INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

The different enhancement roles of covarying cues in Thai and Mandarin tones

Nari Rhee, Jianjing Kuang

University of Pennsylvania
nrhee@sas.upenn.edu, kuangj@ling.upenn.edu

Abstract

Any phonological contrast is distinguished by multiple covarying cues. The role and nature of the cue covariation have been much debated in the literature, identifying the physiological link in production, perceptual integration, and enhancement as the key factors in play. In this study, we test whether the enhancement role of covarying cues are influenced by yet another factor: the interaction between cues at multiple layers of phonological structure. We hypothesize that further enhancement of the cue covariations occurs in contexts where acoustic cues are in competition for phonological contrasts at multiple prosodic levels. To test this, we investigate the enhancement role of the covariation relationship between F0 and spectral cues in Mandarin and Thai tones in different phrasal positions. Exploratory and multidimensional-scaling analyses suggest that in Mandarin, the phrase-final weakening of F0 cues are compensated by enhancing the spectral cues, while in Thai, tonal category-specific enhancement is observed at all phrasal positions. The context- and category- specific enhancement of covarying cues suggests that the language's phonological structure plays an important role in the fine-tuning of cue covariations.

Index Terms: tones, enhancement, Mandarin, Thai

1. Introduction

Any phonological contrast involves more than one acoustic dimensions. Most notably, the stop voicing contrast in English is distinguished by at least 16 different acoustic cues [1]. The contribution of each cue to the distinction between phonological categories varies, but many studies have shown that the covarying cues are indeed linguistically useful in maintaining phonological contrasts. For instance, the voicing contrast in English is primarily distinguished by the voice onset time (VOT), yet the voicing perception can be influenced by other covarying cues, such as the first formant (F1) frequency [2, 3] and transition duration [4], and F0 [5] of the following vowel. Another example of linguistically useful cue covariation is the one between F0 and voice quality, found especially in tonal languages where phonation cues influence the tonal recognition regardless of whether the primary cue, F0, is neutralized (Mandarin: [6]) or present (Mandarin: [7]; Cantonese: [8]), but also more generally in the pitch perception in speech [9].

The role and source of these covarying cues have been hotly debated, yet decades of research has not reached a consensus on the explanation as to how and why cues covary. At one end, it has been proposed that cue covariation is an accidental acoustic consequence of the articulatory gestures associated with distinctive features (see [10] for a review). For example, F0 perturbations during consonant voicing is related to the aerodynamics of consonant phonation, influencing the tension in the cricothyroid muscles [11, 12]. Similarly, voice quality covaries with F0 because non-modal phonation can be driven by extreme F0 targets within a speaker's F0 range [13], as stiffer vocal folds during the production of higher F0 targets cause tenser phonation [14]. However, others have argued that the physiological account is insufficient to explain all cue covariation patterns, such as the cue shifting phenomena between covarying cues (e.g. tonogenesis in Korean [15]) and the cross-linguistic differences in the realization of cue covariations. At the other end, versions of the Enhancement theory were put forward as alternative explanations for cue covariations [16, 17]. Under this view, the role of cue covariations is to enhance phonological contrasts, either by introducing additional acoustic attributes to the contrast, or by enhancing perceptually integral auditory dimensions [18]. Crucially, the Enhancement theoretic explanation of cue covariations predicts that cue covariations are language-specifically driven to enhance phonological contrasts, in comparison with the previous view that cue covariations are universally driven by physiological links.

1.1. The current study

In this paper, in addition to testing whether cue covariations serve the role of language-specific enhancement, we further investigate how a language's phonological structure influences the cue covariations and their enhancement roles. If the purpose of cue covariations is phonological contrast enhancement, one can predict that in contexts where the contrasts are weaker in one acoustic dimension, the available covarying cue dimension should take a more important role in distinguishing the categories. We hypothesize that the fine-tuning of the cue covariation enhancement will depend on the interaction between phonological categories and prosodic levels.

One suitable test case for language- and context- specific enhancement role of cue covariations is the tonal realization of Mandarin and Thai in different phrasal positions. Both languages have been traditionally described to have tones that are primarily distinguished by the F0 cues. The robustness [19] and usefulness (e.g. [20, 7]) of covarying voice quality cues such as spectral tilt, noise, and creaky phonation have been reported for Mandarin; however, little is known about whether Thai tones also exhibit systematic and useful voice quality cues.

Compared to the Mandarin tonal system, which has four tones (T1, high level; T2, mid-rising; T3, low-dipping; and T4, falling) primarily distinguished by F0 height and contour, the tonal system in Thai has more complex realizations and restrictions based on the syllable structures and phrasal positions. The static high tone is realized with a gradual rising contour, while the mid and low tones have gradual falls towards the end of the syllable [21, 22]. The falling tone rises to its highest point early and falls steeply, and the rising tone falls to the lowest point and then rises. Furthermore, syllables with short vowels closed by obstruent codas can only bear high and low tones, and syllables with long vowels closed by obstruent codas can only bear low and falling tones [21, 22]. On top of coda restrictions, it has been found that syllables with unaspirated obstruents in the onset also cannot bear high or rising tones [23].

http://dx.doi.org/10.21437/Interspeech.2020-1685

The two tonal languages also importantly differ in their phrasal prosodies: the phrase-final position is subject to final weakening in Mandarin [24] and steep F0 declination [25], while in Thai, the phrase-final position is a phrasal stress position where full F0 contours of the tones can surface [26]. Consequently, if the enhancement role of cue covariation is language- and context- specific, we expect to observe a context-specific enhancement of covarying cues to compensate for the phrase-final weakening in Mandarin, but not in Thai. In this study, we explore the language- and position- specific enhancement roles of cue covariations in these two languages, to better understand how the language's phonological structure influences the covariation relationship between cues.

2. Methods

2.1. Data

We used the Mandarin [27] and Thai [28] collections from the Global TIMIT project. The present study is based on 3000 utterances from each corpus, produced by 50 Mandarin and 50 Thai speakers. Both corpora were phonetically segmented using an HMM/GMM-based forced alignment system [29].

For every syllable, F0 (STRAIGHT; [30]), Harmonics-to-Noise ratio (HNR), Cepstral Peak Prominence (CPP), and spectral tilt measures via relative amplitude differences of the lower and higher harmonics (H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H4*-H2K*, H2K*-H5K*, corrected for the influence of formant frequencies and bandwidths on the harmonics; [31]) were measured from 10 equidistant subsegments from the sonorant portion of each syllable, using VoiceSauce [32]. Tokens with F0 tracking errors (exhibiting standard deviation greater than 4) or with duration shorter than 50 milliseconds were removed. Furthermore, Mandarin tone sandhi tokens were removed. All extracted measures were normalized by speaker, using min-max normalization to scale values between 0 and 1.

Using the transcriptions and pauses as a proxy for phrasal boundaries, each token was coded for its position in the Intonation-Phrase (IP), as phrase- initial, medial, or final.

2.2. Analysis

2.2.1. *Contour plots*

The time-normalized contours of F0 and spectral cues were plotted at each phrasal positions, for both languages (Figures 1 and 2). Due to the limitation of space, we present contour plots of F0 and only two of the spectral cues, H1*-H2*, a measure of spectral tilt which has been shown to be related to the open quotient [33] and CPP, a spectral aperiodicity measure. For F0 and H1*-H2*, the normalized range of the cues were calculated by subtracting the minimum from the maximum normalized values within the 10 subsegments of the tokens. For CPP, the range was calculated from the two middle subsegments (5 and 6) of the tokens, to exclude the low tails at the beginnings and ends.

2.2.2. *Multidimensional Scaling (MDS)*

To visualize the multidimensional and highly correlated acoustic space of the tones into a more interpretable low-dimensional ordination space, non-metric MDS was performed on three sets of cues: (i) F0 cues, (ii) spectral tilt and noise cues (H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H4*-2K*, 2K*-5K*, CPP, HNR), and (iii) both F0 and spectral cues (Figures 3 and 4). Non-metric MDS was used to calculate the dissimilarity of the categories in each phrasal position, using the metaMDS function [34] in R, version 3.5.1 [35]. MDS failed to converge when pairwise distance matrices were calculated using all tokens. To prevent convergence failure, the cue sets were averaged for every tonal category in each phrasal position.

To further quantify the overall dispersion of the tones in each phrasal position and language, the average pairwise Euclidean distances (d) among tones were calculated for each MDS plot.

3. Results

3.1. Contour Plots

The contour plots of F0, H1*-H2*, and CPP of Mandarin tones are shown in Figure 1 and Thai tones in Figure 2.

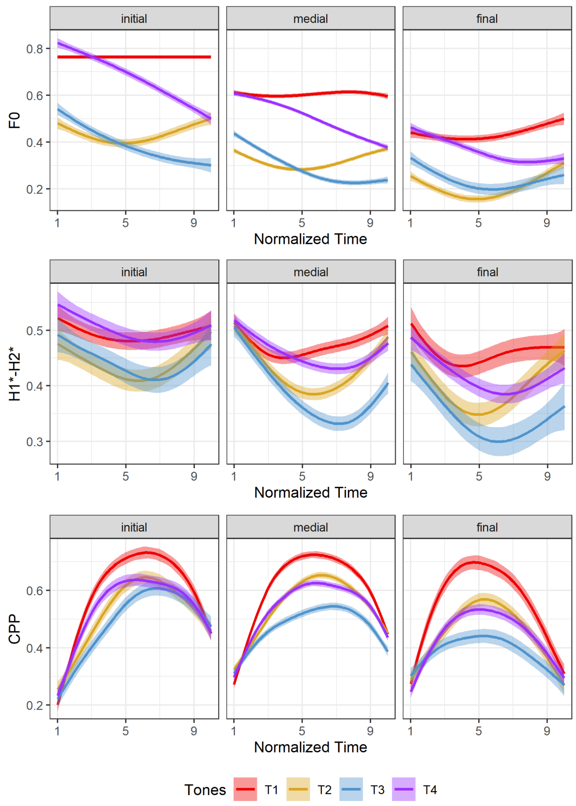

Figure 1: *F0 (top panel), H1*-H2* (middle panel), and CPP (bottom panel) contour plots of Mandarin tones. Phrasal positions are faceted vertically (left: initial; center: medial; right: final).*

The four tones of Mandarin had the most distinct F0 contours and the widest normalized F0 ranges (range: 0.52) in phrase-initial positions. In phrase-medial positions, the tones were distinguished in a smaller F0 range (0.40) but with contours as distinct as those in phrase-initial positions. However, in phrase-final positions, the overall F0 range was the smallest (0.34) and lowest, and the tonal contour distinction between Tones 2 and 3 was particularly weak.

In contrast, phrase-initially, H1*-H2* only had a two-way distinction between higher (T1 and T4) and lower (T2 and T3) tones, and CPP only distinguished T1 from the other tones. Phrase-medially, the tones were distinguished in larger H1*-H2* and CPP ranges (range of H1*-H2*: 0.19; range of CPP:

0.20) than they were in phrase-initial positions (range of H1*-H2*: 0.15; range of CPP: 0.17). The contours of H1*-H2* and CPP were also more distinct phrase-medially (four-way in H1*-H2* and three-way in CPP) than initially. Furthermore, the largest cue ranges and the most distinct tonal contours of H1*-H2* and CPP were observed in phrase-final positions (range of H1*-H2*: 0.24; range of CPP: 0.26).

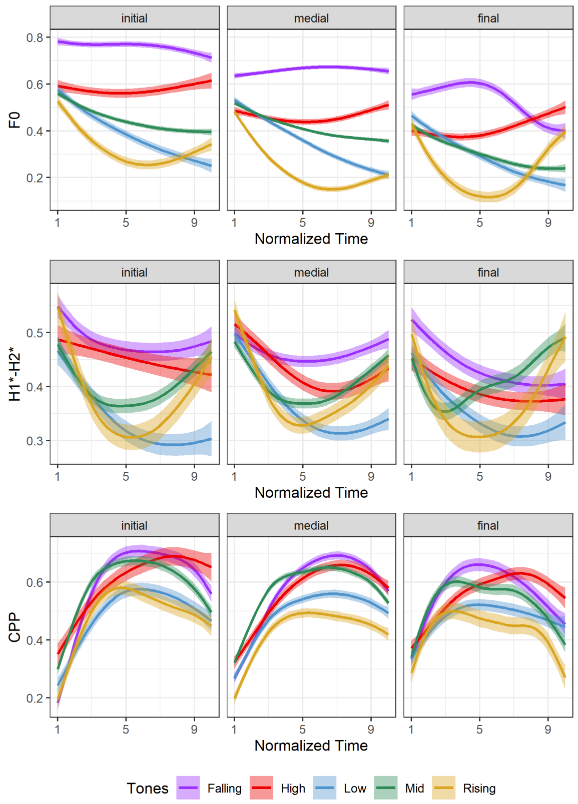

Figure 2: *F0 (top panel), H1*-H2* (middle panel), and CPP (bottom panel) contour plots of Thai tones. Phrasal positions were faceted vertically (left: initial; center: medial; right: final).*

In Thai, the range of the cues did not vary as much by phrasal position as it did in Mandarin (F0 range at initial: 0.54, medial: 0.52, final: 0.49; H1*-H2* range at initial: 0.28, medial: 0.23, final: 0.23; CPP range at initial: 0.15, medial: 0.19, final: 0.19). In phrase-initial and medial positions, the five tones were distinguished more by F0 height than by F0 shape. The falling tone was the highest in F0, which was followed by the high tone, and then by the gradually falling mid and low tones. The only tone that was distinct in terms of F0 contour was the rising tone. In phrase-final positions, the citation F0 contour of the tones largely surfaced, though mid and low tones were overlapping. On the other hand, the values and contours of the spectral cues (H1*-H2* and CPP) clearly distinguished between the low and the mid tones, regardless of the phrasal positions. Furthermore, the rising tone was also distinct from the other tones in its low-rising H1*-H2* contour and low CPP range. Overall, spectral cues did not exhibit enhancement specific to phrasal positions, but contributed to increasing the five-way tonal contrast in all positions.

3.2. Multidimensional Scaling (MDS)

The results of the contour plots were cross-verified by the MDS results. The scaled dimensions their F0 (top panels), spectral (middle panels), and combined (bottom panels) cue spaces for Mandarin and Thai tones are shown in Figures 3 and 4, respectively. For each MDS plot in each phrasal position and cue set, two dimensions ($K = 2$) were sufficient to obtain a good fit of the data ($stress < 0.1$)

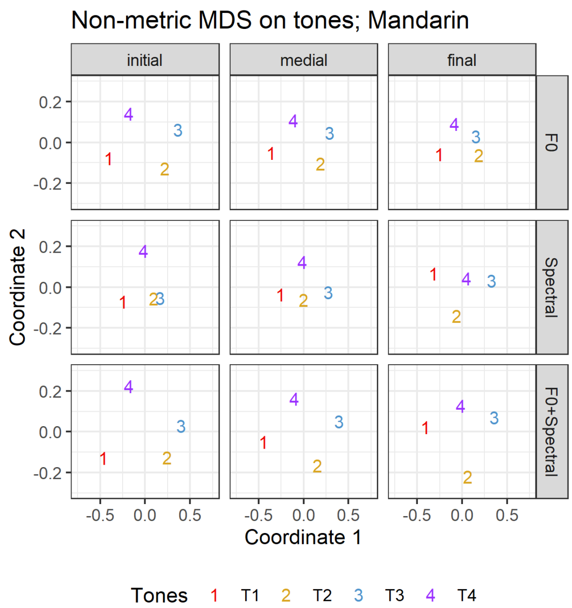

Figure 3: *Non-metric MDS on averages of cues in tonal categories in Mandarin. Phrasal positions were faceted vertically (left: initial; center: medial; right: final), and cue sets were faceted horizontally (top: F0 cues; middle: spectral cues; bottom: both F0 and spectral cues).*

In Mandarin (Figure 3), the four tones were well-separated by just F0 in phrasal-initial (average pairwise distance: $d = 0.50$) or medial positions ($d = 0.41$), but tones 2 and 3 were less distinguished in phrase-final positions, resulting in a lower average pairwise distance ($d = 0.29$). In contrast, spectral cues achieved a worse separation in phrase-initial positions ($d = 0.28$) and phrase-medial positions ($d = 0.31$) than in phrase-final positions ($d = 0.38$). In particular, the reduced tonal space in the F0 cue space in phrase-final positions was compensated by the enlarged spectral cue space, especially for tones 2 and 3. When both F0 and spectral cues were used, all four tones were well-separated in all phrasal positions (initial: $d = 0.57$, medial: $d = 0.52$, final: $d = 0.48$).

In Thai (Figure 4), a different pattern was observed. The dispersion of the tonal categories in the F0 tonal spaces decreased over the course of the utterance, with no clear position-specific contrast enhancement (initial: $d = 0.45$, medial: $d = 0.43$, final: $d = 0.36$), spectral (initial: $d = 0.33$, medial: $d = 0.33$, final: $d = 0.32$), and combined (initial: $d = 0.56$, medial: $d = 0.54$, final: $d = 0.46$). Instead, a more tonal category-specific contrast enhancement was observed for Thai. The falling tone (F) was clearly distinct from all tones in all positions in the F0 space, but mid (M), low (L), and rising (R) tones were less dispersed phrase- initially or medially. While

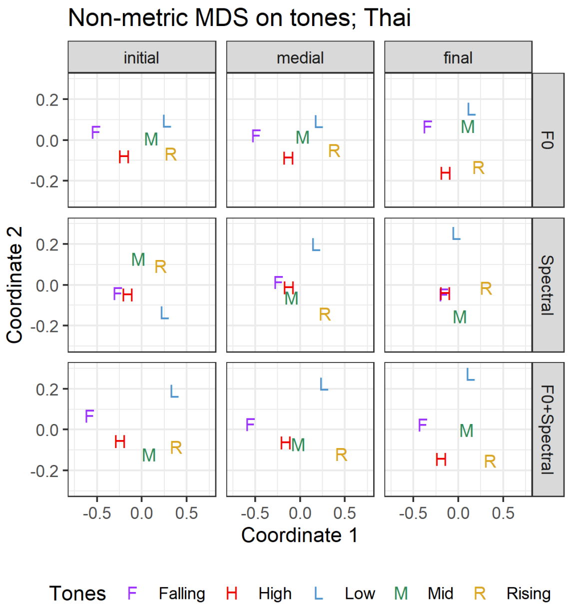

Figure 4: *Non-metric MDS on averages of cues in tonal categories in Thai. Phrasal positions were faceted vertically (left: initial; center: medial; right: final), and cue sets were faceted horizontally (top: F0 cues; middle: spectral cues; bottom: both F0 and spectral cues).*

spectral cues were not sufficient to distinguish all five tones of Thai; the three tones (mid, low, and rising) that were not clearly distinguished by the F0 cues were the most dispersed tones in the spectral cue space. Consequently, just like Mandarin, combining both F0 and spectral cues achieved in a better five-way tonal contrast than using only F0 or only spectral cues in all phrasal positions.

4. Discussion

Visual inspections of the contour plots and the MDS tonal spaces of Mandarin and Thai suggested that manifesting maximal tonal contrasts involved language- and position- specific enhancement through the covarying cues. In both Mandarin and Thai, using a combination of both F0 and spectral cues resulted in maximal dispersion of the tones. In Mandarin, the covarying spectral cues served particularly useful in manifesting the tonal contrast in phrase-final weakening positions, where the range and contrastivity of the F0 cues were most reduced. In comparison, the spectral cues in Thai category-specifically enhanced the tonal contrast of the lower tones (low, mid, and rising) in the spectral space, but this tone-specific enhancement was generally unaffected by phrasal positions. This tonal category-specific enhancement can be partially attributed to the relationship between tones and syllable structures in Thai. In particular, obstruent onsets and codas, which can affect the spectral voice quality cues, limit the tonal types that can surface to tones with a low F0 target [21, 22, 23, 36]. In short, despite their physiologically-natural covariation with F0 [14], spectral cues varied in their roles and usefulness in contrasting tones in the two languages examined, depending on the languages' phonological structures.

More generally, results suggest that while we cannot rule out the hypothesis that cue covariations are initiated by the physiological links, the language's phonological and prosodic structures determine the contrast enhancement role of the covarying cues. Furthermore, the enhancement role of the covarying cues is particularly more important in contexts where the primary cue is jeopardized or weaker, such as in the phrase-final weakening of Mandarin. This pattern resembles a cue trading relationship [37], where a decrease in the contrastivity of one cue can be compensated by increasing the contrastivity of another cue. Such cue-trading relationship elucidates the tight integration of covarying cues in phonological contrasts to maintain maximal contrasts, in a multidimensional and covarying acoustic space where multiple levels and categories are simultaneously distinguished.

While the tonal category-specific cue covariation enhancement in Thai and position-specific enhancement in Mandarin should be verified by perception studies, the results of the present study highlight the role of phonological structure in the shaping of the cue-category mapping. By unveiling the source and nature of the cue covariation, the study contributes a fuller understanding of the organization of the cue space in distinguishing the phonological categories, at multiple prosodic levels.

5. Conclusion

The covariation relationship between F0 and spectral cues in the realization of the tonal contrasts in Mandarin and Thai was examined. In particular, we looked for category- and position-specific enhancement of the cue covariations, to test how the language's phonological structure shapes the covariation relationship between cues. Exploratory and multidimensional-scaling analyses revealed that spectral cues, the covarying cues of F0, serve a position-specific enhancement role in Mandarin, and a category-specific enhancement role in Thai. In conclusion, cue covariations and their contrast enhancement roles are fine-tuned by the phonological and prosodic structures of the languages.

6. References

[1] L. Lisker, ""Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees," *Language and speech*, vol. 29, no. 1, pp. 3–11, 1986.

[2] Q. Summerfield, "Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops," *The Journal of the Acoustical Society of America*, vol. 72, no. 1, pp. 51–61, 1982.

[3] J. R. Benkí, "Place of articulation and first formant transition pattern both affect perception of voicing in English," *Journal of Phonetics*, vol. 29, no. 1, pp. 1–22, 2001.

[4] K. N. Stevens and D. H. Klatt, "Role of formant transitions in the voiced-voiceless distinction for stops," *The Journal of the Acoustical Society of America*, vol. 55, no. 3, pp. 653–659, 1974.

[5] D. H. Whalen, A. S. Abramson, L. Lisker, and M. Mody, "F0 gives voicing information even with unambiguous voice onset times," *The Journal of the Acoustical Society of America*, vol. 93, no. 4, pp. 2152–2159, 1993.

[6] S. Liu and A. G. Samuel, "Perception of Mandarin lexical tones when F0 information is neutralized," *Language and speech*, vol. 47, no. 2, pp. 109–138, 2004.

[7] R. Yang, "The role of phonation cues in Mandarin tonal perception," *Journal of Chinese Linguistics*, vol. 43, no. 1, pp. 453–472, 2015.

[8] K. M. Yu and H. W. Lam, "The role of creaky voice in Cantonese tonal perception," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1320–1333, 2014.

[9] J. Kuang and M. Liberman, "Integrating voice quality cues in the pitch perception of speech and non-speech utterances," *Frontiers in Psychology*, vol. 9, p. 2147, 2018.

[10] P. Hoole and K. Honda, "Automaticity vs. feature-enhancement in the control of segmental F0," *Where do phonological features come from*, pp. 131–171, 2011.

[11] J.-M. Hombert, J. J. Ohala, and W. G. Ewan, "Phonetic explanations for the development of tones," *Language*, pp. 37–58, 1979.

[12] A. Löfqvist, T. Baer, N. S. McGarr, and R. S. Story, "The cricothyroid muscle in voicing control," *The Journal of the Acoustical Society of America*, vol. 85, no. 3, pp. 1314–1321, 1989.

[13] J. Sundberg, I. R. Titze, and R. Scherer, "Phonatory control in male singing: a study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source," *Journal of Voice*, vol. 7, no. 1, pp. 15–29, 1993.

[14] I. R. Titze, "Interpretation of the electroglottographic signal," *Journal of Voice*, vol. 4, no. 1, pp. 1–9, 1990.

[15] Y. Kang and S. Han, "Tonogenesis in early contemporary Seoul Korean: A longitudinal case study," *Lingua*, vol. 134, pp. 62–74, 2013.

[16] K. N. Stevens, S. J. Keyser, and H. Kawasaki, "Toward a phonetic and phonological theory of redundant features," in *Invariance and variability in speech processes*, J. S. Perkell and D. H. Klatt, Eds. Psychology Press, 1986, pp. 426–449.

[17] S. J. Keyser and K. N. Stevens, "Enhancement and overlap in the speech chain," *Language*, pp. 33–63, 2006.

[18] J. Kingston and R. L. Diehl, "Intermediate properties in the perception of distinctive feature values," *Papers in Laboratory Phonology*, vol. 4, pp. 7–27, 1995.

[19] J. Kuang, "Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice," *The Journal of the Acoustical Society of America*, vol. 142, no. 3, pp. 1693–1706, 2017.

[20] A. Belotel-Grenié and M. Grenié, "The creaky voice phonation and the organisation of Chinese discourse," in *International symposium on tonal aspects of languages: With emphasis on tone languages*, 2004.

[21] J. Gandour, "On the representation of tone in Siamese," *UCLA Working Papers in Phonetics*, vol. 27, pp. 118–146, 1974.

[22] B. Morén and E. Zsiga, "The lexical and post-lexical phonology of Thai tones," *Natural Language & Linguistic Theory*, vol. 24, no. 1, pp. 113–178, 2006.

[23] S. Ruangjaroon, "Consonant-tone interaction in Thai: An OT analysis," *Taiwan Journal of Linguistics*, vol. 4, no. 2, pp. 1–66, 2006.

[24] C.-y. Tseng, S.-h. Pin, Y. Lee, H.-m. Wang, and Y.-c. Chen, "Fluent speech prosody: Framework and modeling," *Speech communication*, vol. 46, no. 3-4, pp. 284–309, 2005.

[25] J. Yuan, N. Ryant, and M. Liberman, "Automatic phonetic segmentation in Mandarin Chinese: Boundary models, glottal features and tone," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2539–2543.

[26] P. Peyasantiwong, "Stress in Thai," in *A Conference on Thai Studies In Honor of William J. Gedney*, R. J. Bickner, T. J. Hudak, and P. Peyasantiwong, Eds. Center for South and Southeast Asian studies, the University of Michigan, 1986, pp. 211–230.

[27] J. Yuan, H. Ding, S. Liao, Y. Zhan, and M. Liberman, "Chinese TIMIT: A TIMIT-like corpus of standard Chinese," in *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[28] N. Chanchaochai, C. Cieri, J. Debrah, H. Ding, Y. Jiang, S. Liao, M. Liberman, J. Wright, J. Yuan, J. Zhan, and Y. Zhan, "GlobalTIMIT: Acoustic-phonetic datasets for the world's languages," in *INTERSPEECH*, 2018, pp. 192–196.

[29] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models." in *INTERSPEECH*, 2013, pp. 2306–2310.

[30] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[31] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2283–2295, 2007.

[32] Y.-L. Shue, P. A. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," in *ICPhS*, 2011.

[33] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *the Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[34] J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner, *vegan: Community Ecology Package*, 2018, r package version 2.5-2. [Online]. Available: https://CRAN.R-project.org/package=vegan

[35] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[36] J. M. Perkins, "Consonant-tone interaction in Thai," Ph.D. dissertation, Rutgers University, 2013.

[37] B. H. Repp, "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception." *Psychological bulletin*, vol. 92, no. 1, p. 81, 1982.