52]
$\sigma 2$ (non-sandhi position)	T3			T7		
Sandhi pattern	42 + 35			35 + 213		

Table 1: The tone combinations with the two lexical sandhi patterns in Figures 3 and 4.

Selected references:

- [1] Chien, Y.-F., Sereno, J., & Zhang, J. (2016). Priming the representation of Mandarin Tone 3 sandhi words. *Language, Cognition and Neuroscience*, 31(2), 179–189.
- [2] Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, 27(1), 153–201.
- [3] Zhang, J., & Meng, Y. (2016). Structure-dependent tone sandhi in real and nonce words in Shanghai Wu. *Journal of Phonetics*, 54(1), 169–201.
- [4] Hsieh, H.-I. (1975). How generative is phonology. In Koerner, E. F. K. (eds.), *The Transformational-Generative Paradigm and Modern Linguistic Theory*. John Benjamins, 109–144.
- [5] Zhang, J., Lai, Y., & Sailor, C. (2011). Modeling Taiwanese speakers' knowledge of tone sandhi in reduplication. *Lingua*, 121(2), 181–206.
- [6] Yan, H., & Zhang, J. (2016). Pattern substitution in Wuxi tone sandhi and its implication for phonological learning. *International Journal of Chinese Linguistics*, 3(1), 1–45.
- [7] Chen, M. Y. (2000). *Tone sandhi: Patterns across Chinese dialects*. Cambridge: Cambridge University Press.

The role of pitch contours in Vietnamese learners' perception of Japanese vowel lengths

Trang Le Thi Huyen¹, Mariko Kondo¹, Yixuan Huang¹

¹Waseda University

trangle@moegi.waseda.jp, mkondo@waseda.jp, yixuan.huang@moegi.waseda.jp

Keywords: pitch contour, vowel length, Japanese, Vietnamese, cross-linguistic influence

In Japanese, vowel length is phonemically contrastive, and the pitch patterns associated with long and short vowels vary depending on word position [1, 2]. Previous studies have shown that native Japanese speakers use pitch as a secondary cue for identifying vowel length when duration is ambiguous [3]. In contrast, Vietnamese is a tonal language in which pitch conveys purely lexical tone without any indication of vowel length [4]. This study investigates whether the tonal background of Vietnamese learners' first language (L1) facilitates or interferes with the perception of Japanese vowel length, and whether this is affected by the type of pitch contour.

Twenty minimal pairs of real Japanese words (four contrast types x five word pairs), were selected as stimuli (Table 1). Each contrast type features long vowels with distinct pitch contours: either a pitch drop (HL) or a pitch rise (LH), both in the word-initial position, or flat pitch contours (LL or HH) in the word-final position. The stimuli were recorded and then presented to participants in isolation to eliminate contextual influences on the pitch contours.

Eighty Vietnamese learners of Japanese participated, categorised into three groups by Japanese Language Proficiency Test level: N3 (n = 21, pre-intermediate), N2 (n = 42, intermediate), and N1 (n = 17, advanced). The participants were selected exclusively from Northern Vietnam to control regional tonal variations. In addition, a control group of 13 native Japanese speakers (NJ) from the Kanto region was also included. The experiment used a two-alternative forced-choice identification task, conducted online via the Gorilla platform. The participants were presented with pairs of target words written in *Hiragana* while they listened to an auditory stimulus of one of the paired words, and they had to choose whether it was the word containing a short or a long vowel. Each participant completed 252 trials, comprising 240 experimental trials (4 contrast types × 5 minimal pairs × 2 words × 6 repetitions) and 12 practice trials.

The data were analyzed using two-way ANOVA, and significant effects of both pitch ($F(3, 356) = 14.13, p < .001$) and proficiency ($F(3, 356) = 16.68, p < .001$) on identification accuracy were found. To further examine misperception patterns, a three-way ANOVA was conducted on misheard rates, revealing significant main effects of contrast type ($F(3, 688) = 18.2, p < .001$), proficiency ($F(3, 688) = 20.45, p < .001$) and misheard type ($F(1, 688) = 14.1, p < .001$). There were significant two-way interactions between both proficiency and misheard type ($F(3, 688) = 2.95, p < .05$) and between misheard type and contrast type ($F(3, 688) = 25.37, p < .001$), as well as a three-way interaction ($F(9, 688) = 2.04, p < .05$).

Post hoc analysis showed that the Vietnamese listeners had the greatest difficulty in accurately identifying vowel length in Final.L–LL contrast (Figure 1). This finding is consistent with results from previous Japanese language vowel length perception studies with L1 English [5] and L1 Vietnamese learners [6]. A common error among Vietnamese listeners in the current study was the misperception of long vowels with a flat low pitch as short vowels (Figure 2). This may have been due to the limited pitch movement and low pitch height. Additionally, since it appeared exclusively in word-final position, natural pitch declination may have further obscured the perception of vowel length. In contrast, the long vowels with dynamic pitch contours (i.e., pitch rise or drop) or a flat high pitch were more accurately identified as long. Notably, in the Final.H–HH condition, although the long vowel also exhibited a flat pitch contour, its higher pitch range may have provided clearer acoustic cues, thereby enabling more accurate perception, as suggested in [7]. These findings suggest that Vietnamese learners' perception of Japanese vowel length is influenced by both pitch height and pitch range, with accuracy varying depending on pitch conditions.

These pitch-related effects also varied with listener proficiency. N1 learners approached native-like performance in the more favorable contrast conditions (Initial.H–HL and Initial.L–LH). However, under the most challenging condition (Final.L–LL) there were no significant differences between the three proficiency groups. It is likely that, for Vietnamese listeners, proficiency alone is not sufficient to ensure accurate identification of Japanese vowel length when there is low pitch and minimal pitch movement.

Table 1. Examples of minimal pairs of Japanese words used in the 2AFC task with target sounds underlined.

Contrast	Initial.H-HL		Initial.L-LH		Final.L-LL		Final.H-HH	
Position	Word-initial		Word-initial		Word-final		Word-final	
Duration	Short	Long	Short	Long	Short	Long	Short	Long
Pitch contour	H	HL	L	LH	L	LL	H	HH
Test word	<i>k<u>a</u>do</i>	<i>ka<u>a</u>do</i>	<i>y<u>o</u>koo</i>	<i>y<u>oo</u>koo</i>	<i>k<u>u</u>ro</i>	<i>ku<u>r</u>oo</i>	<i>cho<u>o</u>bo</i>	<i>cho<u>oo</u>bo</i>
Meaning	corner	card	manu- script	sunlight	black	hardship	account book	view

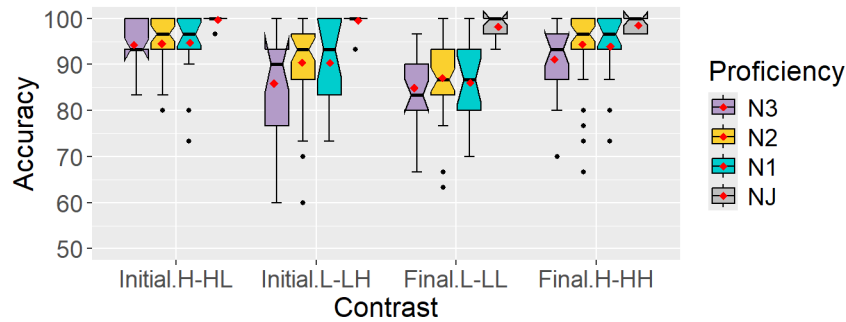


Figure 1. Identification accuracy score (%) of Japanese long and short vowels by four listener groups (N3, N2, N1, and NJ) across four contrast types. The horizontal line indicates the median accuracy, and the red diamond marks the mean accuracy for each group and contrast.

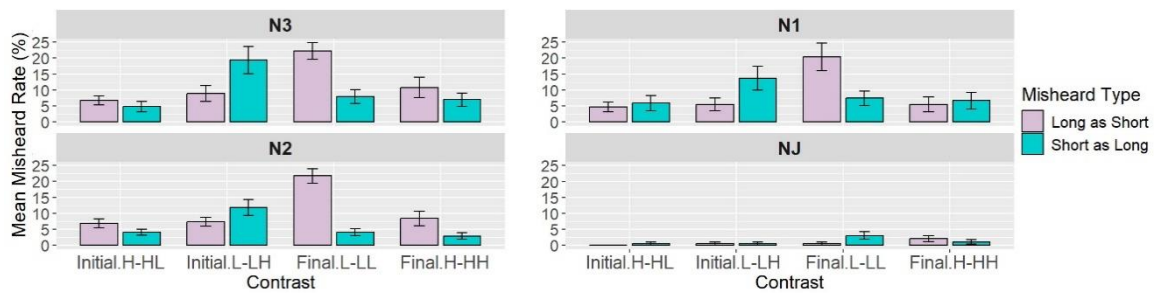


Figure 2. Mean misheard rate (%) in the identification of Japanese long and short vowels by four listener groups (N3, N2, N1, and NJ) across four contrast types.

References

- [1] Vance, T. J. (2008). *The sounds of Japanese*. Cambridge University Press.
- [2] Deguchi, M. (2015). The role of pitch contours in teaching vowel length distinctions in Japanese. *Pronunciation in Second Language Learning and Teaching Proceedings*, 6(1).
- [3] Hui, C. T. J., & Arai, T. (2020). Pitch and duration as auditory cues to identify Japanese long vowels for Japanese learners. *Acoustical Science and Technology*, 41(5), 796–799.
- [4] Kirby, J. (2011). Illustrations of the IPA: Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association*, 41(3), 381–392.
- [5] Oguma, R. (2000). Perception of Japanese long vowels and short vowels by English-speaking learners. *Japanese Language Education around the Globe*, 10, 43–54.
- [6] Do, H. N. (2015). Issues regarding the perception of long vowels and geminate consonants in Japanese among Vietnamese learners. *VNU Journal of Foreign Studies*, 31(2), 31–38.
- [7] Gussenhoven, C., & Zhou, W. (2013). Revisiting pitch slope and height effects on perceived duration. *Interspeech 2013: 14th Annual Conference of International Speech Communication Association*, 1365–1369.

A Preliminary Study on the Productivity of Tone Sandhi in the Baotou Jin Dialect by Child and Adult Speakers

Xinyue LIU, Peggy MOK

The Chinese University of Hong Kong

xinyueliu@cuhk.edu.hk, peggymok@cuhk.edu.hk

Keywords: *tone sandhi; productivity; accidental gap syllable; language contact; Baotou Jin*

This study tested the application of two types of tone sandhi rules in the Baotou Jin dialect in two age groups (children vs. adults), and the results showed different application accuracy regarding the two sandhi rules and various application strategies of the speakers in the two age groups.

As the largest city by population in Inner Mongolia, China, Baotou belongs to the Da-Bao Jin-speaking area. The Baotou Jin dialect has five lexical tones: T1 [24], T2 [44], T3 [312], T4 [52], and T5 [ʔ4] (checked tone). There are mainly two tone sandhi patterns in the Baotou Jin dialect: T1 sandhi and T3 sandhi [1]. Baotou T1 sandhi occurs when there comes two T1 [24], and the second T1 [24] becomes T2 [44] (e.g., [fei]24 + [tei]24 → [fei]24 + [tei]44 “plane”). Baotou T3 sandhi is very similar to the Mandarin T3 sandhi: in a sequence of two T3 [312], the first T3 [312] changes to the rising T1 [24] (e.g., [ma]312 + [ji]312 → [ma]24 + [ji]312 “ant”). Baotou T1 sandhi is a left-dominant sandhi with an initial tone that spreads rightward, while Baotou T3 sandhi is a right-dominant sandhi with a non-final tone that undergoes a local paradigmatic change [2].

The application of tone sandhi to novel materials involves the tacit knowledge of the sound system regarding the “productivity” of tone sandhi. Following the nonce-probe test design [3, 4, 5], both real words and novel words like pseudowords and words with accidental gap syllables (AG words) were tested in this study. A picture-naming task with four conditions was used: 1) disyllabic real words with one picture (Rw1P); 2) disyllabic real words combining two pictures (Rw2P); 3) pseudowords (Pw) which were meaningless combinations of two real syllables; 4) AG words (AGw) each with an AG syllable (represented by a bizarre picture) and a real syllable. Examples of the pictures and words used in the experiment are shown in Figure 1. Each condition consisted of 5 disyllabic words with T1 sandhi and another 5 with T3 sandhi. Two groups of participants were tested: 14 sex-balanced children (aged 5 to 12), and 10 sex-balanced adults (aged 39 to 60). The speech production of the participants was recorded and manually checked. Each token’s accuracy of tone sandhi application was auditorily judged by two Baotou Jin native speakers with phonetic training, and the inter-rater consistency was around 94%. There were three types of accuracy judgments: sandhi (the correct sandhi tone), citation (the citation tone), and wrong (neither “sandhi” nor “citation”).

The mean accuracy of the tone sandhi application is shown in Figure 2. In Rw1P and Rw2P, the accuracy was almost 100% for both T1 and T3 sandhi in both age groups. The Pw test showed a few mistakes, mainly in the T1 sandhi produced by children. The most disparate results were from the AGw test. The AGw with Baotou T3 sandhi showed an overall high accuracy of correct sandhi across the two age groups. In contrast, the accuracy of AGw with Baotou T1 sandhi in both age groups was much lower than that with Baotou T3 sandhi. The AGw with T1 sandhi also showed significant individual variation (as illustrated in Figure 3), which was not found in those with T3 sandhi. The high accuracy even for AG words in Baotou T3 sandhi is very similar to the findings of Mandarin T3 sandhi [6, 7, 8]. This result might be interpreted by the similarity of the T3 sandhi patterns in both Baotou Jin and Mandarin and the frequent contact between the two languages in Baotou, which has enhanced the productivity of Baotou T3 sandhi. On the other hand, the uniqueness of Baotou T1 sandhi as a left-dominant sandhi might further weaken its productivity.

The wrong production had very consistent patterns. Firstly, the wrong productions in T1 sandhi were mostly 24+312 (i.e., the surface form of T3 sandhi) with very few exceptions. This wrong T1 sandhi pattern of 24+312 seemed to result from the wrong application of the T3 sandhi, which had much higher productivity than T1 sandhi. Secondly, the wrong productions in T3 sandhi were mostly 44+312, and this consistently wrong T3 sandhi pattern was not found in previous studies on Mandarin T3 sandhi [6, 7, 8].

As for the differences between the two age groups, the results showed that children tended to keep the citation tones more than using the wrong sandhi, while the adults preferred to change the target tones and even applied the wrong sandhi patterns (i.e., 44+312 for T3 sandhi and 24+312 for T1 sandhi). The results suggest that the children’s phonological awareness of tone sandhi was lower than that of the adults. Nevertheless, two children aged 7 could reach 100% accuracy of T1 sandhi application in AG words, while only one adult could achieve it (Figure 3). Clearly, data from more participants are needed to further explore the differences in productivity between the two sandhi patterns in Baotou Jin, and for a better understanding of the tone sandhi mental mechanisms [9, 10].

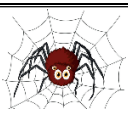
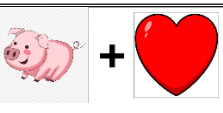


 Part 1: Real words with one picture tsi24tsu24 = tsi24tsu44 “spider”	 Part 2: Real words with two pictures tsu24 + ɛiŋ24 → tsu24ɛiŋ44 “pig heart”
 Part 3: Pseudowords tsu24 + ts ^h ɿ24 → tsu24ts ^h ɿ44 “pig”+ “car”	 Part 4: AG words with recordings tsu24 + t ^h ɿ24 → tsu24 t ^h ɿ44 “pig”+ AG syllable

Figure 1: Example of pictures and words with T1 sandhi in the four conditions of the picture-naming task

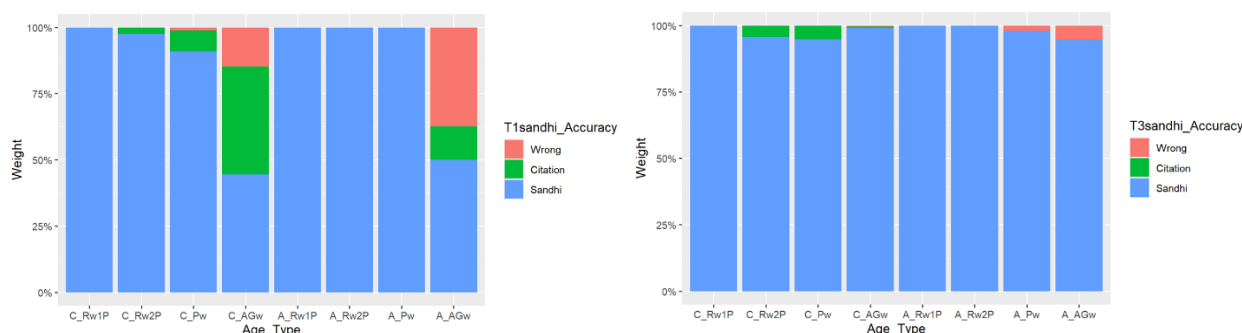


Figure 2: Mean accuracy of tone sandhi application for real words with one and two pictures (Rw1P and Rw2P), pseudoword (Pw), and AG word (AGw) by age groups (C_children, A_adults) and sandhi rules.

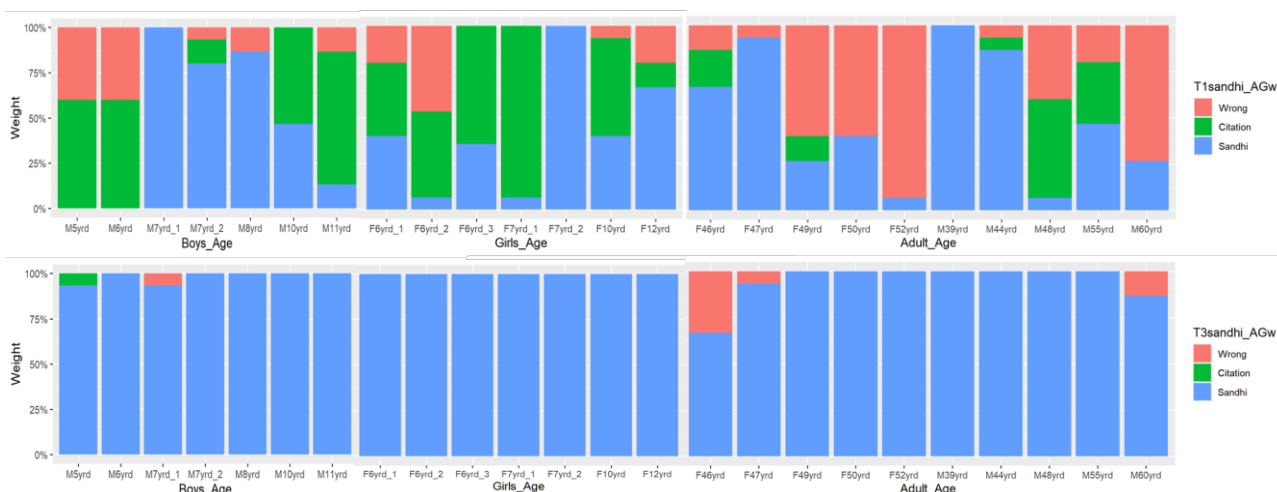


Figure 3: Individual accuracy of AG words with T1 and T3 sandhi by all speakers (M5yrd = 5-year-old male)

Selected references

- [1] Cai, W., & Zhang, Y. (2017). Phonological overview and the phonological diversities in the old vs. the new styles of the Dongsheng City, *Journal of Language and Literature Studies*, 37(1), 97-105. [2] Zhang, J. (2022). Tonal processes defined as tone sandhi. *The Cambridge handbook of Chinese linguistics* (Chapter 14), 291-312. Cambridge, UK: Cambridge University Press. [3] Zhang, J. (2016). Using nonce-probe tests and auditory priming to investigate speakers' phonological knowledge of tone sandhi. In *Proceedings of the 5th International Symposium on Tonal Aspects of Languages* (pp. 12-18). [4] Li, X., & Mok, P. P. K. (2020). The acquisition of Xiamen tone sandhi by children. In *Proceedings of the 10th International Conference on Speech Prosody 2020* (pp. 479-483). [5] Ge, C., & Mok, P. (2024). The effect of phonotactic constraints on tone sandhi application: A cross-sectional study of Xiamen Min. In *Proceedings of the 12th International Conference on Speech Prosody 2024*. [6] Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, 27(1), 153-201. [7] Zhang, C., & Peng, G. (2013). Productivity of Mandarin third tone sandhi: a wug test. *Eastward flows the great river: Festschrift in honor of Prof. William S.Y. Wang on his 80th birthday*, (pp. 256-282). [8] Huang, X., Zhang, G., & Zhang, C. (2018). A preliminary study on the productivity of Mandarin T3 sandhi in Mandarin-speaking children. In *Proceedings of the 6th International Symposium on Tonal Aspects of Languages*. [9] Zhang, C., Xia, Q., & Peng, G. (2015). Mandarin third tone sandhi requires more effortful phonological encoding in speech production: Evidence from an ERP study. *Journal of Neurolinguistics*, 33, 149-162. [10] Zhang, J., Zhang, C., Politzer-Ahles, S., Pan, Z., Huang, X., Wang, C., Peng, G. & Zeng, Y. (2022). The neural encoding of productive phonological alternation in speech production: Evidence from Mandarin Tone 3 sandhi. *Journal of Neurolinguistics*, 62, 101060.

Bulgarian Judeo-Spanish Calling Contours: A Cluster Analysis

Mitko Sabev¹, Bistra Andreeva¹, Christoph Gabriel², Jonas Grünke³

¹*Saarland University*, ²*Johannes Gutenberg University Mainz*, ³*University of Regensburg*

In a production experiment, 17 bilinguals (6 female, mean age 87.33, SD 9.99; 11 male, aged 78.00 ± 9.32) performed a production task in Judeo-Spanish (JSp) and Bulgarian (BB), along with 14 Monolingual Bulgarian (MB) speakers (10 female, aged 78.91 ± 8.70; 4 male, aged 83.00 ± 10.49). They were asked to call a girl named /ka'lina/ from a distance in three contexts: neutral (for dinner), positive (to open her presents) and negative (for a telling-off). Fundamental frequency and vowel midpoint formant frequencies were measured in Praat [1]. Hierarchical clustering was applied to identify distinct f_0 -contour types, and an optimal number of clusters was determined based on information cost measures [2, 3]. A multivariate conditional inference tree [4] was constructed to assess how CLUSTER (contour type) and DURATION vary with situational CONTEXT, linguistic VARIETY and GENDER. Linear mixed effects models were used to compare F1 and F2 frequencies of initial- vs final-syllable /a/ vowels.

Cluster analysis and evaluation identified four optimal clusters, plotted in Fig. 1, and the tree in Fig. 2 shows that CLUSTER and DURATION are sensitive to all predictors – CONTEXT, GENDER and VARIETY. One contour dominates the negative context for all varieties and genders: cluster **A**, which starts with a slight rise into the stressed (second) syllable, followed by a steep and deep fall (H* L-L%). Neutral and positive contexts are not differentiated, but contours vary across genders and varieties. In all varieties, men use overwhelmingly contour **B**, an appreciable climb into the stressed syllable and then a smooth fall (L-H* !H-L%). Monolingual women use three contours in non-negative contexts: contour **B**, which is favoured by all men, along with the cross-linguistically well-attested calling tune **C** (L-H* !H-%), and contour **D**, a slow and rather shallow rise that doesn't peak until the final syllable (L+H* H-L%). In both their languages, bilingual women's first choice of non-negative contour is **B**, which is the one typical for men and one out of the three contours used by monolingual women. Intriguingly, the bilingual women's second choice is contour **A**, which is prevalent in negative contexts for everyone. Crucially, even though they may use the same contour shape in both negative and non-negative contexts, non-negative realisations have considerably longer durations, which substantially softens the call. One reason for bilingual women not employing more 'cheerful'-sounding contours in non-negative contexts may have to do with childhood stigma: several female Judeo-Spanish speakers pointed out—in separately recorded biographical interviews—that they were mocked by peers for their sing-song intonation.

Bulgarian is known for its pronounced and neutralising vowel reduction: when unstressed, the non-high vowels /a ɔ/ raise and merge with their higher counterparts, /ɤ u/; the process is categorical and not merely duration-driven [5]. After centuries of contact with Bulgarian, the local variety of Judeo-Spanish also exhibits appreciable reduction [6]. It has been observed that Bulgarian unstressed vowels resist reduction in the final syllables of calling phrases, which has been explained with the addition of an extra beat [7]. We report significantly higher F1 frequencies of /a/ in the final than in the initial syllables of all calling contours, confirming that a special post-lexical prominence is in place there, which causes unstressed vowels to behave like lexically stressed ones, retaining a degree of openness that no other unstressed vowels do.

Cluster analysis performed to objectively identify contour categories yielded different results from an analysis of a smaller data set based on auditory impression and f_0 -span measurements [8]: clearer effects of context, gender and variety were revealed, and an additional contour was identified. There is little variation in the marked, negative context. Men show little variation in non-negative contexts as well, while women have more varied repertoires: in addition to men's default non-negative contour, monolingual women use two extra tunes, while bilingual women supplement the default contour with a lengthened version of the negative tune.

Keywords: Bulgarian Judeo-Spanish, calling contours, cluster analysis, recursive partitioning, context and gender variation.

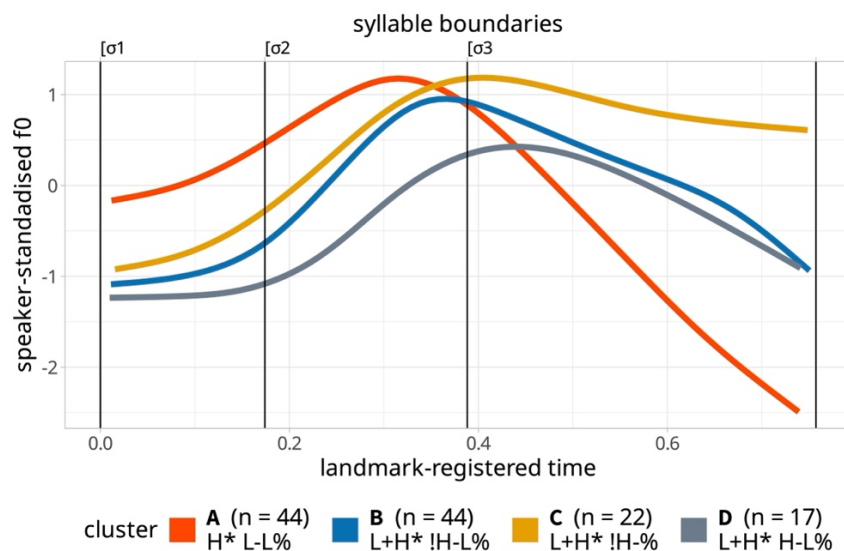


Figure 1. Contour types identified by cluster analysis (GAM curves).

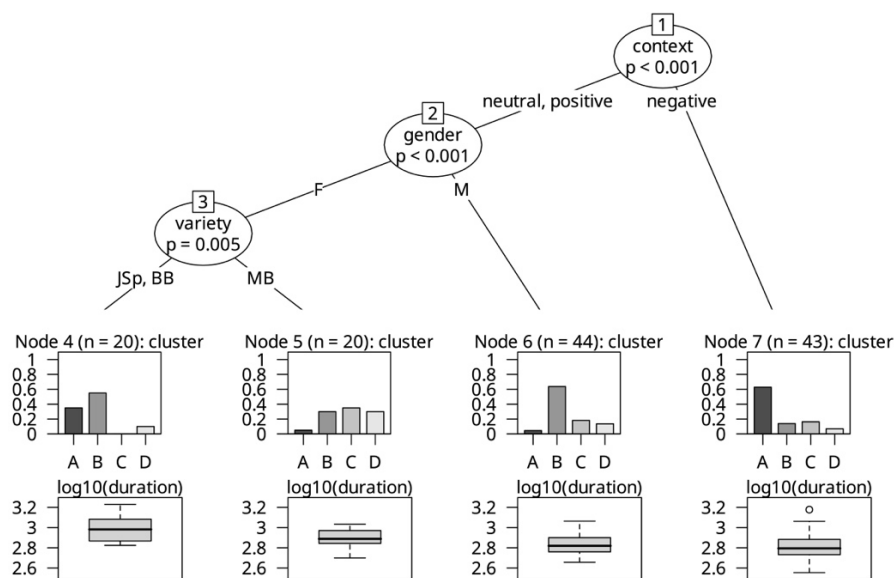


Figure 2. Conditional inference tree: $cluster + duration \sim context + variety + gender$

- [1] Boersma, P., & Weenink, D. 2024. *Praat: doing phonetics by computer* [Software], v. 6.4.17.
- [2] Kaland, C. 2023. Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *JIPA* 53(1), 159–188.
- [3] Kaland, C. & Ellison, T. M. 2023. Evaluating cluster analysis on f0 contours: an information theoretic approach on three languages. *Proc. 20th ICPhS*, 3448–3452.
- [4] Hothorn, T. & Zeileis, A. 2015. partykit: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research* 16, 3905–3909.
- [5] Sabev, M. 2023. Unstressed vowel reduction and contrast neutralisation in western and eastern Bulgarian: A current appraisal. *JPhon* 99, 101242.
- [6] Grünke, J., Sabev, M., Gabriel, C. & Andreeva, B. 2023. Vowel reduction in spontaneous Bulgarian Judeo-Spanish. *Proc. 20th ICPhS*, 2820–2824.
- [7] Andreeva, B. & Dimitrova, S. 2023. Vocative Intonation in Bulgarian. *Proc. 2nd Intl. Conf. on Tone and Intonation*, 29–33.
- [8] Grünke, J., Andreeva, B., Gabriel, C. & Sabev, M. 2023. Vocative Intonation in Language Contact: The Case of Bulgarian Judeo-Spanish. *Languages* 8(4), 284.

Melodic prominence of verbs in Czech spoken narratives

Jan Volín, Adléta Hanzlová, Michaela Svatošová – Institute of Phonetics, Prague

Key terms: prominence, full verb, accent-group, information structure, fundamental frequency

Verbs are often assigned a special status in the syntactic structure of sentences: they are regarded the central constitutive elements. However, in our previous study, we found quite a substantial number of autosemantic verbs (i.e., content verbs, not the auxiliary or modal ones) unaccented [1]. Our objective in the current study is to investigate relative prominence of the accented verbs in terms of their f_0 values and their phonological patterning relative to neighbouring accented words. Their association with theme/rheme status is also examined.

Professional recordings of narratives were obtained from audiobooks. A sample of 16 speakers (8f+8m) was collected with the requirement of at least 1000 words per speaker. Full (autosemantic) verbs were sought in phrase-internal positions, which is an arrangement typical of Czech [2]. Phrases with verbs in initial and final positions were omitted at this stage. We devised six measures pertaining to differences between a verb-related f_0 value and a reference value from the prosodic neighbourhood. The differences (D) can be sorted into three pairs: $DV1$, $DV2$ = differences of vowel based metrics; $DG1$, $DG2$ = differences of accent-group based metrics; $DS1$, $DS2$ = supplementary measures. $DV1$ is a plain difference between the accented syllable nucleus of the verb and the nearest preceding accented nucleus. For $DV2$, the accented nucleus of the verb is related to a line connecting the nearest preceding and nearest following accented syllable. The parameters of the line (gradient, intersect) and the timing moment were used to calculate where the verb value would be if it were directly on the connecting line (Fig. 1). $DG1$ was the difference between the mean f_0 of accent-group containing the verb and the preceding accent group. $DG2$ related the mean f_0 of the accent-group containing the verb to the average f_0 of accent-groups flanking the verb (Fig. 2). As to supplementary measures, $DS1$ was calculated identically to $DV2$, but with the values of the post-stress syllable nuclei, since a common stress-group pattern in Czech is L^*+H (post-stress rise, see [3]). Naturally, this measure was relevant only for consecutive L^*+H accents. Finally, to map the regular downtrends in Czech narratives, we also captured the difference between the mean f_0 of the accent-group with the verb and the following accent group. This measure ($DS2$) was also useful only in a subset of data. Apart from f_0 measures, we annotated phonological status of the three accented syllables in our sample phrases following [4]. (The verb was always between two other accents.) Also, the information status of the verbs was established. We agreed with [5] that despite certain transitional features of verbs, each can be classified as either thematic (Th) or rhematic (Rh) given its wider narrative co-text.

The material provided 473 verbs (about 30 verbs per speaker), of which 276 were accented with L^*+H and 197 with H^* . However, unlike [6: p. 9] we cannot claim that one or the other accent is more prominent in Czech. Moreover, our concern was the relative prominence in a phrasal neighbourhood. That could be established primarily for parallel phonological arrangements ($L^*L^*L^*$ or $H^*H^*H^*$ for short). These occurred in about 37% of the phrases, i.e., in 174 cases. Tables 1 and 2 display percentages of verbs with lower f_0 values for $L^*L^*L^*$ and $H^*H^*H^*$ phrases, respectively. The individual metrics clearly produce disparate results and $DG1$ produced the highest counts of verbs with lower f_0 prominence. How they relate to human perception needs to be established in follow-up research. Nevertheless, it is clear that verbs do not dominate in their immediate prosodic context. Another notable fact is that while in L^* condition the rhematic verbs produce higher counts of lower values, in H^* condition the situation is generally opposite.

The analyses of our data will be expanded. First, it remains to be established what the behaviour of our metrics was in $L^*L^*H^*$ and $H^*H^*L^*$ cases ($n = 100$). Second, the perceptual impact of verbs with reversed pattern in the context ($L^*H^*L^*$ or $H^*L^*H^*$) should be investigated ($n = 111$). Finally, although $DV2$ and $DV4$ took into account downtrends in prosodic phrases computationally, a perceptual model for Czech should be developed before any generalizations are made.

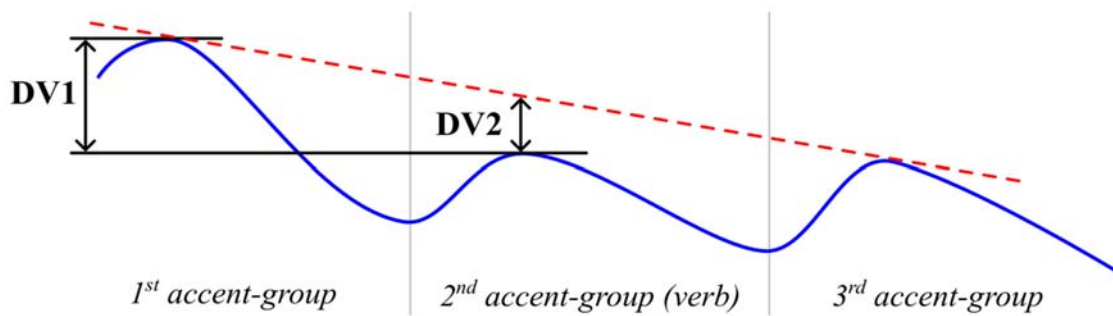


Fig. 1. An illustration of DV1 and DV2 metrics.

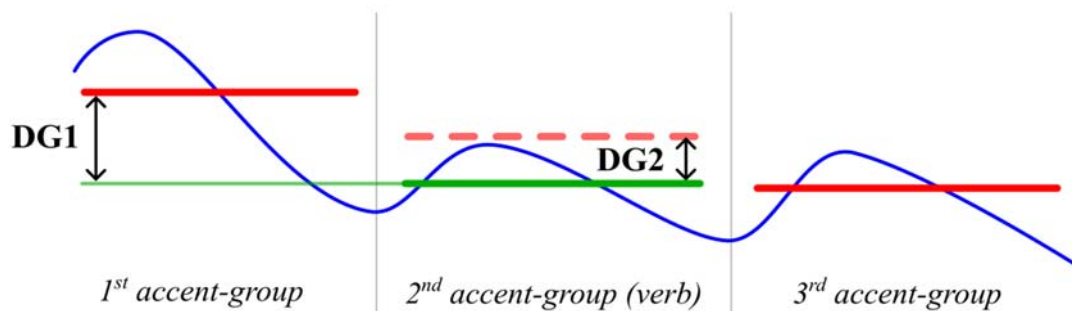


Fig. 2. An illustration of DG1 and DG2 metrics.

<i>L*L*L*</i>	<i>DV1</i>	<i>DV2</i>	<i>DG1</i>	<i>DG2</i>	<i>DS1</i>	<i>DS2</i>
THEMATIC (n = 40)	42.5	35.0	62.5	42.5	52.5	20.0
RHEMATIC (n = 73)	61.6	49.3	68.5	53.4	54.8	30.1

Table 1. Percentages of L* verbs with negative difference (i.e., lower) in f0 values split by information status. For explanation of metrics, see text.

<i>H*H*H*</i>	<i>DV1</i>	<i>DV2</i>	<i>DG1</i>	<i>DG2</i>	<i>DS1</i>	<i>DS2</i>
THEMATIC (n = 21)	85.7	85.7	90.5	76.2	71.4	47.6
RHEMATIC (n = 40)	90.0	67.5	90.0	45.0	45.0	12.5

Table 2. Percentages of H* verbs with negative difference (i.e., lower) in f0 values split by information status. For explanation of metrics, see text.

References

- [1] Volín, J. & Hanžlová, A. (2024). Deaccented verb as an element in the utterance information structure. *Proc. Speech Prosody 2024*, pp. 911-915.
- [2] Uhlířová, L. (1984). Sloveso určité v aktuálním členění větěném (The verb in functional sentence perspective). *Naše řeč*, 67/1, pp. 1-10.
- [3] Volín, J., & Skarnitzl, R. (2022). The impact of prosodic position on post-stress rise in three genres of Czech. *11th Speech Prosody* (Lisbon), pp. 505-509.
- [4] Beckman, M. E., & Ayers Elam, G. (1997). *Guidelines for ToBI labeling, version 3*. Ohio State University, retrieved from https://www.ling.ohio-state.edu/research/phonetics/E_ToBI/.
- [5] Sgall, P., Hajičová, E., & Buráňová, E. (1980). *Aktuální členění věty v češtině (Functional Sentence Perspective in Czech)*. Academia.
- [6] Im, S., Cole, J., & Baumann, S. (2023). Standing out in context: Prominence in the production and perception of public speech. *Laboratory Phonology*, 14/1, pp. 1-62.

Ethnic variation in perceptual sensitivity to stress-based f0 in Singapore English

Jasper Sim¹, Julia Schwarz², Katrina Kechun Li³, Adam Chong⁴

¹*Nanyang Technological University, Singapore;* ²*University of Cambridge*

³*Leiden University;* ⁴*Queen Mary University of London*

In mainstream English varieties, a lexically stressed syllable has higher f0 than an unstressed syllable. Contrastingly, due to long-term language contact and extensive structural changes to the native English variety spoken in Singapore [1] (Singapore English; SgE), f0 rises across single words in SgE regardless of the underlying lexical stress. In addition, stress-initial words are distinguished from non-stress-initial words through a higher mean f0 at word onset and possibly an earlier f0 elbow [2], [3] (Figure 1). Whether SgE speakers can also perceive these stress-based f0 patterns observed in their production, and consequently, whether these are used as discriminative cues in spoken word recognition and lexical access remains unclear [4].

In this pilot study, we asked (i) whether SgE-speaking listeners are perceptually sensitive to the f0 differences that distinguish SgE words with different underlying stress, and (ii) whether their ability to discriminate is modulated by individual bilingualism. We swapped the pitch contours of SgE disyllabic word pairs that have different underlying lexical stress but similar segmental properties (e.g. carTOON ↔ CARpet). We expected that SgE-speaking listeners would judge the stimuli with the original/expected pitch contour (Matching) as more typically Singaporean than those with the Mismatched pitch contour (Task 1). We also expected that, in a paired comparison task, listeners would show a preference for the Matching stimuli over the Mismatched ones as being more typical (Task 2).

Pilot participants were 16 native Singaporeans who were bilinguals of SgE and an ethnic mother tongue [EMT; Mandarin ($n=6$), Malay ($n=8$), or Tamil ($n=2$)]. They were between 21 and 36 years old ($M_{age}=28$, $SD_{age}=4.8$; males=10), with no reported hearing or language impairments. The stimuli were 15 SgE word pairs, produced by a male 32-year-old native English-Mandarin bilingual Singaporean. The smoothed f0 contours of the pre-synthesised stimuli were overall consistent with the previously reported production patterns (Figure 1; [2]). Participants completed the tasks on Gorilla.sc following a headphone screening test [5].

In Task 1, all stimuli (i.e., Matching and Mismatched) were rated as highly typically Singaporean on a seven-point likert scale, and the difference in ratings between Matching, $M(SD)=5.47(1.42)$, $Mdn=6$, and Mismatched stimuli, $M(SD)=5.53(1.30)$, $Mdn=6$, was small. Mixed-effects ordinal regression did not reveal significant differences in ratings between conditions (Matching, Mismatched) nor an effect of listener's EMT (Mandarin, Malay, Tamil). In Task 2 (paired comparisons), Matching stimuli were regarded as more typical on average 58% of the time. Mixed-effects logistic regression revealed that an interaction between lexical stress and listener's EMT significantly predicted outcomes (Figure 2): For stress-initial words only, Mandarin-speaking listeners were more likely to prefer the Matching stimuli compared to Malay-speaking ($b=1.02$, $p=0.005$) and Tamil-speaking listeners ($b=1.29$, $p=0.02$). Mandarin-speaking listeners were also more likely to prefer the Matching stimuli of stress-initial words than stress-final words (76% v. 54%, $b=0.97$, $p=0.004$; Figure 2).

Our preliminary findings suggest that stress-based f0 manipulations may not be particularly salient to some listeners, even when directly comparing Matching–Mismatched pairs, which indicate that stress-related f0 changes may not significantly influence naturalness, and/or that SgE speakers are highly tolerant of small pitch modulations. Interestingly, Mandarin bilinguals were much more sensitive to f0 manipulations, but only for stress-initial words (e.g. CARpet with carTOON's f0), which suggests a potential influence of tonal language background in prosodic processing in SgE. In our full-scale study, we will consider other social factors, their production patterns, and other correlates of stress in perception and also in online processing.

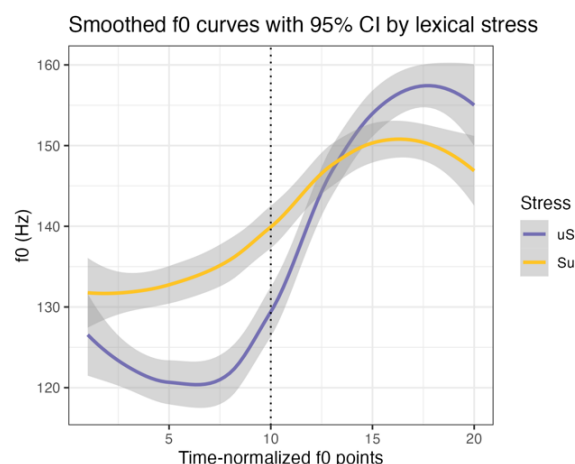


Figure 1: Time normalised plots of f_0 averaged over each stress condition (Su: stress-initial; uS: stress-final) for the disyllabic stimuli. Dotted line represents syllable boundary.

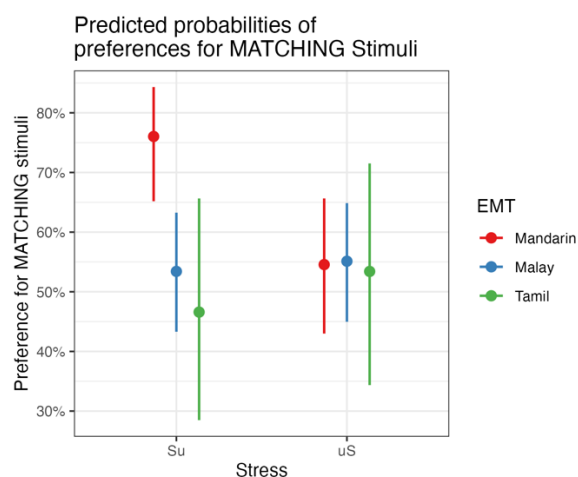


Figure 2: Predicted effects of stress condition and ethnic mother tongue (EMT) on choice in the paired comparison task (Task 2).

References

- [1] J. R. E. Leimgruber, *Singapore English: Structure, Variation, and Usage*. Cambridge: Cambridge University Press, 2013. doi: 10.1017/CBO9781139225755.
- [2] A. J. Chong and J. S. German, “Prominence and intonation in Singapore English,” *Journal of Phonetics*, vol. 98, p. 101240, May 2023, doi: 10.1016/j.wocn.2023.101240.
- [3] J. S. German and A. J. Chong, “Stress, tonal alignment, and phrasal position in Singapore English,” in *TAL2018, Sixth International Symposium on Tonal Aspects of Languages*, ISCA, Jun. 2018, pp. 150–154. doi: 10.21437/TAL.2018-30.
- [4] K. Connell *et al.*, “English Learners’ Use of Segmental and Suprasegmental Cues to Stress in Lexical Access: An Eye-Tracking Study,” *Language Learning*, vol. 68, no. 3, pp. 635–668, 2018, doi: 10.1111/lang.12288.
- [5] K. J. P. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, “Headphone screening to facilitate web-based auditory experiments,” *Atten Percept Psychophys*, vol. 79, no. 7, pp. 2064–2072, Oct. 2017, doi: 10.3758/s13414-017-1361-2.

Keywords

Lexical stress; pitch; variation; language contact; bilingualism; ethnic variation

The iconic and metaphorical representations of lexical tones: Evidence from Hong Kong Cantonese

Feier Gao (feiergao@seu.edu.cn), Southeast University, Nanjing, China

Chun Hau Ngai (cngai059@uottawa.ca), University of Ottawa, Ottawa, Canada

He Zhou (he.zhou@polyu.edu.hk), Hong Kong Polytechnic University, Hong Kong, SAR

Motion (upward vs. downward) and emotional valence (positive vs. negative) are mentally represented in a congruent manner (Casasanto & De Bruin, 2019; Casasanto & Dijkstra, 2010). In Chinese languages, contour tones also have upward or downward pitch trajectories. While previous studies have revealed an audiospatial binding in the production and perception of lexical tones (Garg et al., 2019), its symbolic potentials remain to be explored. **The present study** investigates the crossmodal correspondence demonstrated by the high-rising and the low-falling tones in Hong Kong Cantonese, both in an iconic domain (e.g., spatial motion) and a metaphorical domain (e.g., emotional valence). Unlike Mandarin Chinese, there is no spatially-relevant tonal metaphors or marks that are used by Hong Kong Cantonese speakers, which better helps us tease apart the influence of language experience on shaping the mental representation of tones.

Stimuli. For each type of correspondence, seven word pairs were used as visual stimuli (Table 1), thus giving us 14 different pairs of visual word. Each visual pair was displayed twice ($14 \times 2 = 28$). Within each pair, one word is labelled as “upward” and the other as “downward” with matched intensity, confirmed in a norming task. For auditory stimuli, each of the seven Cantonese vowels /i, y, e, oe, a, o, u/ was produced with two Cantonese tones—Tone 2 (high-rising, [35]) and Tone 4 (low-falling, [21]). All stimuli were recorded by a female and a male native speaker, thus giving us 28 tokens in total. Each audio was paired with a visual-word pair (samples given in Table 1), counterbalanced across subjects.

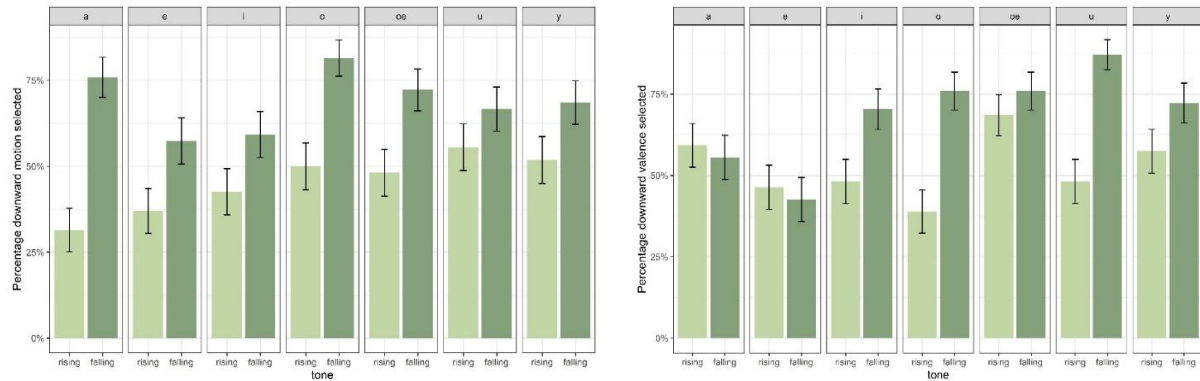
Procedure. Sixty native speakers of Hong Kong Cantonese participated in the study online via Prolific. Subjects were told they were about to participate in a vocabulary game of an “alien language”, by clicking the button to listen to the audio only once and selected from the two options the one they think that was congruent with the word they just heard.

Results. A series of logistic regression models were constructed to examine the effects of *tone* and *vowel* as well as their interaction on the selection. A maximum model was first constructed and sequentially pruned based on the results of model comparison. For **the tone-motion correspondence**, the removal of the interaction term didn’t hurt model fit, and the pairwise comparisons showed that the falling tone was more likely to be matched with downward motions (68.8%) than the rising tone ($p < 0.001$), and such pattern was not significantly modulated by vowel category. For **the tone-valence correspondence**, a significant *tone* \times *vowel* interaction was found. Specifically, at each contour level, certain vowels (i.e., mid vowels /e/ and /o/) diverged in terms of the valence association compared with others; also, relative to the rising tone, the falling tone is only significantly more likely to be matched with negative emotions when combined with /i/, /o/ and /u/ (Table 2).

Discussion. The current study is in line with prior findings on tone-motion mapping in Mandarin (Gao, 2024), showing a robust iconic motion correspondence in a tonal language which does not employ spatial metaphor. This suggests that the crossmodal correspondence between lexical tones and motion is likely to be audiospatial by nature. In addition, we provide novel evidence that the symbolic potential of tones can be even extended to abstract domain—which indicates that lexical tones are also metaphorically represented in our mind, although such correspondence is affected by vowel category.

Table 1. Sample word stimuli

Type	Directionality	Word	Translation
motion	upward	爬山	“hiking”
	downward	潛水	“diving”
valence	upward	開心	“cheerful”
	downward	害怕	“terrified”

FIGURE. Percentages of “downward” options selected; left: motion; right: valence)**Table 2.** Pairwise comparisons for tone-valence correspondence (*p*-values corrected)

		β	SE	z	<i>p</i> -values
Level = tone					
rising	/o/ - /oe/	-1.31e+00	0.420	-3.119	0.038 *
falling	/e/ - /o/	-1.447	0.421	-3.439	0.012 *
	/e/ - /oe/	-1.447	0.421	-3.439	0.012 *
	/e/ - /u/	-2.2027	0.490	-4.498	< 0.001 ***
	/e/ - /y/	-1.254	0.410	-3.059	0.047 *
Level = vowel					
/a/	rising - falling	-0.152	0.390	-0.389	0.697
/e/		-0.170	0.413	-0.412	0.681
/i/		0.973	0.427	2.277	0.023 *
/o/		2.200	0.677	3.248	0.001 **
/oe/		0.427	0.468	0.911	0.362
/u/		15.012	2.934	5.116	< 0.001 ***
/y/		0.678	0.424	1.600	0.110

*Note: some non-significant results are not listed, due to limited space.

References

- Casasanto, D., & De Bruin, A. (2019). Metaphors we learn by: Directed motor action improves word learning. *Cognition*, 182, 177–183. <https://doi.org/10.1016/j.cognition.2018.09.015>
- Casasanto, D., & Dijkstra, K. (2010). Motor action and emotional memory. *Cognition*, 115(1), 179–185. <https://doi.org/10.1016/j.cognition.2009.11.002>
- Gao, F. (2024). Crossmodal correspondence between lexical tones and visual motions: A forced-choice mapping task on Mandarin Chinese. *Linguistics Vanguard*, 10(1), 721–729. <https://doi.org/10.1515/lingvan-2023-0151>
- Garg, S., Hamarneh, G., Jongman, A., Sereno, J. A., & Wang, Y. (2019). Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. *Speech Communication*, 113, 47–62. <https://doi.org/10.1016/j.specom.2019.08.003>

The Effect of Orthography on the Representation of Mandarin Tone in L1 English Listeners | Jamie Adams | University of York

The influence of orthography on L2 phonology has received increased attention in recent years (Bassetti, 2023; Hayes-Harb & Barrios, 2021), although most of the research in this area focusses on orthographic effects on segmental acquisition. This PhD project explores the influence of diacritics on tone discrimination in L1 English learners of Mandarin. It compares three orthographic systems: pinyin, a Latin-based script with diacritics above the main vowel (fiū, fiù, fiu); zhuyin, a non-Latin script with right-aligned diacritics (ㄈㄩˊ, ㄈㄩˋ, ㄈㄩ); and modpin, a novel system combining pinyin spelling with right-aligned diacritics (fiuˊ, fiuˋ, fiu).

90 L1 English listeners with no Mandarin experience were randomly assigned to an orthographic group and took part in a three-phase experiment. In the exposure phase, they were exposed to 24 auditory Mandarin non-words accompanied by an orthographic transcription. In the threshold phase, they completed a cross-modal AX discrimination task, indicating whether the auditory and spoken word matched. 50% of trials were matches, 50% were segmental mismatches. Participants repeated the exposure and criterion phases until they reached 90% accuracy. The test phase was the same as the criterion phase, but mismatches were now tonal.

Responses are analysed using Signal Detection Theory and a generalised linear mixed effects model (GLMM). Results of the threshold phase indicate that zhuyin learners require significantly more exposure trials to reach 90% than pinyin ($\beta = -3.943$, $SE = 0.123$, $df = 3357$, $t\text{-ratio} = -31.979$, $p < 0.0001$) or modpin learners ($\beta = -4.119$, $SE = 0.136$, $df = 3357$, $t\text{-ratio} = -30.246$, $p < 0.0001$). This corroborates findings by Hayes-Harb and Cheng (2016) and suggests that any facilitative effect of zhuyin is modulated by the difficulty in acquiring it. Figure 1 shows that the number of exposure cycles within each group is highly variable, suggesting that factors other than orthography may be at play here. No difference is found between pinyin and modpin learners, indicating that the position of the diacritics has no effect on the ease with which learners acquire the system.

Each participant's final criterion trial was analysed to assess the impact of orthography on segmental discrimination. Figure 2 shows mean d-prime scores for each group. Despite considerable variability, ANOVA confirms that pinyin ($\beta = 0.314$, $SE = 0.053$, $df = 1053$, $t\text{-ratio} = 5.897$, $p < 0.0001$) and modpin ($\beta = 0.241$, $SE = 0.056$, $df = 1053$, $p < 0.0001$) learners scored significantly higher than those in the zhuyin group. This runs contrary to Hayes-Harb and Cheng's (2016) finding that zhuyin proved facilitative in the discrimination of segments. Given that modpin and pinyin are orthographically identical with regard to their segments, it is not surprising that no difference is found between these two ($\beta = -0.073$, $SE = 0.049$, $df = 1053$, $t\text{-ratio} = -1.476$, $p = 0.303$).

Binary accuracy scores (correct / incorrect) were predicted in a generalised linear mixed effects model with group, tone and their interaction as fixed effects, random intercepts for participant, auditory item and written item and a random slope allowing participants to vary by tone. Model estimates are visualised in Figure 3. No effect of orthographic group was found, suggesting that neither tone diacritic placement, nor orthographic familiarity have an effect on tone perception. However, the neutral tone has a significant effect on accuracy ($\beta = -0.94304$, $SE = 0.24799$, $z = -3.803$, $p < 0.001$), resulting in lower scores across the board. This is likely due to variability in the realisation of NT, making it more difficult for learners to create a stable category for it. The model suggests that the vast majority of variation comes from the random effects, indicating a need for further research on individual cognitive differences which may help to explain it.

The Effect of Orthography on the Representation of Mandarin Tone in L1 English Listeners | Jamie Adams | University of York

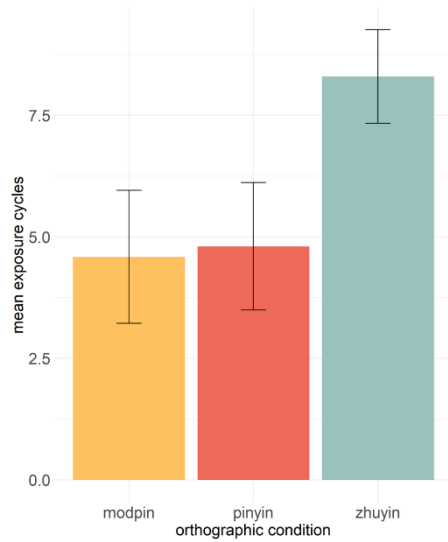


Figure 1. Barplot of exposure cycles by orthographic group.

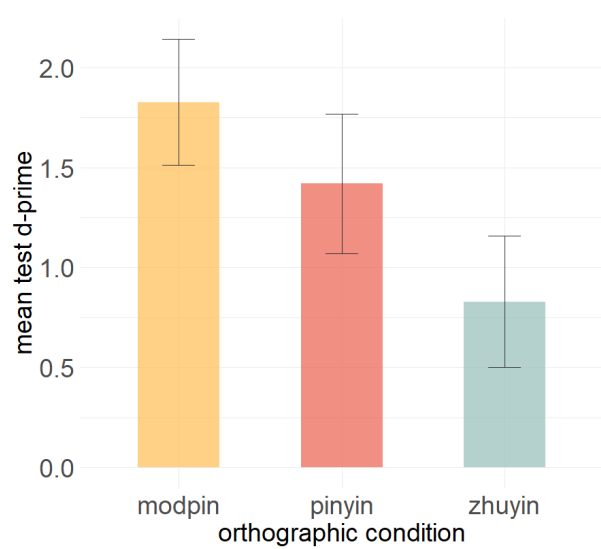


Figure 2. Mean segment discrimination d-primes by orthographic group.

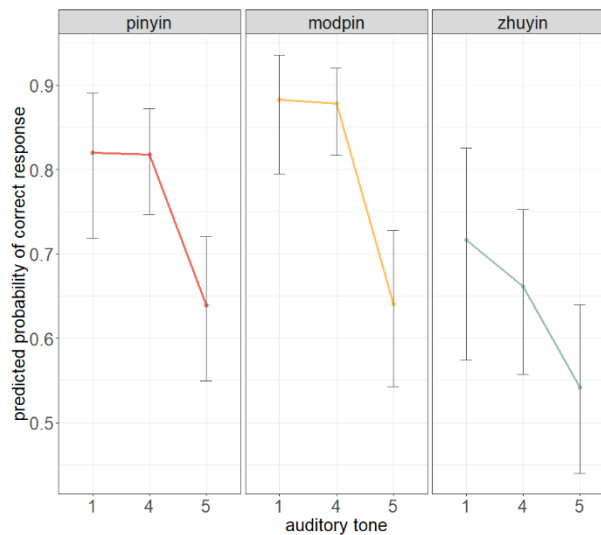


Figure 3. Model predictions of tone perception accuracy by tone and orthographic group.

References

- Bassetti, B. (2023). *Effects of orthography on second language phonology: Learning, awareness, perception and production*. Routledge.
- Hayes-Harb, R., & Barrios, S. (2021). The influence of orthography in second language phonological acquisition. *Language Teaching*, 54(3), 297–326.
- Hayes-Harb, R., & Cheng, H.-W. (2016). The Influence of the Pinyin and Zhuyin Writing Systems on the Acquisition of Mandarin Word Forms by Native English Speakers. *Frontiers in Psychology*, 7, 785.

Prosodic focus in Hebrew: An Acoustic study

Nitzan Bechar Marom¹ Vered Silber Varod,² Noam Amir¹

¹*Department of Communication Disorders, Faculty of Medicine, Tel Aviv University*

²*Center of AI and Data Science (TAD center), Tel Aviv University*

One of the common uses of prosody is to assign prominence to a particular element of an utterance. This prominence draws the listener's attention to the information most important to the speaker. The prosodically highlighted element usually marks the *new* and updated information of the utterance, as opposed to the *given* information [1]. It is common to distinguish between three types of focus: Broad focus (BF); Narrow focus (NF); and Contrastive focus (CF), which is a subcategory of narrow focus [2].

Very few studies have examined production of focus in Hebrew [3], [4], [5], [6]. In the present study, we examined how acoustic measures of prosody (F0, intensity, and duration) align with perception of focus as determined in a previous study [5] on the same data. We also examine how these measures depend on word position in the sentence (first, middle or last position in three-word sentences). Finally, we searched for evidence of Post focal compression, as found in some languages [6], [7].

For this purpose, five target sentences of three disyllabic words were recorded by eighteen speakers (9 women), who were asked to read them aloud in response to a question presented to them on the screen. Each speaker recorded 30 variations, resulting in 450 recordings in total. Ten listeners (women only) who participated in a preliminary experiment [5] performed a listening test via Zoom (Due to COVID-19 restrictions). The listeners were asked to a) mark the word in focus (or refute the existence of a narrow/contrastive focus); and b) to rate the degree of prominence on a scale of 1 (very weak) to 5 (very strong).

Out of the original 450 recordings, 221 were judged consistently, so only these were analyzed acoustically. Praat software was used to annotate the stressed vowel in each of three bisyllabic words in each utterance, and the comparisons below were conducted only on these three segments. This was performed to avoid any biases due to differences between stressed and unstressed vowels, intensity differences between vowels and consonants, etc. Praat was also used to compute F0, after which all pitch contours were examined visually and corrected manually if necessary. F0 and intensity were normalized relatively to the unfocused part of each utterance.

The main results of the acoustic analysis are presented in Figure 1. Figure 1a shows mean measurements in intensity for five conditions: 1- NF on word 1, 2- NF on word 2, 3- NF on word 3, 4 – BF, 5- CF on word 2. The blue columns are the measurements on the first word, red are on the second, and yellow on the third. Condition 4 (BF) can be regarded as a baseline, exhibiting the well known natural tendency for intensity to decrease throughout an utterance. Condition 1 demonstrates PFC: the first word starts with approximately the same intensity as in BF, but words 2 and 3 have lowered intensity relative to BF. Indeed, the intensity of the first word is approximately the same in all conditions, with changes mainly in words 2 and 3. When NF is on word 2, its intensity rises *above* that of word 1, and again PFC causes the third word to have very low intensity. When NF is present on the last word of the utterance, the “opposing forces” of NF vs. Final weakening cause all the words in the utterances to have approximately the same intensity. Finally, CF, condition 5, is an exaggeration of NF: more heightening of intensity on word 2, and more weakening of word 3. These observations were borne out by statistical analyses (ANOVA), which were omitted here due to lack of space.

Results for F0, in Figure 1b, show results very similar to those of intensity. The main difference is that patterns for NF and CF (conditions 2 and 5) are nearly identical, in contrast to intensity, where CF caused a more extreme pattern.

Finally, results for vowel duration, in figure 1c, show that duration has nearly no role in producing narrow focus. Interestingly, it was affected only by focus on word 3, the last one, producing a noticeable degree of final lengthening.

In conclusion, intensity and F0 can be regarded as the main acoustic markers of NF/CF in Hebrew, whereas duration appears to play a very minor role in this respect. This is not surprising given that previous studies have shown that duration is the main acoustic indicator of word level stress in Hebrew [8]. In addition, the results provide a clear indication that PFC is present in Hebrew.

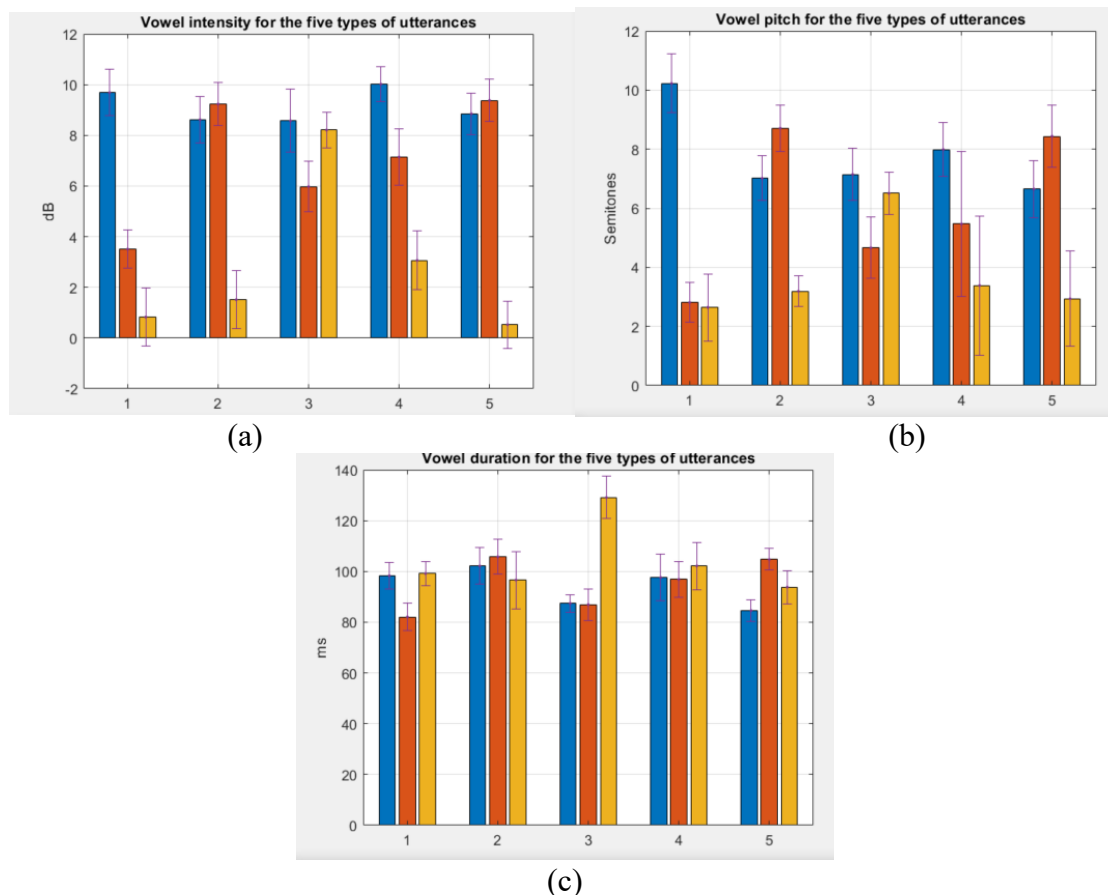


Figure 1. The effect of narrow/contrastive focus on (a) intensity, (b) pitch and (c) duration. Conditions 1-5 in each plot are: 1- NF on word 1; 2-NF on word 2; 3-NF on word 3; 4-BF; 5-CF on word 2. Blue bars represent the first word, red bars represent the second word, yellow are the third. Error bars indicated 95% confidence intervals.

- [1] Birch, S., & Clifton, C. 2002. Effects of varying focus and accenting of adjuncts on the comprehension of utterances. *Journal of Memory and Language*, 47(4), 571-588.
- [2] Sityaev, D., & House, J. 2003. Phonetic and phonological correlates of broad, narrow and contrastive focus in English. In *15th ICPHS* (Vol. 1822), 1819-1822.
- [3] Amir, N., Almogi, B. C., & Gal, R. 2004. Perceiving prominence and emotion in speech - A cross lingual study. In *Speech Prosody 2004 (SP-2004)*, 375-378.
- [4] Mixdorff, H., & Amir, N. 2002. The Prosody of Modern Hebrew - A Quantitative Study. In *Speech Prosody 2002 (SP-2002)*, 511-514.
- [5] Bechar Marom, N., Hechter, N., Tahayu, Y., Amir, N., & Silber-Varod, V. (2021). Prosodic focus in Hebrew: A Perception study. *Tone and Intonation 2021 (TAI 2021)*, The 1st International conference on TAI, Sønderborg, Denmark.
- [6] Amir, N., Silber-Varod, V., Anavi, O., Deouell, E., & Mixdorff, H. (2023). The effect of f0 and post-focal compression on the perception of narrow focus in hebrew. *ICPhS*
- [7] Xu, Y. (2011). Post-focus compression: Cross-linguistic distribution and historical origin. In *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong (pp. 152-155).
- [8] Silber-Varod, V., Sagi, H., & Amir, N. (2016). The acoustic correlates of lexical stress in Israeli Hebrew. *Journal of Phonetics*, 56, 1-14. doi: 10.1016/j.wocn.2016.01.003

Keywords: Narrow focus, broad focus, prominence, post focal compression, prosody

Lexical tone and stress in Kera'a, an endangered Trans-Himalayan language

Kirsten Culhane (University of Canterbury),
Naomi Peck (University of Freiburg)
Uta Reinöhl (University of Freiburg)

Key words: Trans-Himalayan, lexical tone, stress, phonetic cues, variation

Kera'a (glottocode: idum1241) is an understudied Trans-Himalayan (*aka* Tibeto-Burman) language of the Eastern Himalayas. It is spoken by about 10 000 people in northeastern Arunachal Pradesh (India; see Figure 1) as well as by a smaller community in Tibet, China (Reinöhl 2022). While rich evidence suggests that Kera'a is tonal, it is an open question whether it has stress. In this talk, we investigate the question of stress and relate it to our research of Kera'a tone, given that both phenomena are linked to partially identical phonetic correlates.

Kera'a has not been described as having stress. However, in order to better understand the prosody of this understudied language, we consider the possibility that it has non-lexical stress (i.e. that stress regularly falls on a particular syllable). We conduct an acoustic analysis of Kera'a spontaneous speech, focussing on duration, intensity, spectral tilt and measures of vowel quality. We investigate if any particular syllable in the word has greater duration, greater intensity or lower spectral tilt values. We also examine if there is evidence of vowel reduction in any particular syllable, as in many languages, vowels undergo reduction in unstressed syllables (Gordon and Roettger 2017). We do not find, however, any consistent patterns which would support a stress analysis. Phonotactic patterns of the language also do not support a stress analysis. These findings lead us to the tentative conclusion that stress is probably not a relevant feature of Kera'a phonology, and that we can turn our full attention to the tonal system of Kera'a.

We present a preliminary phonological analysis of a five-way level-tone system in Kera'a based on evidence from mono- and multi-syllabic lexemes in both isolation contexts and connected speech (e.g. /da¹/ 'to tie up', /da²/ 'to return', /da⁴/ 'to loan, borrow'; /shi²/ 'to trample', /shi⁵/ 'to die'; /a²tā¹/ 'elephant, /a⁴tā⁵/ 'food'). We take into account the significant sociolinguistic variation of how tone is realised, in particular relative to dialects and the lower social level of 'clan-lects' (see also Kaland et al. 2021). The lexical tone system of Kera'a shows rare complexity both areally and typologically. A five-way level-tone system is not only, to the best of our knowledge, unattested in other Trans-Himalayan languages of the eastern Himalayas, but rare in the world's languages (Yip 2002: 26-7). We conclude by suggesting a link between Kera'a tonal complexity and the ongoing, wide-ranging loss of its codas and onsets (Reinöhl 2022, Culhane et al. 2023).



Figure 1: Map of districts in Arunachal Pradesh, India, in which Kera'a is widely spoken (dark grey)

References:

- Culhane, Kirsten, Naomi Peck, and Uta Reinöhl. 2023. The Mystery of Word-initial Consonant Loss in Kera'a. *SocArXiv*. doi:10.31235/osf.io/zvfdw.
- Gordon, Matthew, and Timo Roettger. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard* 3:1–11.
- Himmelman, Nikolaus P. 2022. On the comparability of prosodic categories: why 'stress' is difficult. *Linguistic Typology* 27(2): 341-361.
- Kaland, Constantijn, Naomi Peck, T. Mark Ellison & Uta Reinöhl. 2021. An initial exploration of the interaction of tone and intonation in Kera'a. *Proc. 1st International Conference on Tone and Intonation (TAI)*, 132-136.
- Morey, Stephen. 2014. Studying Tones in North East India: Tai, Singpho and Tangsa. *Language Documentation and Conservation* 8:637-671.
- Post, Mark W. & Robbins Burling. 2017. The Tibeto-Burman Languages of Northeastern India. In: Graham Thurgood and Randy J. LaPolla (eds.), *The Sino-Tibetan Languages*, 213-242. 2nd edn. London & New York: Routledge.
- Reinöhl, Uta. 2022. Locating Kera'a (Idu Mishmi) in Its Linguistic Neighbourhood: Evidence from Dialectology. In: Mark W Post, Stephen Morey & Toni Huber (eds.), *Ethnolinguistic Prehistory of the Eastern Himalaya*, 232–263. Leiden: Brill.
- Yip, Moira. 2002. *Tone*. Cambridge: Cambridge UP.

Is F0 a correlate of sentence-level prominence in Ukrainian?

Janina Mołczanow¹, Beata Łukaszewicz¹, Anna Łukaszewicz¹

¹*University of Warsaw*

As is well known, F0 is a widely attested acoustic correlate of both word-level and sentence-level prominence. Recent research on Ukrainian ([1], [2], [3]), based on citation forms embedded in a frame, revealed no significant differences in F0 maximum between lexically stressed and unstressed syllables, pointing to increased vocalic duration as the main cue to word-level prominence in accented positions. In addition, mean F0 slopes have been shown to be less steep in lexically stressed syllables than in unstressed syllables, which may be related to the enhanced duration of the former. The failure to detect expected differences in F0 indices between stressed and unstressed syllables may call into question the presence of intonational pitch accents which are associated with a stress-prominent syllable. The present study reappraises the question of whether F0 changes play a role in cueing sentence-level prominence in Ukrainian by (i) zooming in on the F0 shifts across the syllables immediately preceding lexical stress and those in lexically stressed syllables, and (ii) taking into account the existence of polar opposite (rising vs falling) intonational patterns in Ukrainian. The study is based on segmentally identical minimal pairs, which differ only in the position of lexical stress, i.e. $\sigma_1\sigma_2\sigma_3$ vs $\sigma_1\sigma_2'\sigma_3$, e.g. [dɔ'b'ihati] 'run' (perf.) – [dɔb'i'hiati] 'run' (imperf.), occurring in focus position.

14 native speakers of standard Ukrainian (10F, 4M; aged 24-62, M = 39) participated in the experiment. The recordings were done using an H4 Zoom portable recorder set at a sampling rate of 44.1 kHz and an AT897 microphone. In sum, 588 recorded word tokens (14 speakers \times 14 words \times 3 repetitions) have been obtained. Measurements were conducted in 1176 vowels.

Analyses of F0 max and F0 slope corroborate previous results (Fig. 1). A potential shift in F0 occurring across adjacent syllables is further investigated using the version of Pairwise Variability Index (PVI; [4], [5]; cf. [6]) which expresses the magnitude and direction of local F0 change: $PVI_{F0} = 12 \cdot \log_2(F0_k/F0_{k+1}) = ST_k - ST_{k+1}$ (ST stands for 'semitones').

We hypothesise that if focus is expressed by F0 excursions aligned with the position of lexical stress, then different PVI values are expected for the comparison $\sigma_1'\sigma_2\sigma_3\sigma_4$ vs $\sigma_1\sigma_2'\sigma_3\sigma_4$, but similar PVI values for the comparison $\sigma_1'\sigma_2\sigma_3\sigma_4$ vs $\sigma_1\sigma_2'\sigma_3\sigma_4$.

Previous descriptions of the Ukrainian intonation [7] as well as a preliminary inspection of the data point to the presence of either a falling or a rising pattern of F0 change in the vicinity of the lexically stressed syllable. We thus divided the data using the k-means classification in SPSS (v. 29), run to fit a two-group model.

The statistical analysis (linear mixed effects models including speaker- and item-specific random slopes) revealed a significantly larger difference in the magnitude of F0 change between the first two syllables in $\sigma_1'\sigma_2\sigma_3\sigma_4$ than in $\sigma_1\sigma_2'\sigma_3\sigma_4$, both in the falling and in the rising pattern of F0 change ($p < .001$) (Fig. 2). This result indicates that F0 plays a role in signalling sentence-level prominence in Ukrainian. As for the comparison $\sigma_1'\sigma_2\sigma_3\sigma_4$ vs $\sigma_1\sigma_2'\sigma_3\sigma_4$, no statistically significant difference was observed between words with stress located on the second and the third syllables in the rising intonation group ($p = .09$) (Fig. 3). In turn, the falling intonation group exhibited a small difference of 0.8 ST which turned out to be statistically significant ($p < .05$) (Fig. 3). However, as the difference of 0.8 ST can be considered negligible from the point of view of perception, we conclude that the present findings lend support to the hypothesis that intonational pitch accent is associated with lexically stressed syllables in Ukrainian.

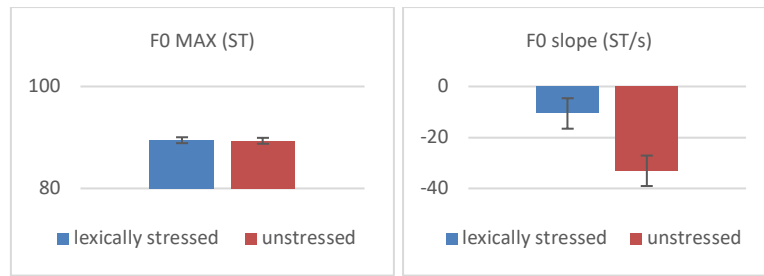


Figure 1. Mean F0 maximum (left panel) and F0 slope (right panel).

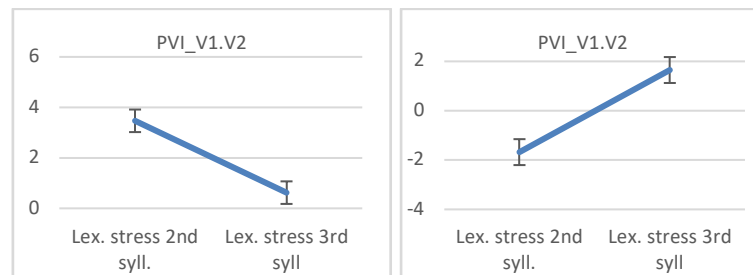


Figure 2. Mean PVIs of the first two syllables (in $\sigma 1' \sigma 2 \sigma 3 \sigma 4$ vs $\sigma 1 \sigma 2' \sigma 3 \sigma 4$ words) in the group with a falling F0 pattern (left panel) and a rising F0 pattern (right panel).

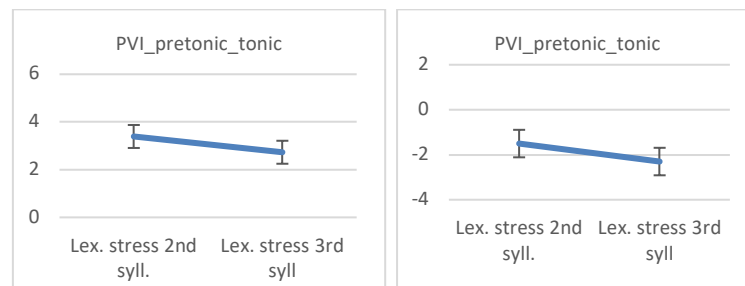


Figure 3. Mean PVIs of the pretonic and tonic syllables (in $\sigma 1' \sigma 2 \sigma 3 \sigma 4$ vs $\sigma 1 \sigma 2' \sigma 3 \sigma 4$ words) in the group with a falling F0 pattern (left panel) and a rising F0 pattern (right panel).

- [1] Łukaszewicz, B., & Mołczanow, J. (2025). *Issues in Metrical Phonology: Insights from Ukrainian*. Cambridge: Cambridge University Press.
- [2] Łukaszewicz, B., & Mołczanow, J. 2018. Rhythmic stress in Ukrainian: acoustic evidence of a bidirectional system. *Journal of Linguistics* 54, 367-388.
- [3] Łukaszewicz, B., & Mołczanow, J. 2018. The role of vowel parameters in defining lexical and subsidiary stress in Ukrainian. *Poznan Studies in Contemporary Linguistics* 54, 355-375.
- [4] Arciuli, J., Simpson, B. S., Vogel, A. P., & Ballard, K. J. 2014. Acoustic changes in the production of lexical stress during Lombard speech. *Language and Speech* 57, 149-162.
- [5] Łukaszewicz, B. 2018. Phonetic evidence for an iterative stress system: the issue of consonantal rhythm. *Phonology* 35, 115-150.
- [6] Low, E. L., Grabe, E., & Nolan, F. 2000. Quantitative characterizations of speech rhythm: syllable-timing in Singapore English. *Language and Speech* 43, 377-401.
- [7] Nikolaeva, T. M. 1977. *Frazovaja Intonacija Slavjanskix Jazykov*. Moscow: Izdatel'stvo Nauka.

The prevalence of L+H* and H* in Basque Spanish conversational statements

Gorka Elordieta¹, Magdalena Romera², Asier Illaro¹

¹University of the Basque Country (UPV/EHU), ²Public University of Navarre

The present study aims to contribute to the growing work on prosodic aspects of languages or language varieties in contact, where prosodic features of one language or language variety are present in the other. It focuses on the variety of Spanish in contact with Basque (Basque Spanish), in parts of the provinces of Bizkaia, Gipuzkoa and Navarre, in northern Spain. [1]-[3] analyze the nuclear contours in conversational speech of information-seeking absolute interrogatives in Basque Spanish, showing that the prevailing contour is rising-falling: L+(j)H* (H)L% (a rising nuclear pitch accent that can be upstepped, followed by a L% or HL% boundary tone). This type of contour is found in Basque, and contrasts with the rising final contour L* H% reported for Castilian Spanish in read speech (cf. references in [1]-[3]). [4] also finds a majority of rising final contours in Madrid Spanish conversational speech. [1]-[3] attribute the falling contour of Basque Spanish polar interrogatives to influence from Basque.

As for declaratives, [5] showed that in declarative utterances of speakers from the province of Gipuzkoa the majority of rising pitch accents had their peaks strictly aligned within the tonic syllable, L+H*. This contrasts with Castilian Spanish, where rising pitch accents have delayed peaks, L+<H* [6-9]. In Basque, the most common prenuclear and nuclear pitch accent in conversational speech is L+H* (cf. [10], [11]). Thus, a plausible hypothesis is that Basque L+H* is being transferred to Basque Spanish.

In the present study, we analyze the pitch accents in declarative utterances of conversational speech (interviews with a member of the research team) of 12 speakers from the province of Bizkaia, 6 from a city (Bilbao) and 6 from a small town (Lekeitio). They add to the 12 speakers from Gipuzkoa analyzed in [5], also from a city (San Sebastian) and a smaller town (Ibarra), for a total of 24 speakers (12 from two cities, and 12 from two small towns). The interest of considering urban and non-urban areas is that previous work on information-seeking polar questions in Basque Spanish has shown that there is a higher degree of presence of the Basque type of contour in areas where Basque is used in daily life consistently, compared to urban areas, where the use of Basque is very reduced (cf. [2]). 886 declarative utterances with 4,621 pitch accents have been analyzed, and compared to 1,051 pitch accents from 210 declarative utterances of 7 speakers of Central Castilian Spanish (from the city of Madrid) analyzed in [5]. The experimental methodology is the same for Basque Spanish and Madrid Spanish, so the results are directly comparable. We divide words in two types, depending on their information status: those that convey background or known information, and those that convey relevant information, responding to what was asked by the interlocutor. We call the accents in the two types *prenuclear* and *nuclear*, respectively (although the parallelism with the terms used in the literature on read speech is not exact). Three ToBI transcribers annotated the pitch accents and boundary tones (85% agreement).

Four separate logistic regression models were computed using the glmer function from the lme4 package [12] in RStudio [13] for the four most frequent tones: L+H*, L+<H*, H*, and L*. The results plotted in Figure 1 show that L+H* has a substantially higher probability of occurrence in non-urban and urban varieties of Basque Spanish compared to Madrid Spanish, with a higher probability in non-urban varieties. Since L+H* is the dominant pitch accent in Basque, the results would be compatible with the hypothesis that the use of Basque-like intonational features in Spanish is higher in areas where Basque is more present. Conversely, the predicted probability of L+<H* is much higher in Madrid Spanish. Finally, the predicted probability of L* is higher in Madrid Spanish than in Basque Spanish. The differences are less clear for H*.

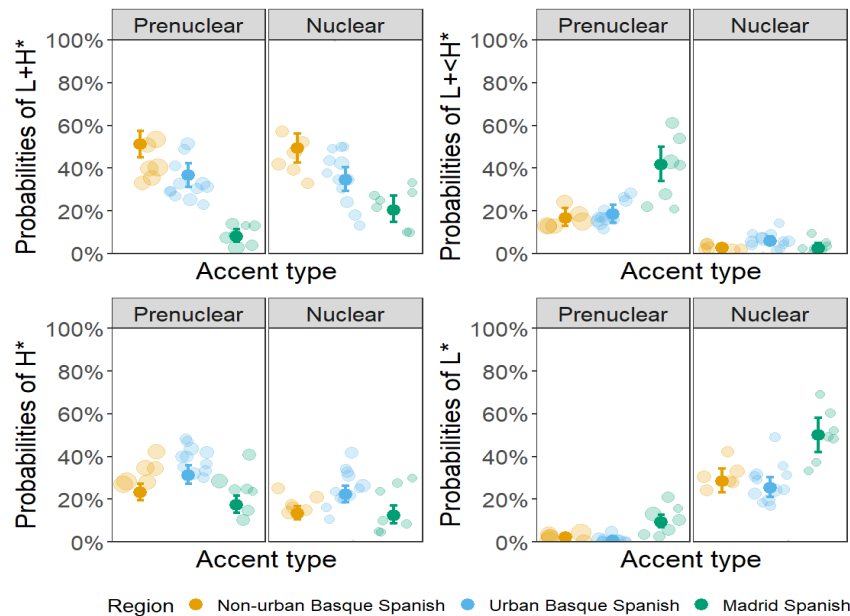


Figure 1. Probabilities of pitch accents by accent type and region. Full dots indicate model estimates with 95% CIs, and lighter dots indicate individual speaker means.

- [1] Elordieta, G. & Romera, M. 2021. The influence of social factors on the prosody of Spanish in contact with Basque. *International Journal of Bilingualism* 25, 286–317.
- [2] Romera, M. & Elordieta, G. 2020. Information-seeking question intonation in Basque Spanish and its correlation with degree of contact and language attitudes. *Languages* 5, 70.
- [3] Elordieta, G. & Romera, M. 2024. Contornos ascendentes-descendentes en el castellano del País Vasco y Navarra. In Elvira-García, W. & Roseano, P. (Eds.), *Avances metodológicos en fonética y prosodia*. Madrid: UNED, 85-96.
- [4] Elordieta, G. & Romera, M. 2020. The intonation of information-seeking absolute interrogatives in Madrid Spanish. *Estudios de Fonética Experimental* XXIX, 195-213.
- [5] Elordieta, G., Romera, M. & Illaro, A. 2024. Pitch accents in Basque Spanish declarative utterances. *Proceedings of Speech Prosody 2024* (Leiden, The Netherlands), 627-631.
- [6] Sosa, J.M. 1999. *La entonación del español*. Madrid: Cátedra.
- [7] Face, T. 2008. *The intonation of Castilian Spanish declaratives and absolute interrogatives*. Munich: Lincom Europa.
- [8] Estebas-Vilaplana, E. & Prieto, P. 2010. Castilian Spanish intonation. In Prieto, P. & Roseano, P. (Eds.), *Transcription of intonation of the Spanish language*. Munich: Lincom Europa, 17-48.
- [9] Hualde, J.I. & Prieto, P. 2015. Intonational variation in Spanish, European and American varieties. In Frota, S. & Prieto, P. (Eds.), *Intonation in Romance*. Oxford: Oxford University Press, 350-391.
- [10] Igartua, J. 2023. Ibarako euskararen azentuari eta intonazioari hurbilpen bat. Undergraduate thesis, University of the Basque Country.
- [11] Ezama, M. 2023. Azpeitiko euskara: Azentuaren eta intonazioaren azterketarako ekarpen txiki bat. Undergraduate thesis, University of the Basque Country.
- [12] Bates, D., Maechler, M., Bolker, B. & Walker, S. 2015. Fitting linear mixed-effects models using lme4 [R package]. *Journal of Statistical Software* 67, 1-48.
- [13] R Core Team. 2024. *R: A language and environment for statistical computing* [software]. Viena: R Foundation for Statistical Computing. Available: <http://www.r-project.org>.

Social and situational variability in the distribution of melodic contours

Tatiana Kachkovskaia

Helsinki Collegium for Advanced Studies, University of Helsinki

This paper addresses intonational variation in collaborative dialogues depending on the relationship between the interlocutors. This is part of an on-going research where the general hypothesis is that (a) the same speaker uses different prosodic features depending on who he/she is talking to, and that (b) within the same language, this variation can be systematic, i.e. similar tendencies can be observed for different speakers. This particular paper discusses variation in the frequency distribution of phonologically different melodic contours.

Testing such a hypothesis requires a carefully designed speech database. Here, the speech corpus SibLing was used [4], which is composed of collaborative dialogues where each of the 20 “core” Russian speakers talked to 5 other people in settings ranging in social distance: with (1) a same-gender sibling; (2) a same-gender friend of similar age; (3, 4) strangers of similar age and the same gender (labelled “stranger1”) vs. the opposite gender (“stranger2”); (5) a stranger of the same gender and greater age whose job requires leadership skills (labelled “boss”). The interlocutors were performing a collaborative task (card-matching game). In total, the analysis included 90 dialogues. The corpus contains manual prosodic annotation.

Russian intonation system consists of 5–7 phonologically distinct melodic patterns observed within the nucleus of each intonational phrase. The most frequent melodic patterns correspond to C. Odé’s [5] pitch accents L*, HL*, H*L, L*H, and H*H (Bryzgunova’s ICs 1, 2, 3, 4 and 6). Based on the Sibling’s prosodic annotation, the frequency of these patterns was calculated for each dialogue. Mean number of melodic contours per speaker per setting was 243 (sd = 85).

For several melodic patterns, statistical analysis revealed significant influence of the factor of social distance. Figure 1 shows the distributions for HL*, often used for contrast or emphasis (other uses, e.g. wh-questions, are extremely rare in our data). The factor of social distance was significant (ANOVA: $F(4,72)=7.46$, $p<0.0001$, $\eta^2=0.213$), and there was a tendency for HL* frequency to decrease with the growth of social distance. We might suggest that smaller social distance is associated with more arousal or emotionality (as well as more laughter, as shown previously [3]). Figure 2 shows the distributions of L*H, a rising contour used in ordered listing and often described as characteristic of formal speech [6]. The factor of social distance was significant (Friedman test: $\chi^2(4)=12.9$, $p=0.012$), suggesting that distant relationships may require some features of formal speaking style.

Other differences between the 5 settings were more visible within the male subgroup, see Fig. 3. With the growth of social distance, we observe a decrease in the frequency of L* (ANOVA: $F(4,28)=3.486$, $p=0.02$, $\eta^2=0.231$), a non-emphatic fall used in final and non-final IPs; more frequent use of H*H (ANOVA: $F(4,36)=3.596$, $p=0.014$, $\eta^2=0.182$), which often occurs in lists and, in some descriptions, associated with “active mental processes” [6]; somewhat less frequent use of H*L (borderline significance) reserved for neutral non-finality, listing, and yes/no questions¹.

The results presented above confirmed that variation in the use of melodic patterns could be explained by social distance between the interlocutors and bears similarities to stylistic variation, e.g. in formal vs. informal speech [1, 2]. With the growth of social distance, we observe more intonational features typical for formal settings.

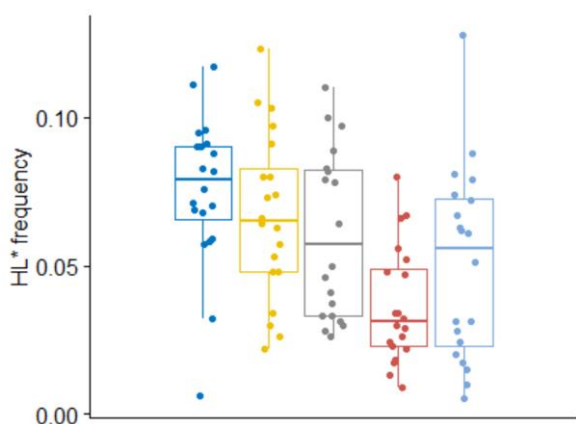


Fig. 1. Variation in the frequency of HL^* depending on the interlocutor; social distance increases from left to right

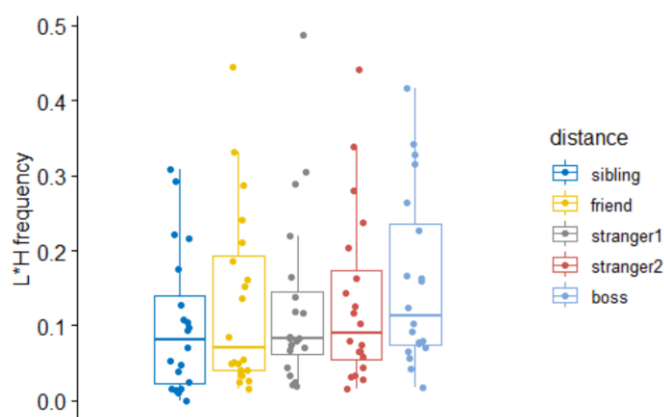


Fig. 2. Variation in the frequency of L^*H depending on the interlocutor; social distance increases from left to right

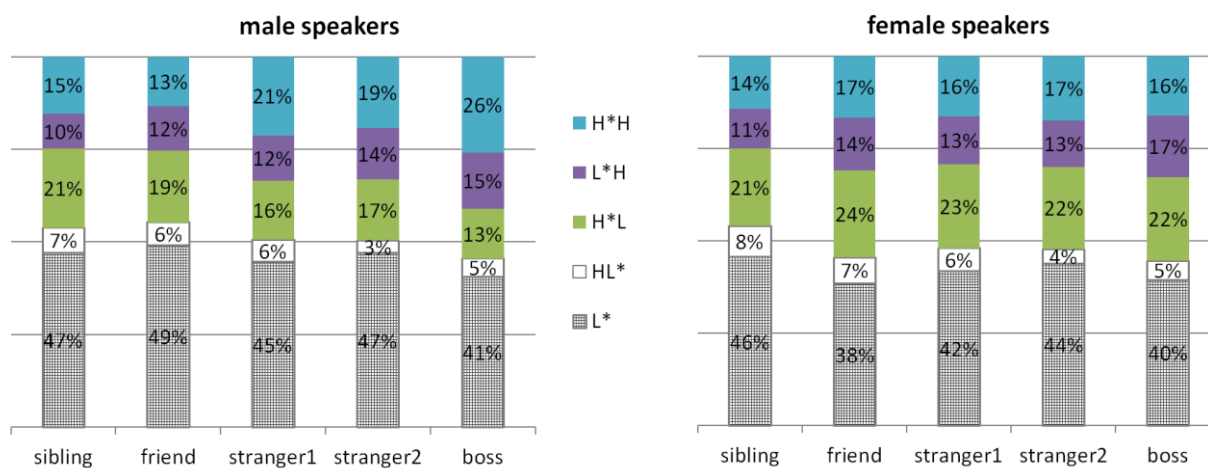


Fig. 3. Mean frequency of the five melodic contours in male's (left) and female's (right) speech depending on the interlocutor; social distance increases from left to right

References

- Henriksen, N. (2013). Style, prosodic variation, and the social meaning of intonation. *Journal of the International Phonetic Association*, 43, pp. 153–193
- Hirschberg, J. (2000). A Corpus-Based Approach to the Study of Speaking Style. In: Horne, M. (eds) *Prosody: Theory and Experiment. Text, Speech and Language Technology*, vol 14. Springer, Dordrecht, pp. 335–350
- Kachkovskaia T., Mamushina A., Kocharov D., Kholiavin P., Menshikova A., Guseva D., Portnova A., Evdokimova V., and Zimina S. (2022): *Kommunikativnaja adaptatsija v dialoge: vzgl'ad fonetista* (Communication Accommodation in Dialogues. A Phonetician's Perspective), Kachkovskaia T. (ed.) [in Russian], Saint Petersburg University Publishing House
- Kachkovskaia T., Chukaeva T., Evdokimova V., Kholiavin P., Kriakina N., Kocharov D., Mamushina A., Menshikova A., and Zimina S. (2020). Sibling corpus of Russian dialogue speech designed for research on speech entrainment. *Proc. of the 12th LREC. Marseille, France: ELRA*, 2020, pp. 6556–6561
- Odé, C. (2008). Transcription of Russian intonation, ToRI, an interactive research tool and learning module on the internet. *Dutch Contributions to the Fourteenth International Congress of Slavists. Ohrid, North Macedonia, September 10–16; vol. 1*, pp. 431–50
- Yanko T. (2008). *Intonational strategies of the Russian speech from a contrastive perspective [Intonatsionnye strategii russkoj rechi v sopostavitel'nom aspekte]*. Yazyki slavyanskikh kul'tur, Moscow

ⁱ Further testing revealed that frequency of yes/no questions per se decreases with the growth of social distance.

AREM: A REGISTER LANGUAGE WITHOUT TONE

TẠ Thành Tấn

Hanoi National University of Education, and Hankuk University of Foreign Studies

tantt@hnue.edu.vn, tathanhtan@hufs.ac.kr

Arem is a Vietic language of the Austroasiatic phylum spoken in a single hamlet located inside the Phong Nha-Kẻ Bàng National Park in Quảng Bình province, north-central Vietnam. This language is dying, as nowadays only seven Arem people can speak it [1].

Arem has a binary phonological contrast between high and low registers characteristic of many Austroasiatic languages. Register is assumed to be derived from the devoicing of onset obstruents in the proto-language, a process usually known as *registrogenesis* [2-6]. Typically, the high register (< *voiceless onsets) has a higher f₀ and more open vowels than the low register (< *voiced onsets), and the high register also has a modal phonation compared to a breathy phonation of the low register [5, 7]. The Arem register has been described as signaled by the expected differences in f₀, vowel quality, and phonation type between the high and low registers, like in other register languages in the Austroasiatic phylum [8-11]. However, there has not been an empirical study on the phonetic properties of the register contrast in Arem.

In June 2019, I recorded three male and two female Arem speakers to study the register contrast in the language. I designed a wordlist consisting of 40 lexical items made up of four onsets /t, k, r, l/ combined with five vowels /i:, u:, ε:, ɔ:, a:/ in the two registers. In September 2021, another production study was conducted on glottal codas and tones in Arem. I recorded two male and two female speakers pronouncing a wordlist of 81 lexical items consisting of five different types of codas, the glottal fricative -h, the glottal stop -ʔ, the lateral -l, the glottalized lateral -lʔ, and open syllables combined with various vowels in the two registers. The data was annotated in Praat using TextGrid, and f₀, spectral tilt, and H1-H2 measurements were subsequently extracted by PraatSauce [12].

The results show that Arem contrasts a high and a low register by differences in vowel quality and phonation. The proto-vowel system was doubled in the two registers, with high-register vowels more open than their low-register counterparts (Fig 1.). The low register has a breathier phonation compared to the high register; however, no f₀ difference between the two registers was found (Fig. 2), contra previous studies [8, 13]. The results also refute the treatment of Arem as tonal [8, 13]. This is because the glottal finals -h, -ʔ and glottalized sonorants are still preserved in the language and because the duration, f₀, and phonation perturbations they condition on the preceding rhyme are highly variable across registers and speakers (Fig. 3).

The results of this study, therefore, contribute to a better understanding of register contrast in Arem, an extremely endangered language. These findings also shed light on hypotheses about registrogenesis in Southeast Asian languages, questioning previous assumptions and calling for modifications.

Keywords: *Arem, register, tone, tonogenesis, registrogenesis*

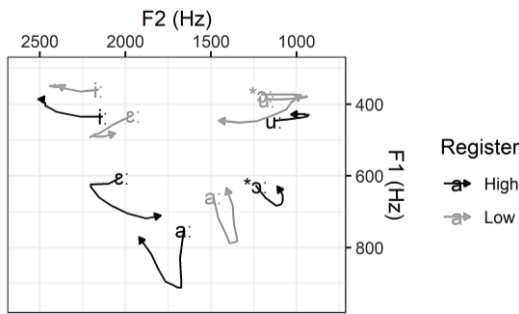


Fig. 1: Vowels in the high and low register in Arem from their onsets (IPA symbols) to midpoints (arrows).

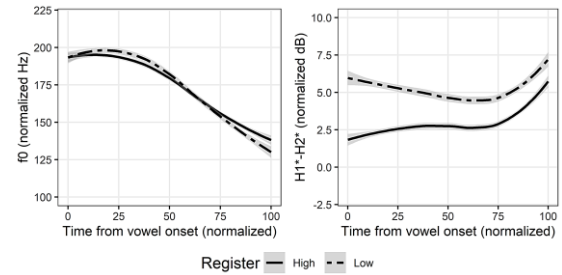


Fig. 2: Mean f_0 (left panel) and $H1^*-H2^*$ (right panel) of the high and low registers in Arem.

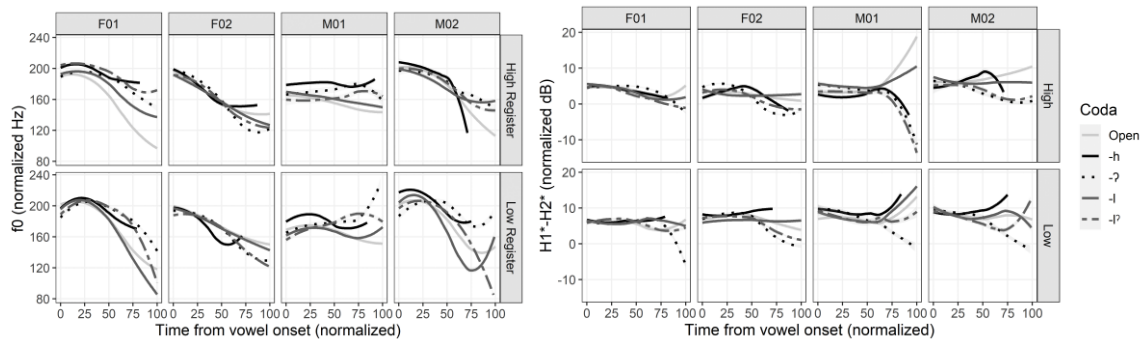


Fig. 3: f_0 (left) and $H1^*-H2^*$ (right) on the syllables having different coda types, subgrouped by *speaker* and *register*.

References

- [1] T. T. Ta, "Register and Tone Developments in Vietic languages," Ph.D Dissertation, Linguistics, University of Ottawa, Ottawa, 2023.
- [2] E. J. A. Henderson, "The Main Features of Cambodian Pronunciation," *Bulletin of the School of Oriental and African Studies*, vol. 14, no. 1, pp. 149-174, 1952.
- [3] F. E. Huffman, "The Register Problem in Fifteen Mon-Khmer Languages," *Oceanic Linguistics Special Publications*, no. 13, pp. 575-589, 1976.
- [4] M. Brunelle and T. T. Ta, "29 Register in languages of Mainland Southeast Asia: the state of the art " in *The Languages and Linguistics of Mainland Southeast Asia: A comprehensive guide*, S. Paul and J. Mathias Eds.: De Gruyter Mouton, pp. 683-706, 2021.
- [5] M. Ferlus, "Formation des registres et mutations consonantiques dans les langues Mon-Khmer," *Mon-Khmer Studies*, vol. VIII, pp. 1-76, 1979.
- [6] A.-G. Haudricourt, "Les mutations consonantiques des occlusives initiales en môn-khmer," *Bulletin de la Société de Linguistique de Paris*, vol. 60, 1, pp. 160-172, 1965.
- [7] J. Kirby and M. Brunelle, "Southeast Asian tone in areal perspective," *The Cambridge Handbook of Areal Linguistics*, pp. 703-731, 2017.
- [8] M. Ferlus, "Arem, a Vietic Language," *Mon-Khmer Studies*, vol. 43, 1, pp. 1-15, 2013.
- [9] V. L. Nguyễn, *The Rục language*. Hà Nội: KHXH, 1993.
- [10] V. P. Đoàn, "The phonetic system in Arem," in *Những vấn đề ngôn ngữ học về các ngôn ngữ Phương Đông*, vol. 47-53. Hà Nội: Viện Ngôn ngữ học, 1986.
- [11] T. D. Trần, "Issues on lexicon and phonetics of the Chứt language contributing to the study of historical phonetics of Vietnamese," Ph.D Dissertation, Trường Đại học Tổng hợp Hà Nội, Hà Nội, 1986.
- [12] PraatSauce-master. (2018). <https://github.com/kirbyj/praatsauce>.
- [13] T. D. Trần, "Notes on tones in Arem dialect," *Khoa Học*, vol. 1990, no. 2, pp. 37-40, 1990.

Bidirectional Consonantal Effects on Vowel F0 Production in Modern Burmese

Tianyi Ni
The Ohio State University

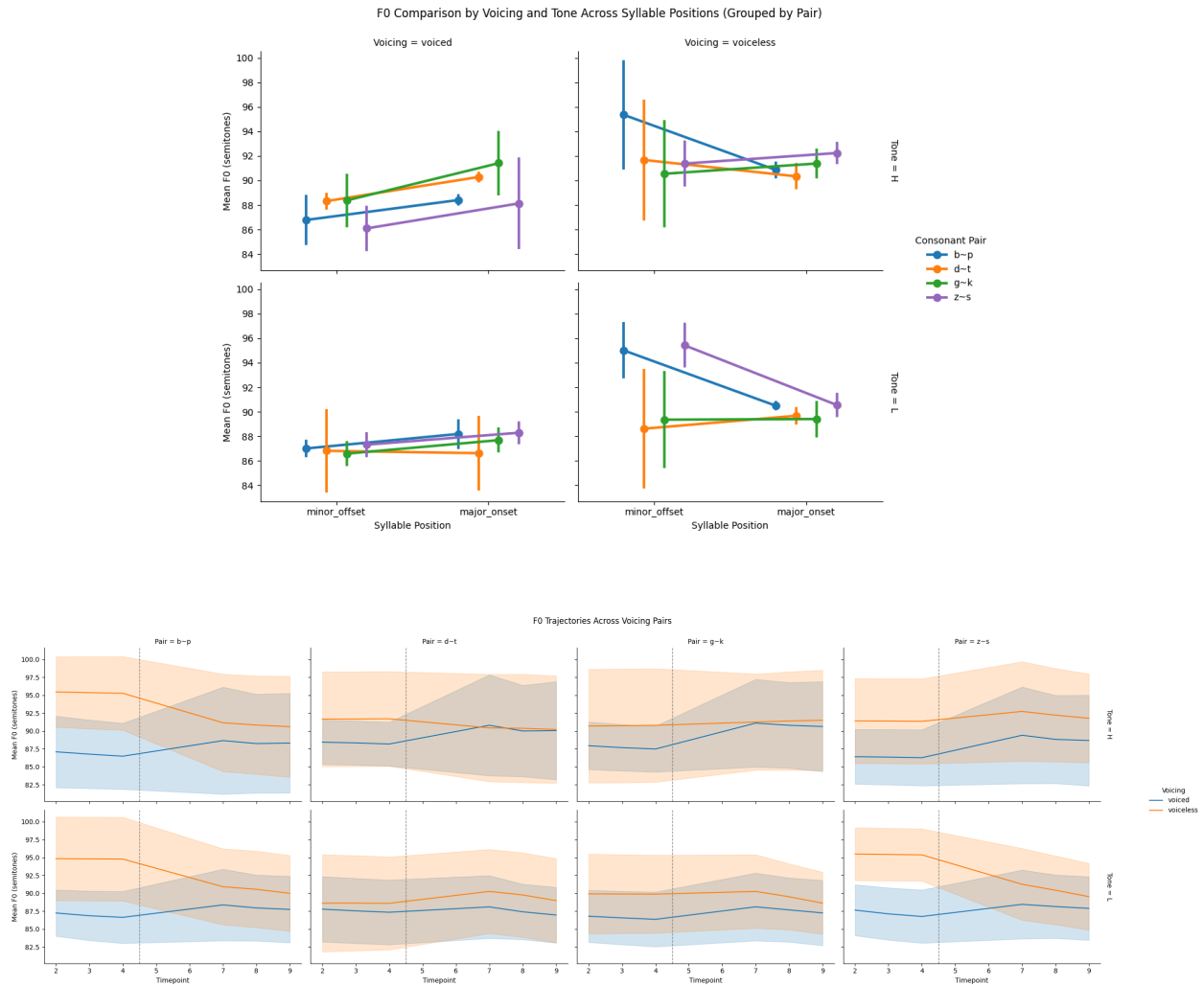
Issue—The effect of consonantal voicing on the fundamental frequency (F0) of surrounding vowels has been widely studied. Cross-linguistically, voiceless and aspirated consonants have been shown to elevate the onset F0 of following vowels in languages such as Cantonese (Zee, 1980), Danish (Jeel, 1975), and Eastern Khmu (Kirby et al., 2024), whereas in languages like Mandarin (Xu and Xu, 2004) and Hindi (Kagaya and Hirose, 1975), aspiration lowers F0. Most previous studies, including recent work by Kirby et al. (2024), primarily examine the impact of consonants on the immediately following major (tonally specified) syllables. By contrast, minor syllables, often described as unstressed and toneless, have received relatively little attention in the context of consonant-induced F0 perturbations, despite exhibiting distinct pitch properties in some languages (Yip, 2002; Zhu, 2022). Modern Burmese, a tonal language with four distinctive tones (high, low, creaky, and checked), realizes these tones exclusively in major syllables, while minor syllables—occurring only in non-final positions—are traditionally considered toneless and largely overlooked. Crucially, the question of whether consonant voicing affects pitch in preceding minor syllables, reflecting anticipatory effects, remains unexplored. This study addresses this gap by investigating bidirectional consonant-induced pitch effects, focusing on whether voicing distinctions influence F0 not only at the onset of following major syllables but also at the offset of preceding minor syllables.

Experiment Design—Four native speakers of Burmese from Yangon (two males, two females), aged between 18 and 35 years, participated in this study. Each speaker produced 8 minimal pairs contrasting consonant voicing in stops (/p, b, t, d, k, g/) and fricatives (/s, z/). Each consonant was embedded between a preceding minor syllable /ʔə/ and a following major syllable /a/, which carried either a high or low lexical tone (e.g., /ʔə-pá/ high vs. /ʔə-pà/ low, with acute and grave accents denoting high and low tones, respectively). These minimal pairs were embedded in the carrier sentence: *mín ____ sá-dè* ("You eat ____."). Each token was repeated three times per speaker. F0 contours were extracted at 10 equally spaced time points per syllable using ProsodyPro, then normalized within each speaker and converted to semitones for comparability. Statistical analysis targeted minor syllable offsets (time points 8–10) and major syllable onsets (time points 1–3) to investigate bidirectional consonantal voicing effects. A linear mixed-effects model was implemented in Python (statsmodels).¹

Results and Discussion—Our model reveals that consonantal voicing significantly modulates F0 in Burmese. Voiceless consonants consistently elevated F0 compared to voiced consonants, producing significant local pitch perturbations ($\beta = 1.609, SE = 0.628, z = 2.563, p = 0.010$). Importantly, the magnitude of this voicing-related F0 raising differed significantly across syllable boundaries, with notably stronger anticipatory effects at the offset of minor syllables compared to carryover effects at the onset of major syllables ($\beta = 3.604, SE = 0.828, z = 4.351, p < 0.001$). The prosodic differences between minor and major syllables were reflected in significantly different baseline F0 values, with minor syllable offsets exhibiting lower overall pitch than major syllable onsets ($\beta = -2.451, SE = 0.645, z = -3.799, p < 0.001$). Crucially, the voicing effect remained stable across tonal contexts, as indicated by the non-significant Voicing \times Tone interaction ($\beta = 0.468, SE = 0.825, z = 0.567, p = 0.571$), suggesting consonantal pitch perturbations are robust regardless of lexical tone beared by the following major syllable. Gender differences in F0 were not statistically significant.

This study also challenges the traditional view that toneless syllables have either default or polarized F0. Instead, we show that F0 on minor syllables is systematically influenced by the voicing of the following consonant. Moreover, the results reveal that consonant voicing can simultaneously affect both adjacent vowels, not just one directionally. This bidirectional effect offers a novel perspective on tonogenesis, suggesting that tonal contrasts may arise not only from effects on the following vowel but also from anticipatory effects on the preceding vowel.

¹Semitone \sim Voicing \times SyllablePosition \times Tone + Gender + (1 | Speaker).



References

- V. Jeel. An investigation of the fundamental frequency of vowels after various danish consonants, in particular stop consonants. *Annual Report of the Institute of Phonetics University of Copenhagen*, 9:191–211, 1975.
- R. Kagaya and H. Hirose. Fiberoptic electromyographic and acoustic analyses of hindi stop consonants. *Annual Bulletin of the Research Institute of Logopedics and Phoniatics*, 9:27–46, 1975.
- J. Kirby, R. Puggaard-Rode, F. Burroni, and S. Maspong. Effects of coda consonants on preceding vowel f0 in eastern khmu. *Proceedings of Speech Prosody 2024*, 2024. Leiden, The Netherlands.
- Y. Xu and C. Xu. Effects of consonant aspiration on mandarin tones. *Journal of the International Phonetic Association*, 34(2):165–181, 2004.
- M. Yip. *Tone*. Cambridge University Press, 2002.
- E. Zee. The effect of aspiration on the f0 of the following vowel in cantonese. *UCLA Working Papers in Phonetics*, 49:90–100, 1980.
- Y. Zhu. Variable pitch realization of unparsed moras in suzhou chinese: Evaluation through f0 trajectory simulation and classification. *Proceedings of the Annual Meetings on Phonology*, 2022.

The intonational spectrum: A Fourier-based method of quantifying tonal rhythm

Argyro Katsika¹, Louis Goldstein², Khalil Iskarous²

¹University of California at Santa Barbara, ²University of Southern California

For many decades, discussion of *speech rhythm* referred to quasiperiodicity in temporal and/or amplitude properties of speech. More recently, in work on prosodic typology, a new dimension of speech rhythm has been increasingly explored—quasiperiodicity in the tonal realm, which has been named macro-rhythm [1]. This is defined as phrase-medial regularity in tonal patterns, and has been argued to be an important dimension of typological differentiation of prosodic systems, some of which, like Korean and Greek, have a strong macro-rhythm, with frequent alternation of Low and High tones, while English, for instance, has a weaker one [1]. Towards the goal of understanding the prosodic structure of languages, therefore, it is important to be able to quantify tonal rhythm, and to relate it to quantification of amplitude rhythm [2]. One approach to quantifying tonal rhythm is local and investigates variability between successive troughs and peaks of F0 [1,3,4], while another takes a more global approach quantifying the path length of the F0 contour, since a flatter rhythm would have a shorter path length than a more oscillatory one [5]. We propose here a new global method, previously used for quantifying amplitude rhythm [2], and based on the oldest signal processing method, the Fourier Transform. At this early stage of quantification of tonal rhythm, we believe it is important for the researcher to have several approaches to choose from, and which can be compared. Our goal is to measure global F0 regularity, regardless of whether the fundamental units of the tonal rhythm are tied to the accentual phrase or to prosodic words and phrases [1]. **Methods:** We used the CommonVoice database, used for Automatic Speech Recognition system training, and selected Greek, Korean, and English datasets. We selected 5000 random sentences from each language, and used Matlab® pitch estimation NN to extract F0. An LPC-smoothed spectrum with 1 pole was then computed for each F0 contour. LPC-smoothing was used since the DFT spectrum by itself is extremely noisy due to the inexactness of F0 estimation. No attempt was made to pre-fit the F0 contours with a smoothing curve to avoid the sharp drops in the voiceless portions, since this could bias the spectrum estimation. Voiced-Voiceless alternation contributes to the high frequency aspect of these intonational spectra, therefore we left these alternations and only quantified differences between the languages at the lowest frequencies, where the tonal rhythm is expected. For each LPC-spectrum, we calculated the mean of the Fourier amplitude at the lowest 2% of frequencies. **Predictions:** Since Korean and Greek are expected to have stronger tonal rhythm than English, the expectation is that they would have a higher peak at the lowest frequencies. **Results:** Figure 1 shows the distributions of the mean amplitude of the lowest 2% of frequencies across the large database of 5000 utterances per language. We used a robust Cohen's d measure of inter-distributional distance, which measures the effect size of difference, to quantify the differences between languages. The Cohen's d distance between English and Greek was .25 standard deviations, which is considered small, and the Cohen's d distance between English and Korean was .71 standard deviations, which is considered large. There was also a Cohen's d distance of .5 between Greek and Korean. **Discussion:** The intonational spectrum method shows the expected difference in tonal rhythm between English and Korean, but for Greek, the results are less clear. One possibility is that aspects other than phrase-medial tonal regularity are at play in Greek. We speculate that the position of stress in the word could play a role in differentiating Greek from Korean, which doesn't have stress. We believe the results, however, are promising

enough for this method to be compared with others, and opens the door for using other spectral methods which may give smoother estimates.

References

[1] Jun, S-A. 2014. Prosodic typology: by prominence type, word prosody, and macro-rhythm. In: Jun, S-A. (ed), *Prosodic Typology II*. Oxford University Press, 520-539. [2] Tilsen, S. and K. Johnson (2008). Low-frequency Fourier analysis of speech rhythm. *JASA Express Letters*, 124 (2). [3] L. Polyanskaya, M. G. Bus, and M. Ordin, “Capturing Crosslinguistic Differences in Macro-rhythm: The Case of Italian and English,” *Language and Speech*, vol. 63, no. 2, pp. 242–263, Jun. 2020. [4] C. Prechtel, “Macro-rhythm in English and Spanish: Evidence from Radio Newscaster Speech,” in *Speech Prosody 2020*. ISCA, May 2020, pp. 675–679. [5] Kaland, C. (2022). Bending the string: intonation contour length as a correlate of macro-rhythm, *Interspeech 2022*. [6] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, M., and Weber, G. (2019). Common Voice: A Massively-Multilingual Speech Corpus. CoRR, Vol. abs/1912.06670 (2019). arxiv: 1912.06670 <http://arxiv.org/abs/1912.06670>.

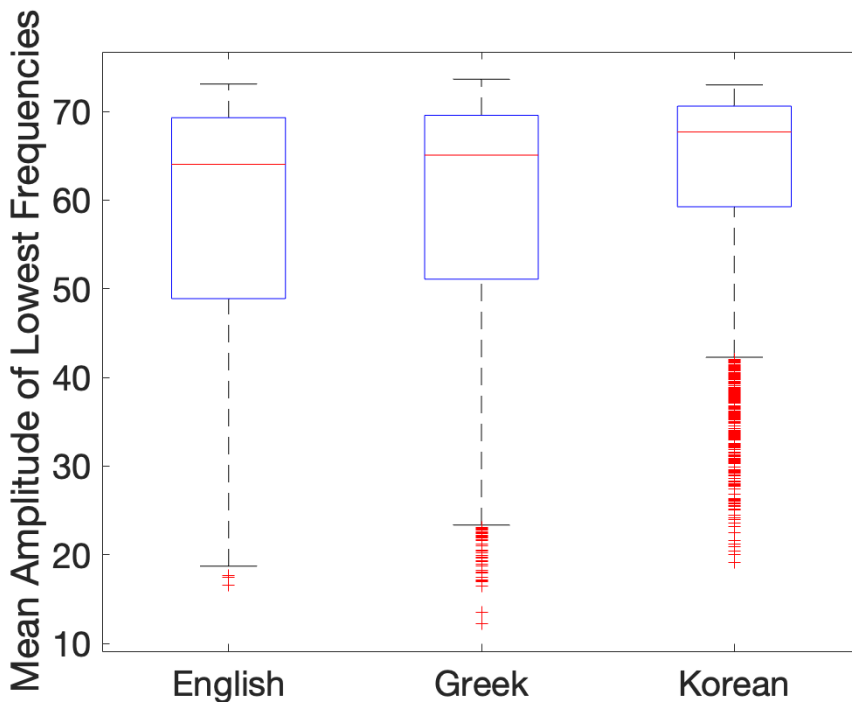


Figure 1 Mean amplitude of lowest 2% of frequencies for the 3 languages.

Phonologization of pitch in Phnom Penh and Standard Khmer for cluster and disyllabic words

Sothornin Mam¹, Pittayawat Pittayaporn¹ & Sireemas Maspong²

¹*Department of Linguistics & Center of Excellence in Southeast Asian Linguistics,
Chulalongkorn University, Bangkok, Thailand*

²*Institute for Phonetics and Speech Processing (IPS), LMU Munich, Germany
6681006122@student.chula.ac.th, pittayawat.p@chula.ac.th,
s.maspong@phonetik.uni-muenchen.de*

In Phnom Penh Khmer (PPK), the colloquial variety used in daily speech of the Phnom Penh area, the [r] of Standard Khmer (SK), the variety used in formal settings, education, and by news reporters, has been replaced with an aspiration [h] accompanied by a falling-rising pitch in the following vowel, thus creating minimal pairs such as [c^hu:k] ‘lotus’ and [c^hù:k] ‘pig’, corresponding to SK [c^hu:k] and [cru:k], respectively [1]. This change was viewed as an incipient of tonogenesis as their distribution is very limited [2]. However, a previously undocumented phenomenon—pitch lowering in the second syllable after voiced onsets in disyllabic words—suggests that tonogenesis may be well underway. Although lowering in pitch is an intrinsic property of voiced consonants, it can become phonologized if the effect is further exaggerated by the speakers to the point that it cannot be explained by universal phonetics [3]. In this study, we investigate f₀ patterns in PPK monosyllables with onset clusters and disyllables, comparing them with SK pitch realizations. We aim to determine whether the pitch pattern has been phonologized in the PPK sound system.

Data. The wordlist includes cluster and disyllable pairs where the last onset consonant and rime match. We examine four types of words according to their last onset consonants: those with 1) r>h, 2) voiced stops, 3) voiceless stops, and 4) nasal, totaling 50 words. Examples of target words are shown in Table 1. Five native PPK speakers (two males and three females) who are all literate in SK were first asked to say the words in a frame sentence [ɲom ɲij tha: ____ tiət] ‘I speak the word ____ again’ in the way they usually talk with their friend or family, [ɲom] and [ɲij] being PPK forms. They were then asked to say the word within the frame sentence [kɲom ni.jij tha: ____ tiət] as if they were a news announcer, [kɲom] and [ni.jij] being SK forms.

Analysis. F₀ values were extracted from the rime of the target words, which are time-warped into 51 equidistant points. All the trajectories are fitted into a Functional Principal Component Analysis (FPCA) model and transformed into principal component (PC) scores as the representation of each trajectory. A linear mixed-effects regression (LMER) model was performed with PC score as the dependent variable. The model included onset (r>h, voiced, voiceless, nasal), syllable structure (cluster, disyllable), language variety (SK, PPK), and their interaction terms as the fixed effects, with speaker as a random effect.

Results. After fitting all the curves to FPCA, the first PC score, which represents pitch height, is implemented. This PC score captures 81% of all the curves’ variance. LMER models’ results reveal a significant effect of the three-way interaction coefficient of the fixed effect variable, confirming the difference in f₀ across linguistic dimensions. The cluster with voiced onset is significantly lower than the other onset in SK, shown as having overlapping curves in Figure 1, suggesting the lowering of f₀ with voiced onset. While in disyllable, the model predicted voiceless onset as having the highest f₀ height, which is also significantly greater than other onsets. This difference is magnificently larger in PPK, clearly illustrating two categorically different f₀ groups. As expected, in PPK, the r>h cluster patterns with the lower f₀ group of disyllable structure, although its predicted height in SK is one of the highest. Significant effects were also found when comparing language varieties for the cluster of r>h and disyllables of r>h, voiced and nasal onsets, as shown in Figure 2.

Discussion. Our study reveals that, in PPK, in addition to words with r>h, disyllables with voiced stops and nasal onset form a distinct group of pitch contour compared to other word groups, where pitch is manifested in the higher range. The small difference between voiceless and other onset types in SK suggests a phonetic effect of voicing and syllabicity on the following pitch contour, while the huge categorical difference in PPK suggests the pitch difference has been phonologized into a binary phonological distinction between a high and a low tone. Although whether the speaker makes use of this acoustic cue in perception remains an important question, these results suggest that tonogenesis in PPK may be more advanced than originally believed.

Keywords: tonogenesis, phonologization, Khmer, sound change, language variety, syllable structure

Table 1 Example of word category implemented (pitch and cluster transition are not specified)

Onset type	Cluster		Disyllable	
	SK	PPK	SK	PPK
r>h	[criəw] ‘to stir’	[c ^h iəw]	[kaŋ.criəw] ‘to be loud’	[k ^ə c ^h iəw]
voiced stop	[tḅouŋ] ‘south’	[tḅouŋ]	[ḁm.ḅouŋ] ‘first’	[t ^ə ḅouŋ]
voiceless stop	[pka:] ‘flower’	[pka:]	[prɔː.ka:] ‘type’	[p ^ə ka:]
nasal	[cniəŋ] ‘scoop net’	[cniəŋ]	[cum.niəŋ] ‘spirit’	[c ^ə niəŋ]

Figure 1 PC1 (representing pitch height) predicted by the LMER model for SK and PPK, color-coded by onset type, where solid and dashed lines represent syllable type.

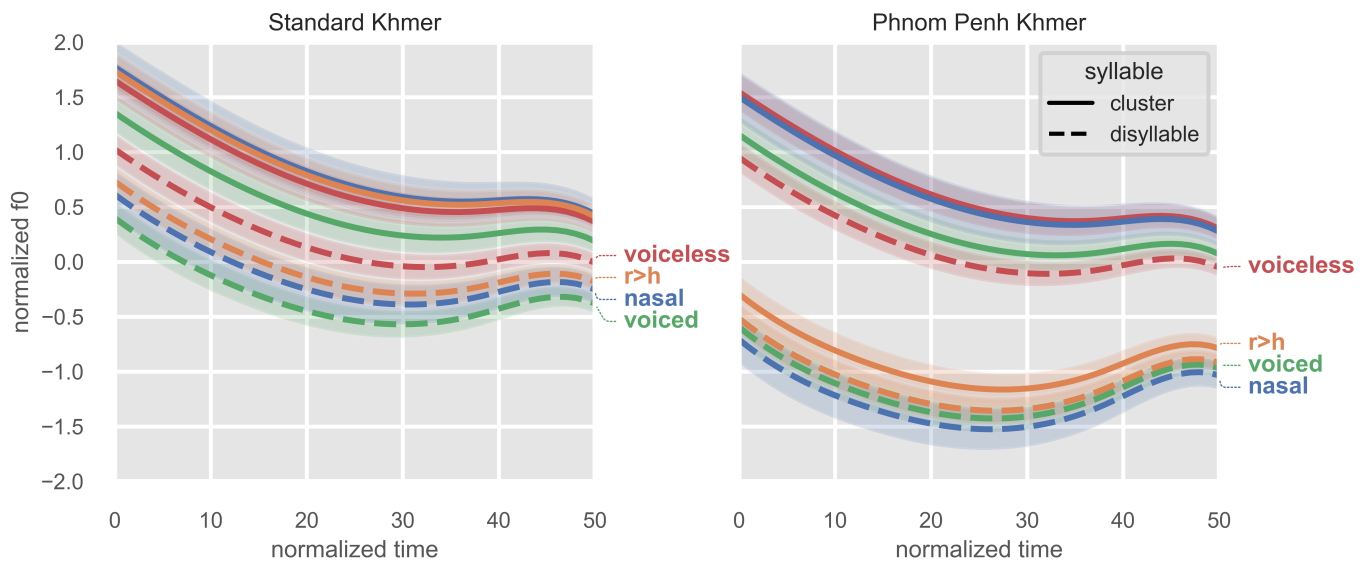


Figure 2 Boxplot of PC1 (representing pitch height) for all word categories comparing SK and PPK.

References

- [1] N. Pisitpanporn, "On the r>h shift in Phnom Penh Khmer," *Mon-Khmer Studies*, vol. 24, pp. 105-113, 1995.
- [2] J. Kirby, "Incipient tonogenesis in Phnom Penh Khmer: Acoustic and perceptual studies," *Journal of Phonetics*, vol. 43, pp. 69-85, 2014/03/01/ 2014, doi: <https://doi.org/10.1016/j.wocn.2014.02.001>.
- [3] L. M. Hyman, "Enlarging the scope of phonologization," in *Origins of Sound Change: Approaches to Phonologization*, A. C. L. Yu Ed. United Kingdom: Oxford University Press, 2013.

The Role of Prosodic and Semantic Cues in Preschoolers' Perception of Emotional Speech

Ruowei Wu¹, Zheyu Song¹, Yi Li¹ and Ping Tang^{1*}

¹*School of Foreign Studies, Nanjing University of Science and Technology, China*

ping.tang@njjust.edu.cn

Emotional speech perception relies on two key cues: semantics and prosody. It has been shown that prosody dominates over semantics in adults' emotion perception, as revealed by the emotional Stroop task, which requires participants to focus on one cue (e.g., prosody or semantics) while ignoring the other when cues are incongruent [1-3]. In contrast, preschoolers in Western cultures relied more on semantic cue [4-5]. However, the role of these two cues among Mandarin-speaking preschoolers remains unclear. So far studies on Mandarin-speaking preschoolers have primarily used the emotional Stroop task, but findings have been inconsistent [6-7]. Furthermore, given the cognitive demands of the Stroop task, which may overwhelm preschoolers due to developing attention and inhibition skills [8], alternative paradigms are needed to better investigate the role of emotional cues in them.

To address these issues, we have modified the emotional Stroop paradigm following [9]. Instead of requiring preschoolers to focus on one cue while ignoring the other cue, we asked them to evaluate the prosody-semantic incongruent emotional expressions in a holistic way. Given the tonal nature and high-context style of Mandarin, we predicted that prosodic cue would play a more important role than the semantic cue when the two cues are incongruent. Furthermore, as they grow older, the role of prosodic cue would gradually become increasingly pronounced, aligning with the adult pattern of emotional perception.

The experiment recruited 24 3-year-old children, 29 4-year-old children, 28 5-year-old children and 27 adult controls. The stimuli consisted of 48 Mandarin sentences conveying two emotions (i.e., happiness and sadness) through both prosody and semantics, recorded by a native female speaker. They were designed to be either congruent (both cues expressed the same emotion) or incongruent (prosody and semantics conveyed different emotions). In each trial, participants were randomly presented with a target stimulus and then asked to select one of two cartoon faces that matched the emotion of the speech: one showing happiness and the other showing sadness.

The results revealed that all participants accurately identified emotion types in the congruent condition (Figure 1). In the incongruent condition, adults predominantly judged emotion types based on prosody, as reflected by significantly higher prosody-based responses as compared to semantic-based responses, with prosody-based responses significantly higher than 50% chance-level (Figure 2). However, only 4- and 5-year-olds showed more prosody-based responses relative to semantic-based responses, and only 5-year-olds' prosody-based responses were significantly higher than chance-level.

Our findings suggest that Mandarin-speaking preschoolers at 4 years start to show an adult-like cue-weighting strategy in emotional perception, shifting from an unspecialized strategy to a prosody-dominant strategy. To be more specific, 3-year-olds had not yet developed a stable preference for either prosody or semantics, 4-year-olds showed a weak prosodic bias, and only 5-year-olds demonstrated a significant adult-like reliance on prosodic information. These findings enhance our understanding of how Mandarin-speaking preschoolers integrate prosodic and semantic cues in emotional perception and highlight the importance of cross-culture research in emotional perception.

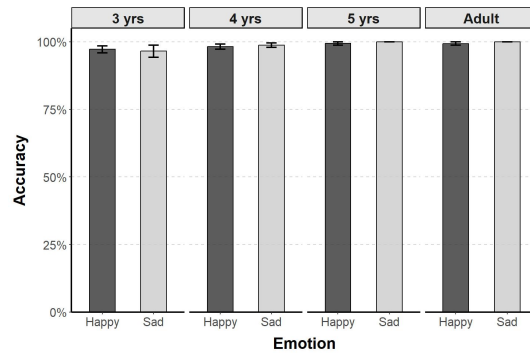


Figure 1. Emotional perception accuracy in the congruent condition.

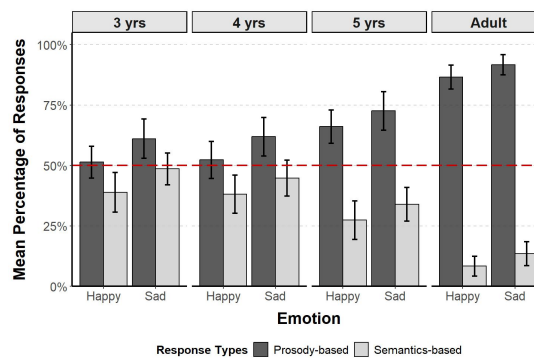


Figure 2. Percentage of prosody-based and semantics-based responses across different age groups in the incongruent condition.

References

- [1] Ben-David, B. M., Multani, N., Shakuf, V., Rudzicz, F., & van Lieshout, P. H. (2016). Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research*, 59(1), 72-89.
- [2] Filippi, P., Ocklenburg, S., Bowling, D. L., Heege, L., Güntürkün, O., Newen, A., & De Boer, B. (2017). More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing. *Cognition and Emotion*, 31(5), 879-891.
- [3] Lin, Y., Ding, H., & Zhang, Y. (2020). Prosody dominates over semantics in emotion word processing: Evidence from cross-channel and cross-modal Stroop effects. *Journal of Speech, Language, and Hearing Research*, 63(3), 896-912.
- [4] Aguert, M., Laval, V., Le Bigot, L., & Bernicot, J. (2010). Understanding expressive speech acts: The role of prosody and situational context in French-speaking 5-to 9-year-olds.
- [5] Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child development*, 72(3), 834-843.
- [6] Kuang, C., Chen, X., & Chen, F. (2024). Recognition of Emotional Prosody in Mandarin-Speaking Children: Effects of Age, Noise, and Working Memory. *Journal of Psycholinguistic Research*, 53(5), 68.
- [7] Yang, Y., Wang, L., & Wang, Q. (2021). Take Your Word or Tone for It? European American and Chinese Children's Attention to Emotional Cues in Speech. *Child development*.
- [8] Adleman, N., Menon, V., Blasey, C., White, C., Warsofsky, I., Glover, G., & Reiss, A. (2002). A developmental fMRI study of the Stroop color-word task. *Neuroimage*, 16(1), 61-75.
- [9] Kikutani, M., & Ikemoto, M. (2022). Detecting emotion in speech expressing incongruent emotional cues through voice and content: investigation on dominant modality and language. *Cognition and Emotion*, 36(3), 492-511.