THE EFFECT OF MANDARIN ACCIDENTAL GAPS ON PERCEPTUAL CATEGORIZATION

Tzu-Hsuan Yang, Shao-Jie Jin, Yu-An Lu

National Chiao Tung University
thy2105@tc.columbia.edu; courtney0419003@gmail.com; yuanlu@nctu.edu.tw

ABSTRACT

Perceptual categorization has been shown to be biased by various linguistic knowledge, including native inventory, lexicon, phonotactic restrictions, and morphological alternation. Building on previous studies on segmental categorization, the present study investigates whether tonal categorization may be biased by knowledge of accidental gaps (i.e., syllable-tone combinations that could but do not exist) in Mandarin. In a forced-choice identification experiment, Mandarin listeners were presented with CV syllables carrying a tone along one of two continua, T1-T2 (level to rising) or T1-T4 (level to falling), with a word on one end and an accidental gap on the other, or with both ends being words or gaps (e.g., *T1-T2, T1-*T2, T1-T2, *T1-*T2). The results showed that Mandarin listeners' responses, though heavily guided by syllable frequencies, were biased towards the non-gap endpoints on the continua. The results suggest that listeners' perceptual categorization is sensitive to segmental as well as suprasegmental (tonal) information.

Keywords: accidental gaps, categorical perception, speech perception, tone, Mandarin

1. INTRODUCTION

Previous studies have demonstrated that speech perception may be shaped by a speaker's linguistic knowledge. For example, in a study examining the influence of the lexical status of a phonetic sequence on phonetic categorization, Ganong [6] employed a continuum of stops varying in VOT in which one end was a word and the other was a non-word (e.g., *task*-**dask, *tash*-*dash*), and asked participants to identify the word they heard. The results showed that English speakers were more likely to identify ambiguous stimuli as the real words (*task* or *dash*) along the continuum.

Phonotactic restrictions have also been shown to affect phonetic categorization. Massaro and Cohen [12] demonstrated this effect through an identification experiment in which segments [l] and [r] were synthesized along a continuum and embedded after [s] and [t], creating [sl]-*[sr] and *[tl]-[tr] combinations. English speakers consistently made more [l] responses when ambiguous sounds were embedded after [s], but more [r] responses when the identical sounds were placed after [t], suggesting an influence of phonotactic knowledge on speech perception.

In a similar study testing speech perception in Korean speakers, Ahn [1] synthesized a continuum from [ti] to [tɕi] and embedded the ambiguous sounds within homo-morphemic (e.g., [tipa]-[tɕipa]) and hetero-morphemic (e.g., [nit-ida]-[nitɕ-ida]) contexts. In Korean, [t] and [tɕ] are phonemic within morpheme boundaries while [t] becomes [tɕ] preceding [i] across morpheme boundaries. The results showed that the boundary of categorization was biased towards [tɕ] in hetero-morphemic contexts, providing evidence for the effect of morphological restrictions on speech perception.

While previous studies have demonstrated a sensitivity to segmental features, there is little evidence of speech sound categorization being shaped by suprasegmental information (c.f., [5]). The present study aims to investigate the role of suprasegmental information in speech perception by examining the effect of accidental gaps on Mandarin speakers' tonal categorization.

Mandarin is a tone language with four phonemic tones: high-level Tone 1 [X^{55}], rising Tone 2 [X^{35}], falling-rising Tone 3 [X^{214}], and falling Tone 4 [X^{51}]. However, not every allowable syllable carries all four tones. For instance, the syllable [ts^{h}u] can be combined with T1 ([ts^{h}u]^{55} "coarse"), T2 ([ts^{h}u]^{35} "die" in Old Chinese), and T4 ([ts^{h}u]^{51} "vinegar"), but not with T3 (*[ts^{h}u]^{214}). These syllable-tone combinations that could but do not exist are termed 'accidental gaps' [4]. Building on previous works, the present study investigates whether tonal categorization may be biased by the knowledge of accidental gaps.

2. EXPERIMENT

To examine Mandarin speakers' tonal categorization, we conducted a forced-choice identification experiment.

2.1. Methodology

2.1.1. Participants

20 Taiwanese Mandarin speakers (16F, 4M; ages 20-37, *M*= 22) were recruited from National Chiao Tung University. None reported any hearing deficiencies. All participants were compensated monetarily for their time.

2.1.2. Materials

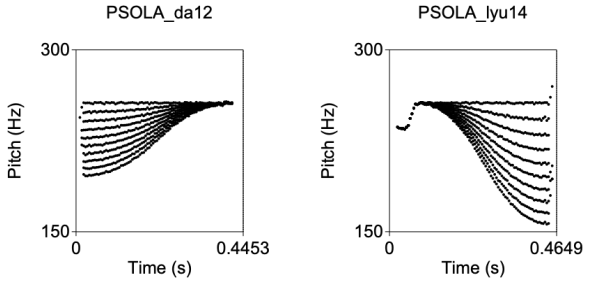
To examine listeners' tonal categorization, Mandarin CV syllables carrying a tone along one of two continua were selected: T1-T2 (level to rising) or T1-T4 (level to falling). Eight pairs were selected so that each pair contained a word on one end and an accidental gap on the other, or with both ends being words or gaps (e.g., *T1-T2, T1-*T2, T1-T2, *T1-*T2). Since only a limited number of tokens fit this criteria, token frequency (Table 1 [15]) was not balanced. The possible effect of the unbalanced token frequency is discussed in Section 3.

Table 1: Syllable frequencies of the selected tokens

T1-T2 continua		T1-T4 continua	
Tokens	Frequency	Tokens	Frequency
*[tɤ]^{55}-[tɤ]^{35}	gap-844	*[ly]^{55}-[ly]^{51}	gap-3,129
[tʰa]^{55}-*[tʰa]^{35}	11,088-gap	[ha]^{55}-*[ha]^{51}	27-gap
[ta]^{55}-[ta]^{35}	146-157	[ta]^{55}-[ta]^{51}	146-3,375
[ny]^{55}-[ny]^{35}	gap-gap	*[tɤ]^{55}-*[tɤ]^{51}	gap-gap

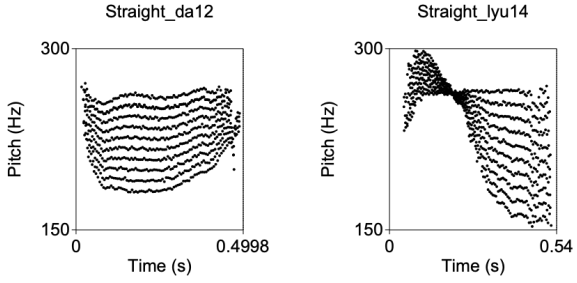
The eight syllables were naturally produced and recorded by a phonetically trained female native Taiwan Mandarin speaker. Two sets of continua were then synthesized. Following the common practice of previous studies involving pitch manipulation, we used the Pitch Synchronous Overlap and Add (PSOLA) algorithm in Praat [3] to create the first set of stimuli. The endpoint stimuli were first manipulated to have the same duration (i.e., the mean duration of the endpoints) to avoid temporal cues that would favor one end over the other (T1-T2: 484 ms; T1-T4: 465 ms). For each pair, the *f0* of T1 endpoints was set as 256.2 Hz, the mean *f0* of all the T1 tokens. Then a 10-step continuum was resynthesized using equal steps along the semitone scale, with 196.2 Hz at the tone onset to 256.2 Hz at the tone offset for the rising contour, and 256.2 Hz at the tone onset to 156.2 Hz at tone offset for the falling contour. Onset and offset values were selected based on the means of the tokens produced. Figure 1 shows the *f0* trajectories of the resynthesized stimuli for the token [ta] on the T1-T2 continuum (left) and for the token [ly] the T1-T4 continuum (right).

Figure 1: *f0* trajectories of tokens synthesized using PSOLA.



This method created continua varying solely in pitch. However, it has been shown that secondary cues such as duration and creakiness can also affect listeners' perception and categorization of tones [17, 18]. Therefore, we generated another set of stimuli using TANDEM-STRAIGHT, a speech analysis, modification and resynthesis framework [8]. This method allows for the resynthesis to be done from the entire CV token; in other words, not only was the pitch manipulated, but the duration and voice quality were also manipulated proportionally. By including this set of stimuli, we hoped to tap into listeners' phonological knowledge of tonal perception, instead of regarding only a single acoustic cue. Figure 2 shows the *f0* trajectories of the resynthesized stimuli for the token [ta] on the T1-T2 continuum (left) and for the token [ly] on the T1-T4 continuum (right).

Figure 2: *f0* trajectories of tokens synthesized using TANDEM-STRAIGHT.



2.1.3. Procedure

The 160 resynthesized stimuli (2 tone continua [T1-T2, T1-T4] x 4 endpoints [*T-T, T-*T, T-T, *T-*T] x 10 steps x 2 methods [PSOLA, STRAIGHT]) were presented in four blocks, using E-Prime [14]. Stimuli resynthesized using the two different methods and along different tone continua were put into separate blocks. The four blocks as well as the trials in them were randomized. Participants were verbally instructed and given written instructions on the monitor to listen to each sound and judge whether they heard T1 or T2/T4 by pressing the corresponding key on the keyboard as soon as they were sure. Participants' responses and response times (RTs) were recorded in E-Prime. Six practice trials were

given to familiarize participants with the task. These trials contained only the endpoint stimuli from the continua.

2.2. Results

2.2.1. Results of tonal categorization

Mixed-effects logistic regression models were first fitted to the data on the continua where one end was a word while the other was a gap, using the lme4 package in R [2]. The dependent variable was the participants' tonal categorization, with T1 responses coded as 0 and T2/T4 responses coded as 1. The models included Endpoint (e.g., T1-*T4, *T1-T4) and Step (1 to 10, normalized), and an interaction term for Endpoint and Step. The models also included random intercepts and slopes for Participant and Token. We ran separate models for each of the methods (PSOLA and STRAIGHT) and tone continuum (T1-T2 and T1-T4). The results showed that when presented with stimuli along the T1-T4 continua resynthesized using SRAIGHT, listeners' responses were significantly biased towards the non-gap endpoints as indicated by the effect of different Endpoints (p=.002) (Figure 3). Nevertheless, the same bias was not observed in the PSOLA-generated stimuli, as indicated by the non-significant effect of Endpoint (p=.272) (Figure 4).

Figure 3: Responses to the STRAIGHT-generated T1-T4 continua with different endpoints

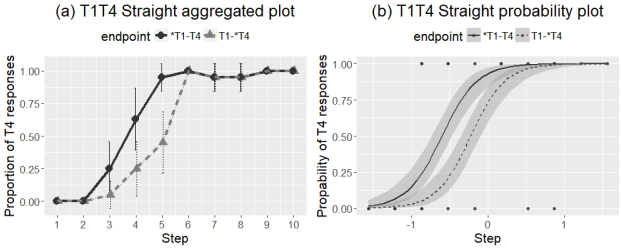
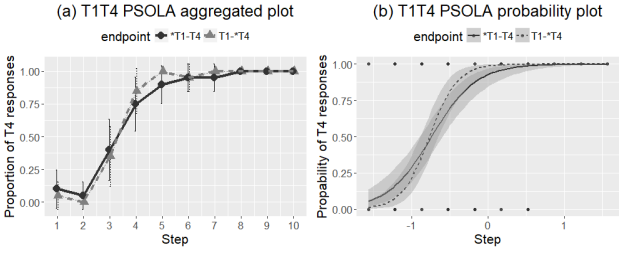


Figure 4: Responses to the PSOLA-generated T1-T4 continua with different endpoint



For the STRAIGHT-generated T1-T2 continua, there was a marginal effect of Endpoint (p=.06), suggesting that the listeners' responses were slightly biased towards the non-gap endpoints (Figure 5). Continua resynthesized using PSOLA, again, did not show this pattern (p= .919) (Figure 6) (but also see the findings in [5]).

Figure 5: Responses to the STRAIGHT-generated T1-T2 continua with different endpoints.

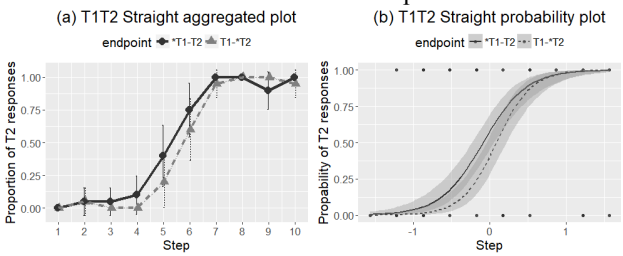
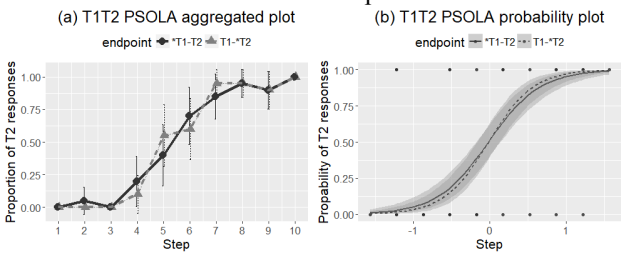


Figure 6: Responses to the PSOLA-generated T1-T2 continua with different endpoints.



Finally, recall that continua with both ends being words or gaps were also included in the stimuli to serve as the baseline. The word-word continuum and gap-gap continuum were not expected to show any bias; however, Figure 7 and 8 show that they actually behaved quite differently. We attribute this to a possible frequency effect which will be discussed in more detail in Section 3.

Figure 7: Responses to continua with both ends being words.

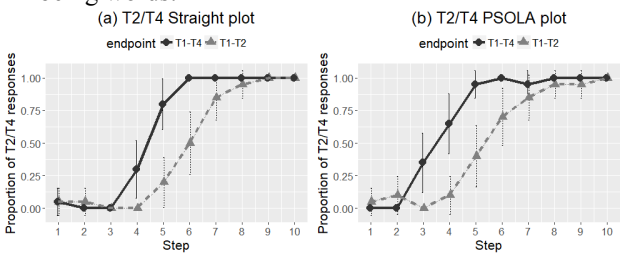
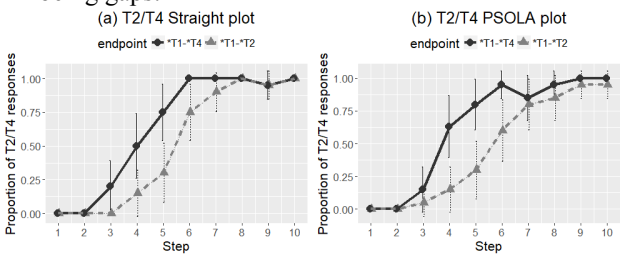


Figure 8: Responses to continua with both ends being gaps.



2.2.2. Results of response time

Figures 9 (T1-T2) and 10 (T1-T4) show the RTs to the stimuli on the tone continua created by the two methods, STRAIGHT (a) and PSOLA (b). We observed two tendencies. First, the RTs peaked at different steps along the continua, with earlier peaks

for the *T-T continua and later for the T-*T continua. In addition, although listeners successfully identified the gap word, they generally spent more time in making the decisions. This tendency was also evident in the PSOLA-generated stimuli, in which no bias was found in the identification results. The RT results were taken to suggest that listeners' perceptual categorization was biased by accidental gaps.

Figure 9: RTs of the identification of T1-T2 continua.

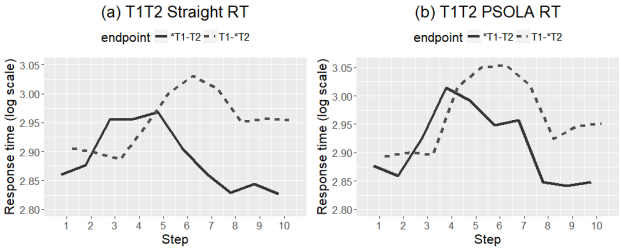
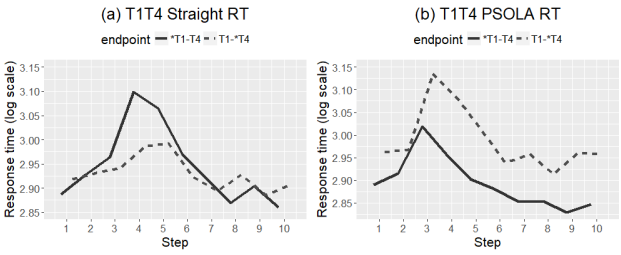


Figure 10: RTs of the identification of T1-T4 continua.



3. DISCUSSION AND CONCLUSION

The results of the identification task showed that Mandarin listeners' tonal categorization was, to some extent, biased by their knowledge of accidental gaps. However, differences between the two tone continua and between the different synthesis methods were evident. The possible sources of these observed differences are discussed below.

First, although our intent was not to compare the two synthesis methods, the results for the two sets of stimuli clearly diverged: the effect of Endpoint was more pronounced when participants were presented with the STRAIGHT-generated stimuli rather than PSOLA-generated ones. We attribute this to the fact that PSOLA only manipulates pitch information, which may have guided participants to attend more to pure phonetic cues. In contrast, the STRAIGHT resynthesis process proportionally manipulated other acoustic features which have been shown to be important secondary cues for tonal perception [7, 10, 11, 17, 18]. Therefore, the STRAIGHT-generated stimuli likely tapped into phonological processing. The results indicated that tonal categorization reflected the listeners' knowledge of accidental gaps, but only when other phonetic cues besides pitch were available.

Another important observation was that the effect of Endpoint was less conspicuous in the T1-T2 continua: only a marginal effect was found for STRAIGHT-generated stimuli, whereas no Endpoint effect was found for PSOLA-generated stimuli. This may be a reflection of T2 ([X^{35}]) having a smaller pitch span than T4 ([X^{51}]), which was taken into consideration when creating the PSOLA stimuli. When presented with a stimulus along the T1-T2 continua, participants may have identified it as T2 as long as it had a slightly rising contour. With a greater span, there was more room for ambiguity along the T1-T4 continua. In addition, T4 offsets may exhibit creaky phonotation [9], and T4 syllables are significantly shorter than the other tones [17], which were available as extra information for tonal processing.

Finally, recall that continua with both ends being words or gaps, which were supposed to serve as baselines, actually showed some bias. In the word-word condition (Figure 7), tokens on the T1-T4 continua had relatively earlier perceptual boundaries than those on the T1-T2 continua. We attribute this to a possible frequency effect. For the T1-T4 continuum where both ends were words, the token frequency was unbalanced (146 vs. 3375). In contrast, token frequencies were relatively balanced for the T1-T2 continuum (146 vs. 157). That is, there was a bias toward T4, a higher-frequency word compared to T1 on the continuum. The earlier boundary for the tokens on the T1-T4 continua lends further support to previous findings that speakers' word judgement may be biased by word frequency [13, 16].

We would, then, expect no bias in the gap-gap condition in that syllables that are gaps contain no frequency. However, the results were in fact biased towards T4 (Figure 8). We attributed this to a tone frequency effect, as T4 generally occurs more frequently than T1 (228182 vs. 105168) [15].

To summarize, the findings in the present study showed that Mandarin listeners' responses, though heavily guided by syllable and tone frequencies, were biased by accidental gaps, but only when all acoustic cues were considered. The results suggest that listeners' perceptual categorization is sensitive to segmental as well as suprasegmental (tonal) information.

4. ACKNOWLEDGEMENT

We would like to thank Sang-Im Lee-Kim, Ho-Hsien Pan, Peggy Mok and the ICPhS reviewers. This work was supported by Ministry of Science and Technology (MOST106-2410-H-009-031) grant to Yu-An Lu.

5. REFERENCES

[1] Ahn, M. 2008. Morphologically conditioned perceptual bias. *Proc. Chicago Linguistic Society* Chicogo, 44:1-15.

[2] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* 67(1), 1–48.

[3] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot Int.* 5:9/10, 341–345.

[4] Duanmu, S. 2000: The phonology of Standard Chinese. Oxford: Oxford University Press.

[5] Fox, R. A. Unkefer, J. 1985. The effect of lexical status on the perception of tone. *J. Chinese Linguistics*, 13(1), 69-90.

[6] Ganong, W. F. 1980. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology Human Perception and Performance* 6(1), 110–125.

[7] Huang, Y. Low f0 as a creak attribute in mandarin tone perception. The 93rd Annual Meeting of the Linguistic Society of America, 4 Jan 2019, Sheraton New York Times Square, NY. Conference Presentation.

[8] Kawahara, H., Morise, M., Takahashi, T., et al. 2008. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *Acoustics, Speech and Signal Processin*g Las Vegas, NV, 3933-3936.

[9] Kuang, J. 2013. Phonation in tonal contrasts. Doctoral Dissertation. University of California, Los Angeles.

[10] Kuang, J., Liberman, M. 2018. Integrating voice quality cues in the pitch perception of speech and non-speech utterances. *Front. Psychol.* 9, 2147.

[11] Lu, Y., Lee-Kim, S. The effect of linguistic experience on perceived vowel duration: evidence from tone language speakers. The 93rd Annual Meeting of the Linguistic Society of America, 5 Jan 2019, Sheraton New York Times Square, NY. Conference Presentation.

[12] Massaro, D. W., Cohen, M. M. 1983. Phonological context in speech perception. *Perception and Psychophysics* 34(4), 338–348.

[13] Myers, J., Tsay, J. 2013. Modeling universal and lexical influences on phonotactic judgments. *Proc. 4th International Theoretical Phonology Conference* Taipei.

[14] Schneider, W., Eschman, A., & Zuccolotto, A. 2002. *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.

[15] Tseng, S.-C. 2013. Lexical coverage in Taiwan Mandarin conversation. *International Journal of Computational Linguistics and Chinese Language Processing* 18(1):1-18.

[16] Wiener, S., Ito, K. 2015. Do syllable–specific tonal probabilities guide lexical access? Evidence from Mandarin, Shanghai and Cantonese speakers. *Language, Cognition, and Neuroscience*, 30(9), 1048–1060.

[17] Wu, F., & Kenstowicz, M. 2015. Duration reflexes of syllable structure in Mandarin. *Lingua,* 164, 87–99.

[18] Yu, K. M. 2010. Laryngealization and features for Chinese tonal recognition. *Proc. INTERSPEECH-2010* Makuhari, 1529-1532.