THE USE OF TONAL COARTICULATION IN SPEECH SEGMENTATION BY LISTENERS OF MANDARIN

Zhe-chen Guo[1], Shu-chen Ou[2]

[1]The University of Texas at Austin, [2]National Sun Yat-sen University
zcadamguo@utexas.edu, sherryou@mail.nsysu.edu.tw

ABSTRACT

Previous research has shown that fine-grained phonetic information is exploited to segment continuous speech into discrete chunks. For example, a sequence of segments is perceived as constituting a unit if they are coarticulated. The current study examines the use of tonal coarticulation as a speech segmentation cue by Mandarin-speaking listeners in an artificial language learning task. We created a mini nonsense language with three level tones and implemented tonal coarticulation in the form of carryover assimilation, a type of tonal coarticulation commonly found across lexical tone languages. Participants first learned the language by listening to utterances in which the words of the language were concatenated with no intervening pauses and then identified the words in a test. The results indicated that pitting the tonal coarticulatory cue against statistical regularities in the utterances significantly reduced identification accuracy in the test, suggesting that tonal coarticulatory information is used in segmenting speech.

Keywords: Speech segmentation, artificial language learning, tonal coarticulation, Mandarin listeners.

1. INTRODUCTION

Understanding spoken language requires segmentation of connected speech into discrete units. Previous studies (e.g., [11, 16, 20, 22, 23]) have demonstrated that listeners tend to use segmentation cues in a way that reflects phonological regularities in their languages. For example, since English lexical stress is mostly word-initial, English listeners are inclined to perceive prominent syllables as word beginnings ([22]). Recent research further reveals that segmentation behavior exhibits sensitivity to phonetic subtleties in the native speech. For instance, Korean listeners exploit an initial low tone cue more effectively as it comes to increasingly resemble the phonetic realizations of the initial low tone in the Korean accentual phrase ([21]). To extend this line of investigation, this study investigates whether prosodic information as fine-grained as tonal coarticulation is used by Mandarin listeners to segment speech.

Despite being a subphonemic detail, coarticulation can be a robust segmentation cue. At the segmental level, coarticulation refers to phonetic change of one segment due to surrounding segments, and the degree of such change is sensitive to boundaries of units in speech. In general, adjacent segments are more strongly coarticulated with each other when they occur within constituents such as words or phrases than when they span the boundaries of these constituents ([6, 8]). Experiments have shown that when listeners learned to extract "words" of an artificial language from continuous speech streams, their performance is facilitated if word-internal segments are coarticulated but impeded if across-word segments are coarticulated ([7]). This indicates that coarticulation of segments is used to discover discrete units.

Coarticulation also refers to phonetic influence of one suprasegmental or prosodic feature on another. A case in point is the so-called "tonal coarticulation," the fact that the F0 realization of a tone is affected by neighboring tones. It has been found that the strength of tonal coarticulation is conditioned by boundary properties. For example, as revealed by analyses of Mandarin tone production, lexical tones within smaller prosodic domains tend to be more strongly coarticulated (e.g., [12]), suggesting that tonal coarticulatory cues can possibly be helpful for parsing speech into smaller chunks. Nevertheless, while recent segmentation studies ([4, 10]) indicate that listeners of a lexical tone language are sensitive to information carried by distinct tone categories (or at least more so than non-tone language listeners), it remains unclear whether they also exploit fine-grained prosodic details such as tonal coarticulation.

The present work sets out to examine this question, focusing on one commonly observed type of tonal coarticulation: progressive (carryover) assimilation. Tonal coarticulation can be assimilatory or dissimilatory, with the direction being progressive or regressive. Among all possible coarticulatory patterns, carryover assimilation (whereby the F0 realization of a lexical tone is at least partially assimilated to the preceding tone) is found in a wide range of lexical tone languages, including Mandarin ([19, 24]), Tianjin Chinese ([25]), Thai ([9]), Cantonese ([13]), Vietnamese ([3]), etc. It is thus of interest to test whether such a common coarticulatory pattern

subserves segmentation. We addressed the question by conducting an artificial language (AL) learning experiment with Mandarin-speaking listeners.

2. METHOD

2.1 Experimental design

Widely employed to investigate the role of prosody in segmentation, the AL learning experiment has two phases. The first one is a learning phase in which participants listen to continuous streams of speech generated by concatenating the words of an AL (which are nonsense syllable sequences); the second one is a test phase in which they identify the words. Identification accuracy serves as a measure of segmentation performance during the learning.

Our AL learning experiment was modeled after the one by [7] and exposed subjects to the AL under three conditions: single-cue, congruent-cues, and incongruent-cues. For all the conditions, the transitional probability (TP) between two adjacent syllables was a basic cue for extracting out the words in the speech streams of the learning phase: syllable sequences contained within an AL word generally had higher TPs than those that straddled a word boundary. In the single-cue condition, TP information was the sole segmentation cue. The other two conditions supplied tonal coarticulation (in the form of carryover assimilation) as an additional cue. In the congruent-cues one, the coarticulatory cue agreed with TP information as syllables within the words of the AL were tonally coarticulated. In the incongruent-cues one, tonal coarticulation was pitted against TPs by letting tonally coarticulated syllables straddle word boundaries. If tonal coarticulation is exploited, it would be expected that compared with that under the single-cue condition, segmentation would be significantly better under the congruent-cues one, or it would be significantly worse under the incongruent-cues one.

2.2 Materials

The AL consisted of six words, which were trisyllabic sequences constructed from three consonants ([p, t, k]), four vowels ([a, i, u, e]), and three level tones (high, mid, and low level tones, denoted by the diacritics [́], [̄], and [̀], respectively). The words were [pémīkù], [tēpátī], [kípītá], [pūkùpē], [tàtūkí], and [kìkēpè]. To implement the tonal coarticulatory cue and control for its magnitude, they were created such that two adjacent syllables in a word differed by one tone level. The component syllables of the words were individually produced in a monotone by a male native Mandarin speaker and then prosodically normalized by using Praat [2]. They were manipulated to have a duration of 335 milliseconds (ms), with their F0 contours flattened at 126 Hz. The F0 and duration values were the mean F0 and duration values of the syllables prior to the manipulation. Following [4], we created the high and low tones by raising and lowering the flat F0 contour (which served as the mid level tone) by 3.5 semitones, respectively. The manipulated syllables were combined to form the words.

Another set of trisyllabic sequences were created: [tápūkù], [pēkípī], [kítēpá], [tīkùtē], [pètēpá], and [kùpētà]. They were "partwords" that occurred in the speech streams of the AL but spanned a word boundary. They served as distractors in the test phase and had the same tone patterns as the AL words.

Tonal coarticulation was implemented in the form of carryover assimilation and by imposing a smooth F0 transition over the initial portion of the F0 contour of a syllable. Specifically, we raised or lowered the F0 at the onset of the syllable to the offset F0 value of the preceding tone; then, we quadratically interpolated the F0 transition between the onset and the 25% time point into the F0 contour of the syllable. This changed the first quarter of the flat contour into an F0 fall (if the preceding tone was higher) or an F0 rise (if the preceding tone was lower), imitative of the assimilatory effect from the previous tone. Yet, acoustic analyses have shown that in actual Mandarin tone production, the assimilatory effect of the preceding tone was still significant even at the 75% time point of the next tone ([24]). We assessed the effects of tonal coarticulation conservatively by manipulating only the initial 25% of the F0 contour.

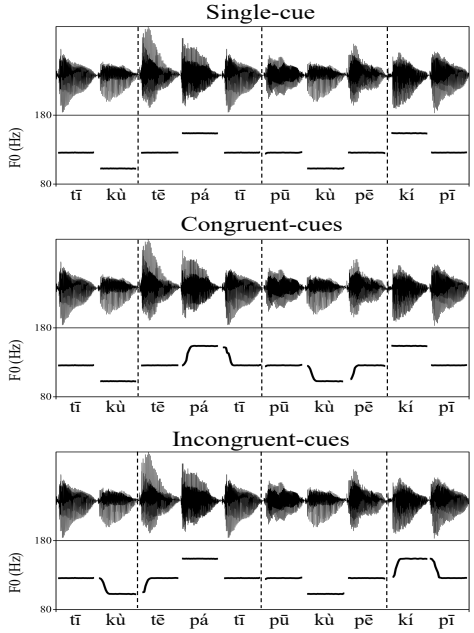
Used in the learning phase were six speech streams in which the six AL words were strung together without pauses. The words each occurred 20 times in each stream, with their occurrences being random but subject to the constraint that tokens of the same word did not appear in a row. In all conditions, the average TP between two adjacent syllables for an AL word ranged from 0.75 to 1.00 (mean: 0.88) while that for a partword ranged from 0.32 to 0.62 (mean: 0.49). As in [22], the first and last five seconds of each stream were faded in and out, preventing access to the syllables at the beginning and the end. The total duration of the six streams was 12 minutes.

To introduce the tonal coarticulatory cue to the AL, we picked out a number of trisyllabic sequences and manipulated them so that the second and third syllables were tonally assimilated to their preceding syllables (i.e., the first and second syllables, respectively) in the fashion described above. These sequences corresponded to all instances of the partwords occurring in the speech streams of the incongruent-cues condition and to 27% of the tokens of the AL words in the speech streams of the

congruent-cues condition. The reason that only 27% of the tokens in the latter condition received the cue was that we followed [7] and kept the numbers of tonally coarticulated syllables equal in the congruent- and incongruent-cues conditions. In this way, the two conditions differed only in the alignment of tonally coarticulated syllables with word boundaries (but not in the number of such syllables). Shown in Figure 1 is a comparison of the AL across the three conditions.

The test was a two-alternative forced-choice test that presented two stimuli in each trial: a word of the AL and a partword. They were presented successively and separated by 500 ms of silence. The stimuli did not have the tonal coarticulation cue; thus, the test was identical for the three conditions. The orders of the words and partwords were balanced, and there were totally 36 trials (6 words × 6 partwords). Stimulus presentation and response recording were conducted by using E-prime 2.0 ([17]).

Figure 1: Examples of speech streams used in the single-cue, congruent-cues, and incongruent-cues conditions. The dashed lines indicate word boundaries.



2.3 Procedure

Tested individually in a sound-attenuated booth, subjects were informed that they were going to learn a novel language by listening to six sound files in which the words of the language were strung together. They were not provided with information about the boundaries or lengths of the words. They were just asked to pay attention to the sound files as best as they could but were made aware of a subsequent test that would assess their knowledge of the language. After the learning phase, subjects proceeded to the test, in which they heard two stimuli in each trial and selected the one that was a word of the language by pressing the button on a response box which corresponded to the presentation order of that stimulus (i.e., "1" or "2"). They had 10 seconds to respond.

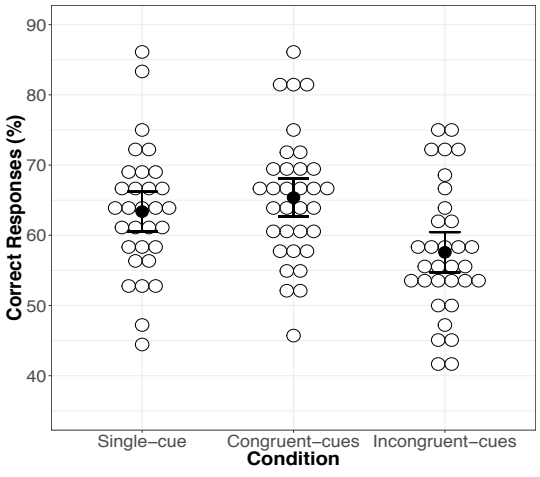
2.4 Participants

A total of 96 native listeners of Mandarin (and Taiwanese Southern Min) were recruited from a university in Southern Taiwan and randomly assigned to each condition. The numbers of listeners of the single-cue, congruent-cues, and incongruent-cues conditions were 31, 33, and 32, respectively. None reported hearing deficits.

3. RESULTS

The listeners' identification accuracy in the forced-choice test was analyzed. Invalid responses (e.g., omissions), which accounted for 0.43% of the data, were discarded. Figure 2 shows the accuracy of individual subjects and the mean of each condition. One-sample *t*-tests against the 50% chance performance revealed that the mean accuracy rates of the three conditions were all significantly above chance (single-cue: $t(30) = 8.149$, congruent-cues: $t(32) = 9.685$, incongruent-cues: $t(31) = 4.576$, all ps < 0.001, two-tailed), indicating that participants could achieve at least some level of learning.

Figure 2: Mean percentages of correct responses for individual subjects (empty dots) and for the three experimental conditions (solid dots). The error bars indicate 95% confidence intervals.



To examine segmentation performance across conditions, the listeners' responses in the test were modeled by using linear mixed-effects logistic regression analysis as implemented by the glmer() function available in the lme4 package ([1]) of R ([18]). The dependent variable was subjects' responses (coded 1 if correct and 0 if incorrect). The fixed effect of primary interest was *Condition* (baseline level: single-cue). We also included the

difference in the average TP of adjacent syllables between the word and partword in each trial (*TPdiff*) as a covariate to capture effects of such difference. In addition, *Trial* was included to partial out potential fatigue or practice effects over trials. The random effects contained a by-subject random intercept and a by-subject random slope for *TPdiff*.

The results of the mixed-effects analysis are summarized in Table 1. The model revealed a significant effect of *TPdiff*. As the TP value of an AL word was always higher than that of a partword, the effect suggested that, not surprisingly, responses were more likely to be correct when the TP difference between the word and partword was larger. *Trial* was also significant, indicating that identification became less accurate over trials. Crucially, it was found that while identification accuracy was not significantly different from that in the single-cue condition when tonal coarticulation was congruent with TP information, it was significantly lower when the two cues were incongruent with each other. Such a finding suggests that pitting the coarticulation cue against TPs reduced the listeners' segmentation performance during the learning, providing partial but clear evidence that tonal coarticulation is exploited to discover units in connected speech.

Table 1: Results of the mixed-effects model fitted to the listeners' responses in the test.

Fixed effects:

	β	SE(β)	z	p
Intercept	0.250	0.132	1.893	0.058
TPdiff	1.394	0.242	5.769	<0.001
Trial	-0.012	0.003	-3.491	<0.001
Condition (congruent)	0.088	0.101	0.871	0.384
Condition (incongruent)	-0.251	0.100	-2.524	0.012

4. DISCUSSION AND CONCLUSION

This study investigates the use of tonal coarticulation in Mandarin listeners' segmentation by testing whether the carryover assimilatory effect of one tone on the following tone guided them to perceive units in continuous flows of speech. The results of the AL learning experiment revealed that their segmentation was disrupted in the incongruent-cues condition, in which the partwords received the tonal coarticulatory cue. Such a finding shows that the cue is exploited, even though in this case it incompatible with TP regularities in the AL speech streams.

However, Mandarin listeners' segmentation did not significantly improve in the congruent-cues condition, in which tonal coarticulation and TP information agreed with each other. This is in contrast with the results of [7]'s AL learning experiment, on which ours was based. They showed that congruence between segmental coarticulation and TPs significantly enhanced segmentation performance (at least in an optimal listening situation). Nevertheless, the lack of significant improvement in our congruent-cues condition may not be a too surprising finding. Previous AL learning studies (e.g., [15]) have showed that while listeners' segmentation is inhibited by cues in conflict with regular prosodic patterns in their native language, it is not necessarily facilitated by cues conforming to these patterns. In fact, the different results of the congruent-cues conditions of [7] and the current study may be attributed to two factors. One is that prosodic information is simply a less effective cue compared with segmental information, as has been suggested in models of cue weighting in speech segmentation (e.g., [14]). The other is the mere methodological fact that we gave the tonal coarticulatory cue to only 27% of the word tokens in the congruent-cues condition (so that it had the same number of tonally coarticulated syllables as the incongruent-cues one). It is thus interesting to see whether there will be a significant performance gain if the number of cue-bearing tokens is increased.

Still, the Mandarin listeners' reduced performance in the incongruent-cues condition indicates that their segmentation is somewhat guided by the tonal coarticulatory cue. One question that may be worth exploring further concerns the extent to which such a cue is cross-linguistically useful for tone-language listeners. Tonal coarticulation is implemented as carryover assimilation since it is a coarticulatory pattern widely attested in lexical tone languages. Yet, there are exceptions. For example, it has been reported that in Malaysian Hokkien, carryover tonal coarticulation is not assimilatory and it is even slightly dissimilatory, possibly due to final prominence in the language's tone sandhi system ([5]). Moreover, recent AL learning studies (e.g., [16]) show that some putatively cross-linguistic segmentation cues such as final lengthening may be overridden by language-specific phonological patterning. Including listeners of various lexical tone languages in future research could help understand how language-specific phonology shapes realizations of phonetic details and how these details in turn modulate listeners' segmentation strategies.

5. ACKNOWLEDGEMENT

The present study is partially supported by a research grant from the Ministry of Science and Technology of Taiwan to the second author (MOST 105-2410-H-110-061-MY2).

6. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48.

[2] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer (Version 6.0.37) [Computer program]. Retrieved from http://www.praat.org/.

[3] Brunelle, M. 2009. Northern and Southern Vietnamese tone coarticulation: A comparative case study. *Journal of Southeast Asian Linguistics* 1, 49–62.

[4] Caldwell-Harris, C. L., Lancaster, A., Ladd, D. R., Dediu, D., Christiansen, M. H. 2015. Factors influencing sensitivity to lexical tone in an artificial language: Implications for second language learning. *Studies in Second Language Acquisition* 37, 335–357.

[5] Chang, Y.-C., Hsieh, F.-F. 2012. Tonal coarticulation in Malaysian Hokkien: A typological anomaly? *The Linguistic Review* 29, 37–73.

[6] Cho, T. 2004. Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal. of Phonetics* 32, 141–176.

[7] Fernandes, T., Ventura, P., Kolinsky, R. 2007. Statistical information and coarticulation as cues to word boundaries: A matter of signal quality. *Perception & Psychophysics* 69, 856–864.

[8] Fougeron, C., Keating, P. A. 1997. Articulatory strengthening at edges of prosodic domains. *J. Acoust. Soc. Am*. 101, 3728–3740.

[9] Gandour, J., Potisuk, S., Dechongkit, S. 1994. Tonal coarticulation in Thai. *Journal of Phonetics* 22, 474–492.

[10] Gómez, D. M., Mok, P., Ordin, M., Mehler, J., Nespor, M. 2018. Statistical speech segmentation in tone languages: The role of lexical tones. *Language and Speech* 61, 84–96.

[11] Kim, S., Broersma, M., Cho, T. 2012. The use of prosodic cues in learning new words in an unfamiliar language. *Studies in Second Language Acquisition* 34, 415–444.

[12] Lai, W., Kuang, J. 2016. Prosodic grouping in Chinese trisyllabic structures by multiple cues–tone coarticulation, tone sandhi and consonant lenition. *Proc. of Tonal Aspects of Languages 2016*, 157–161.

[13] Li, Y, Lee, T, Qian, Y. 2004. Analysis and modeling of F0 contours for Cantonese text-to-speech. *ACM Transactions on Asian Language Information Processing* 3, 169–180.

[14] Mattys, S. L., White, L., Melhorn, J. F. 2005. Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General* 134, 477–500.

[15] Ordin, M., Nespor, M. 2016. Native language influence in the segmentation of a novel language. *Language Learning and Development* 12, 461–481.

[16] Ordin, M., Polyanskaya, L., Laka, I., Nespor, M. 2017. Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Memory & Cognition* 45, 863–876.

[17] Psychology Software Tools. 2012. E-Prime 2.0. Pittsburgh, PA: Author.

[18] R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[19] Shih, C.-L. 1988. Tone and intonation in Mandarin. *Working Papers of the Cornell Phonetics Laboratory* 3, 83–109.

[20] Toro, J. M., Pons, F., Bion, R. A., Sebastián-Gallés, N. 2011. The contribution of language-specific knowledge in the selection of statistically-coherent word candidates. *Journal of Memory and Language* 64, 171–180.

[21] Tremblay, A., Cho, T., Kim, S., Shin, S. 2018. *Proc. of Speech Prosody 2018* Poznań, 65–69.

[22] Tyler, M. D., Cutler, A. 2009. Cross-language differences in cue use for speech segmentation. *J. Acoust. Soc. Am*. 126, 367–376.

[23] Vroomen, J., Tuomainen, J., de Gelder, B. 1998. The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language* 38, 133–149.

[24] Xu, Y. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25, 61–83.

[25] Zhang, J., Liu, J. 2011. Tone sandhi and tonal coarticulation in Tianjin Chinese. *Phonetica* 68, 161–191.