Cognitive factors in perception of Thai tones by naïve Mandarin listeners

Juqiang Chen[1], Catherine T. Best[1,2], Mark Antoniou[1], Benjawan Kasisopa[1]

[1]The MARCS Institute, Western Sydney University, Australia
[2]Haskins Laboratories, New Haven CT, USA
J.Chen2/C.Best/M.Antoniou/B.Kasisopa@westernsydney.edu.au

ABSTRACT

Memory load and task-irrelevant phonetic variations influence discrimination of non-native segmental contrasts. We tested how these factors modulate perceptual assimilation and/or discrimination of non-native lexical tone contrasts, relative to Perceptual Assimilation Model (PAM) [1-2] predictions. When perceptually assimilating Thai tones to their native tone system, Mandarin listeners showed sensitivity to native allophonic differences only if memory load was low, but were unaffected by phonetic variations in talkers and vowels. However, AX discrimination decreased with either talker or vowel variability. Unlike non-native segment perception, where discrimination is poorer under high memory load than lower load, tone discrimination was not diminished by high load (long interstimulus interval). PAM-driven predictions were supported across the cognitive manipulations: when two Thai tones were categorized into two native categories (TC) they were better discriminated than when one or both Thai tones were uncategorized (UC/UU). Overlapping choices for TC assimilations can reduce discrimination accuracy.

Keywords: cross-language perception, cognitive load, lexical tones, Thai tones, Mandarin listeners

1. INTRODUCTION

Perceptual attunement to native speech constrains perception of non-native contrasts. This influence from one's native language is theorized by the Perceptual Assimilation Model (PAM) [1] to occur by way of perceptual assimilation. A given non-native phone may be perceptually assimilated to the native phonological system in one of three ways: (1) as Categorized to a native phoneme; (2) as an Uncategorized phone that falls between native phonemes; or (3) as a Non-Assimilable (NA) non-speech sound. Consequently, PAM claims that discrimination is better if two non-native phones are assimilated into two native categories (Two Category assimilation: TC) than if they are both assimilated into the same native category but differ in their discrepancy from the native "ideal" (Category Goodness difference: CG), which is in turn better than if they are equally good/poor exemplars of one category (Single Category: SC).

Naïve listeners can categorize non-native consonants and vowels [3]–[5] as well as lexical tones [7]–[9] into their native categories, and their perceptual assimilation patterns can predict their performance in discriminating non-native contrasts. However, those studies did not systematically manipulate cognitive factors. A number of studies have identified a distinction between a phonetic mode and a phonemic/phonological mode in speech perception [6]–[8], as a function of cognitive load.

Memory load, the capacity to hold a rapidly decaying memory for limited time [9], can cause a switch between modes. In discrimination tasks it is operationalized by manipulating inter-stimulus intervals (ISI). With long ISIs (1500 ms; high load), English listeners discriminated the Hindi retroflex-dental stop contrast according to L1 phonology which has only an alveolar stop, whereas with short ISIs (500 ms; low memory load) they discriminated phonetic level differences. Similarly, German listeners' discrimination of Japanese segmental length contrasts was negatively affected by high memory load (ISI = 2500ms) [10].

Another cognitive factor explored in previous discrimination studies is attentional control, the ability to allocate attention between task-relevant and irrelevant information [11]. In [10], German listeners discrimination of Japanese consonant length was adversely affected when task-irrelevant information (pitch variation) was added to the task.

Talker variability also leads to attention shifts and increased cognitive load. One theory is that talker variability leads listeners to form multiple phonetic interpretations for a particular acoustic pattern, holding the alternatives in working memory while shifting attention to evaluate them [12]. This suggests that accommodating talker variability demands more working memory resources.

Studies on cognitive factors in cross-language perception have mostly been restricted to consonants and vowels, whereas few cross-language tone per-

ception studies have investigated cognitive factors in discrimination. Those that have did not use theory-driven predictions about L1 influences. In addition, memory load and attention control were most often manipulated in discrimination tasks, which generally involve more low-level phonetic processing than categorization tasks. Thus it is unknown whether these cognitive factors affect cross-language perceptual assimilation similarly.

Our study manipulated memory load and attention control in non-native tone categorization and discrimination tasks. Memory load was operationalized as ISI in the discrimination task and as response interval (time between the end of the stimulus and the signal to select an L1 category) in the categorization task. Attention control was operationalized by manipulating talker and vowel variability.

Discrimination was tested first in each of two sessions to minimize effects of prior categorization on performance. Due to the multiple cognitive load conditions, it was not feasible to test all Thai tone pairs. Based on a previous study using the same stimuli [13], we selected three Thai tone contrasts that met the required PAM assimilation patterns: T33-T45 (TC), T315-T45 (SC), T33-T241 (UC). After the first session, participants were asked to come back two weeks later for a second session in which we tested another two pairs: T21-T241 (UC), T21-T33 (TC), to compare with a Vietnamese group in a larger project.

Based on PAM, we predicted that Mandarin listeners would discriminate T33-T45 (TC) and T21-T33 (TC) better than T33-T241 (UC) and T21-T241 (UC). T315-T45 (SC) should be the most difficult contrast to distinguish. In order to evaluate these predictions, we report the categorization experiment before the discrimination.

2. EXPERIMENT 1: CATEGORIZATION

2.1. Method

2.1.1. Participants

28 native speakers of Mandarin participated in both experiments 1 and 2, divided into two groups for each response interval/ISI condition (ISI_{short}: M_{age} = 24 years, SD = 4; 8 females; ISI_{long}: M_{age} = 25 years, SD = 6; 10 females). Participants completed a background questionnaire before the test. All had normal hearing and none had experience with Thai or more than two years of formal musical training.

2.1.2 Stimulus materials

Two syllables (/ma/, /mi/) were chosen because they are real words for each native tone in both Thai and Mandarin. The target Thai syllables were each read several times by two female native Thai speakers. These informants had no experience with other tone languages. Two tokens of each target item that were judged to be correct and most natural-sounding to a third native Thai speaker were selected.

We used Chao values [14] to provide a priori phonetic descriptions of the tones in each language. In Chao notation, F0 height at tone onset and offset is referenced by numbers 1-5 ranging from low to high. Thai, the target language, has three level tones (characterized as high-level T45, mid-level T33, low-level T21) and two contour tones (rising T315 and falling T241) [15]. Mandarin has four tones: a level tone M55; a rising tone M35; a falling-rising tone M214; and a falling tone M51 [16].

Response interval condition (2000ms vs. 500ms) was a between-subjects factor, while talker variability (same vs. different) and vowel variability (same vs. different vowels in each trial: /ma/, /mi/) were blocked within each group.

2.1.3 Procedure

Participants were tested individually in the testing booth (at Western Sydney University, or UNSW). Stimuli were presented on a Dell Latitude 7280 laptop running E-Prime Professional 2. Stimuli were presented via Sennheiser HD 280 Pro headphones at 72 dB SPL.

Before the test session, participants completed 10 practice trials. The categorization task had 140 trials in total. On each trial, the stimulus token was presented and listeners made a forced-choice categorization judgment to their native tones (four Pinyin options) via a key press as quickly as possible within a 3s timeout. They then heard the tone again and rated how well it fitted their chosen native category on a 7-point scale. (1 = poor, 7 = perfect, 4 = OK).

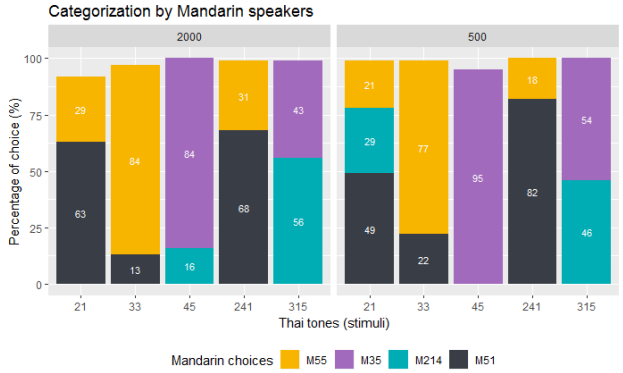
2.2. Results

3897 data points were collected (23 missing points removed). We fitted the data with a multinomial regression model. The full model was built with Mandarin tone choice as a dependent measure, and response interval, talker variability, vowel variability and Thai tones as fixed factors. Fixed factors were subtracted one at a time and compared to the full model to determine the effect of cognitive load. Both response interval ($\chi^2(12) = 92.37$, $p < .0001$) and Thai tone ($\chi^2(12) = 5630$, $p < .0001$) showed significant effects in Likelihood ratio tests. However, talker variability and vowel variability were not significant. Mean percent selection of each Mandarin L1 tone for each Thai tone are plotted in Figure 1.

To determine whether a Thai tone was catego-

rized into a native Mandarin tone category, we set the categorization criterion to 70% of responses [17]. Both T33 and T45 were Categorized in all cognitive conditions; T21 and T315 were Uncategorized, differing from [13]. T241 was Categorized in the short response interval condition but Uncategorized in the long response interval condition.

Figure 1: Categorization of Thai tones by Mandarin listeners in the two response interval conditions.



3. EXPERIMENT 2: DISCRIMINATION

3.1 Method

3.1.1 Participants and Stimulus Materials were the same as in Experiment 1.

3.1.2 Procedure

An AX task ("same-different") was used because it allows better control of ISI; it was used in many previous discrimination studies [10]. ISI (500ms vs. 2000ms) was a between-subjects factor, as we reasoned that listeners may not be able to switch between phonetic and phonological mode within an experiment. Eight blocks (talker × vowel variability) were randomized for each participant.

3.1.2 Data analysis

In order to minimize decision bias, we calculated *d'* for discrimination performance. For each tone pair in each cognitive condition, *d'* scores were calculated using the formula *d'* = Z (hit rate) - Z (false positive rate) with adjustments made for probabilities of 0 (=.01) and 1 (=.99). Hit is defined as the number of correct responses ("different" responses on AB or BA trials). False positive is defined as the number of incorrect responses ("different" responses on AA or BB trials).

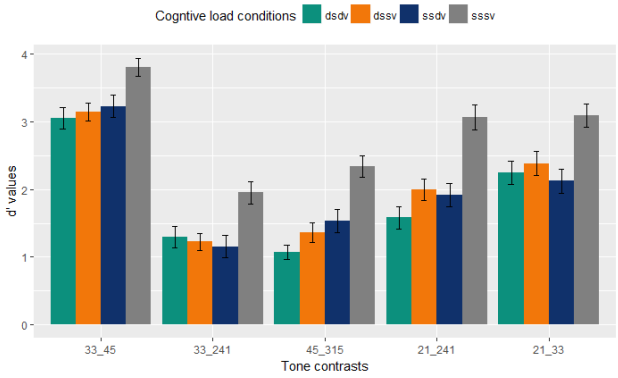
3.2 Results

17847 raw data points were collected (with 73 missing points removed). The *d'* scores for each participant were calculated separately for each tone pair, and for each block, yielding 40 data points for each participant. We fitted the data using a Linear Mixed Effect Regression (LMER) model with *d'* as the dependent variable, and ISI, talker variability, vowel variability and tone pairs as fixed factors, and subject as the random intercept.

To calculate the *p*-values for the fixed effects, we used the Kenward-Roger approximation to the degrees of freedom, as recommend by [18], and the *Anova* function from the *car* package in R, with test specified as "F". The main effect of ISI and all interactions involving ISI were non-significant. However, there were significant main effects of talker variability, $F(1, 1054) = 55.99$, $p < .001$, vowel variability, $F(1, 1054) = 63.80$, $p < .001$ and tone pair, $F(1, 1054) = 109.89$, $p < .001$.

In addition there was a significant interaction between speaker and vowel variability, $F(1, 1054) = 27.67$, $p < .001$. As ISI did not affect the perception, we plotted mean and standard error bars only in terms of vowel and talker variability in Figure 2.

Figure 2: Discrimination of Thai tone contrasts by Mandarin listeners in different cognitive load conditions[a].



[a]Notes: dsdv stands for different speaker different vowel block; dssv stands for different speakers same vowel block; ssdv stands for same speaker different vowels; sssv stands for same speaker same vowel.

Moreover, we did multiple comparisons to test effects of different cognitive factors with the R-package *lsmeans*. Significant differences were found between the same-talker-same-vowel block and the different-talker-different-vowel block, the cognitively easiest versus the most difficult blocks, respectively, β = -1.00, SE = .091, $t(1054)$ = -10.93, p <.001. In addition, when talkers were the same, there was a significant difference between same vowel and different vowels, β = -.86, SE = .091, $t(1054)$ = -9.37 , $p < .001$. Similarly, when the vowel was the same, talker variability had a significant difference, β = -.82, SE = .91, $t(1054)$ = - 9.01 , $p < .001$. Other combinations of talker and vowel conditions were not significantly different.

To test PAM-driven predictions of tone pair contrasts, we did multiple comparisons (Table 1).

Table 1: Multiple comparisons of tone contrasts with PAM-driven predictions

Contrast types	Tone contrasts		t	p
TC_TC/UC	33_45	33_241	18.52	**
TC_UC	33_45	45_315	16.87	**
TC_UC/UU	33_45	21_241	11.37	**
TC_UC	33_45	21_33	8.23	**
TC/UC_UC	33_241	45_315	-1.65	.467
TC/UC_UC/UU	33_241	21_241	-7.15	**
TC/UC_UC	33_241	21_33	10.29	**
UC_UC/UU	45_315	21_241	-5.50	**
UC_UC	45_315	21_33	-8.64	**
UC/UU_UC	21_241	21_33	-3.14	*

Note: * indicates < .05; ** < .001

4. GENERAL DISCUSSION

Generally, we found that different cognitive factors affect discrimination and categorization of non-native tones. First, the response interval effects reflect the influence of memory load in categorization task. The initial information listeners get from speech is low-low level phonetic information. For example, in low memory load, they chose M214 as a response for T21, based on the phonetic similarity between T21 and the allotone of M214 (M21) [13]. But when they were required to wait before selecting their choice answer, the phonetic details faded and all they retained was the more abstract, categorical phonological information. Thus they chose phonologically more similar falling tone M51 for T21, ignoring their phonetic differences. Additionally, phonological processing takes more time than phonetic processing; thus the response interval effect could also reflect two different levels of processing.

Neither vowel nor talker variability affected categorization. Vowel and talker variability in categorization task existed in blocks, not in each trial. Thus, when listeners focused on one tone at a time in each block, the distraction of the talker and vowel variation may have been less than in discrimination. Moreover, the categorization task is more phonological in nature, especially in long response time condition where phonetic details decay and listeners have more time for high-level processing and thus in this case their choices were less susceptible to low level phonetic variations. This supports the argument of phonological constancy that perceivers can assimilate indexical properties of unfamiliar talkers into the key indexical features of their native speech community [19]. Thus they were immune to talker and vowel variation within each block.

Conversely, both vowel and talker variability affected discrimination. AX task involves more phonetic than phonological processing (using more bottom-up stimulus information) and thus is more susceptible to phonetic variation than categorization task. However, ISI did not lead to different performances in tone discrimination, unlike consonant discrimination, where in long ISI listeners fail to distinguish differences that they can do in short ISI. The reason may be that the acoustic cues for consonants are short in duration, and thus are more likely to decay in short-term memory. Tones are longer in duration (in this study extending over the whole syllable). Thus they are less susceptible to decay in short-term memory.

Most PAM-motivated predictions work in different cognitive conditions. TC contrast is better discriminated than the UC and UU contrasts as predicted by PAM. Within UC contrasts, 21_33 (UC) was better discriminated than 45_315 (UC). This could be because 45_315 has a stronger overlapping response choices (M35), leading to more confusion. However, 33_241 (TC/UC) was perceived significantly worse than 21_241(UC/UU) and 21_33 (UC). This could be because both T33 and 241 were assimilated to complementarily overlapping native choices (both yielded M1 and M4 choices), similar to what has been found in unassimilated vowel pairs [20].

5. CONCLUSION

Cross-language tone categorization and discrimination were each affected by different cognitive factors. The longer response interval may have led to decay of low-level phonetic information in the categorization task, shifting listeners to rely more on phonological similarity between native and non-native tones. Categorization was not affected by talker and vowel variability, however, indicating that listeners were to maintain phonological constancy. Showing the opposite pattern, tone discrimination was robust across long and short ISIs, unlike prior findings on perception of non-native segments, but it was affected by the low-level task-irrelevant phonetic variations of talker and vowel variability. PAM-driven predictions (TC>UC>UU) were largely upheld under the different cognitive load conditions. Another novel finding was that overlapping native category choices even for TC assimilation types can decrease discrimination performance on the affected non-native tone pairs. This study has implications for theories of speech perception in general and in particular for other models of non-native and second language speech perception such as the Speech Learning Model (SLM) [21] and the Second Language Speech Perception model (L2LP) [22].

6. REFERENCES

[1] C. T. Best, "A direct realist view of cross-language speech perception.," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.

[2] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language Learning & Language Teaching*, vol. 17, O.-S. Bohn and M. J. Munro, Eds. Amsterdam: John Benjamins Publishing Company, 2007, pp. 13–34.

[3] C. T. Best and W. Strange, "Effects of Phonological and Phonetic Factors on Cross-Language Perception of Approximants," *J. Phon.*, vol. 20, no. 3, pp. 305–330, Jul. 1992.

[4] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *J. Acoust. Soc. Am.*, vol. 109, no. 2, pp. 775–794, Feb. 2001.

[5] M. D. Tyler, C. T. Best, A. Faber, and A. G. Levitt, "Perceptual assimilation and discrimination of non-native vowel contrasts," *Phonetica*, vol. 71, no. 1, pp. 4–21, 2014.

[6] J. F. Werker and R. C. Tees, "Phonemic and phonetic factors in adult cross-language speech perception," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1866–1878, Jun. 1984.

[7] D. B. Pisoni, "Auditory and phonetic memory codes in the discrimination of consonants and vowels," *Percept. Psychophys.*, vol. 13, no. 2, pp. 253–260, Jun. 1973.

[8] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Percept. Psychophys.*, vol. 37, no. 1, pp. 35–44, Jan. 1985.

[9] A. Baddeley and B. A. Wilson, "Prose recall and amnesia: implications for the structure of working memory," *Neuropsychologia*, vol. 40, no. 10, pp. 1737–1743, 2002.

[10] Y. Asano, "Discriminating Non-Native Segmental Length Contrasts Under Increased Task Demands," *Lang. Speech*, pp. 1–21, Oct. 2017.

[11] T. Isaacs and P. Trofimovich, "Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech," *Appl. Psycholinguist.*, vol. 32, no. 1, pp. 113–140, Jan. 2011.

[12] H. Nusbaum and T. M. Morin, "Paying Attention to Differences Among Talkers," in *Speech perception, production and linguistic structure*, Y. Tohkura, Y. Sagisaka, and E. Vatikiotis-Bateson, Eds. Tokyo: Ohmasha Publishing, 1992, pp. 113–134.

[13] J. Chen, C. T. Best, M. Antoniou, and B. Kasisopa, "Cross-language categorisation of monosyllabic Thai tones by Mandarin and Vietnamese speakers: L1 phonological and phonetic influences," presented at the Proceedings of the Seventeenth Australasian International Conference on Speech Science and Technology, 2018, pp. 168–172.

[14] Chao. Y.R., "A system of tone-letters," *Maitre Phon.*, vol. 45, pp. 24–27, 1930.

[15] A. Reid *et al.*, "Perceptual assimilation of lexical tone: The roles of language experience and visual information," *Atten. Percept. Psychophys.*, vol. 77, no. 2, pp. 571–591, Feb. 2015.

[16] M. Yip, *Tone*. Cambridge: Cambridge University Press, 2002.

[17] M. Antoniou, M. D. Tyler, and C. T. Best, "Two ways to listen: Do L2-dominant bilinguals perceive stop voicing according to language mode?," *J. Phon.*, vol. 40, no. 4, pp. 582–594, Jul. 2012.

[18] U. Halekoh and S. Hojsgaard, "A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest," *J. Stat. Softw.*, vol. 59, no. 9, pp. 1–30, 2014.

[19] C. T. Best, "Devil or Angel in the Details?: Perceiving phonetic variation as information about phonological structure," in *Phonetics-Phonology Interface: Representations and Methodologies*, J. Romero and M. Riera, Eds. 2015, pp. 3–31.

[20] M. M. Faris, C. T. Best, and M. D. Tyler, "Discrimination of uncategorised non-native vowel contrasts is modulated by perceived overlap with native phonological categories," *J. Phon.*, vol. 70, pp. 1–19, Sep. 2018.

[21] J. E. Flege, "Second-language speech learning: Theory, findings, and problems," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. 1995, pp. 229–273.

[22] P. Escudero, *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. The Netherlands: Utrecht: LOT, 2005.