Perception of tonal contrasts: high-variability perceptual training with iconic orthographic representations

Yan Chen
University of Arizona
yanchen@email.arizona.edu

Abstract

This study examines the effect of orthographic representations for tones on the perception of five Cantonese tone pairs with high perceptual similarities. Native speakers of American English, Mandarin Chinese, and Standard Thai participated in the study. They completed a high-variability AXB pre-test (day 1), a high-variability AX training (day 2 - 4), and a high-variability AXB post-test followed by two generalization tests (day 5). Visual-trained listeners were presented with tone marks resembling the f0 heights and contours of the tones as feedback, whereas non-visual-trained listeners were not. Visual-trained listeners made correct decisions significantly faster in post-test and two generalization tests.

Index Terms: speech perception, perceptual learning, non-native perception, tones, Cantonese

1. Introduction

Laboratory training studies have revealed that perceptual mechanisms can be modified even within a short period of time with laboratory methods, and that perceptual training can direct listeners' attention to previously ignored acoustic features (e.g., [1], [2], [3], [4]). In addition, several studies have shown that a high-variability phonetic training paradigm [5], where listeners are exposed to phonetic variability within a phonetic category (through various talkers), is more effective than a single-talker paradigm and can lead to a more robust category formation. In the domain of tone perception, a high-variability same-different discrimination task improves non-native listeners' perception of Thai mid tones and low tones [6].

The perception of speech can also be influenced by orthography. Diacritic tone marks (e.g. ¯, ´, ˇ, `) have been shown to be facilitatory in processing Mandarin tones. The diacritics lead to significantly higher accuracy in tone production for native English speakers and native Japanese speakers [7] and they facilitate English speakers' learning of pseudo-words with Mandarin tones [8]. However, Mandarin tones all have different f0 shapes and so do their corresponding diacritics, and thus learning the association between Mandarin tones and diacritics may be easy. In order to see whether orthographic representations are truly effective or not, we examine the effect of tone marks on the perception of five tone pairs in Cantonese (a language with 6 lexical tones, shown in Figure 1) with high degree of phonetic similarities:

T2-T5: Both of them are rising tones starting from relatively the same pitch level but rise to different levels. Even some native Cantonese speakers experience difficulties in perceiving these tones [9] and this pair is most often noted as having merged in production [10].

T3-T6: These two level tones differ in f0 height. For a typical voice with f0 covering a range from 140Hz to 275Hz, the difference between T3 and T6 is just about 30Hz. Such a difference is just marginally sufficient to maintain a phonological contrast [11]. Some native speakers tend to merge T3 and T6 in perception and in production [12].

T4-T5, T4-T6, T5-T6: T4, T5, and T6 have very similar starting pitch levels.T6 is a level tone. T4 and T5 only have slight contours toward the end of the syllable. Some native speakers tend to merge T4 and T6 in perception and in production [12].

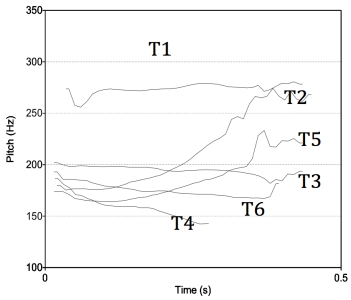

Figure 1: *Real-time Cantonese tones produced by a female native speaker (one token for each tone). The syllable carrying the tones is [si]. T1: High-Level, T2: High-Rising, T3: Mid-Level, T4: Low-Falling, T5: Low-Rising, T6: Low-Level.*

We also examine whether the effect of tone marks is L1-dependent. There languages were chosen because they form a tonality continuum: English (a non-tone language), Mandarin (a tone language with a relatively simple tone system), and Standard Thai (a tone language with five tones, some of which have similar f0 shapes).

2. Methods

2.1. Participants

97 Cantonese-naive participants were recruited as listeners (29 native speakers of American English and 30 native speakers of Mandarin at the University of Arizona, U.S.; 38 native speakers of Standard Thai at Mahidol University, Thailand). None of them had had musical training in the past 6 years (by self-report) at the time of testing. Eight female native speakers of Cantonese were recruited as talkers.

2.2. Stimuli

There were two kinds of stimuli: auditory and visual. For auditory stimuli, 20 Cantonese words were selected: 4 syllables ([fɐn], [si], [jim], [fu]) each carrying each of the five Cantonese

tones (T2, T3, T4, T5, and T6). The talkers produced the sentences [ha:22 jɐt5 kɔ:33 t͡si:22 hɐi22] _____ 'Next character is _____', where the target words were embedded at the end of the sentences. The target words were excised and normalized for peak amplitude using PRAAT [13]. Words containing the syllables [fɐn] and [si] produced by four of the 8 talkers were used in pre-test, training, and post-test. In generalization test 1 (new syllables with familiar talkers), words with [jim] and [fu] produced by the same four talkers in the training were used. Words with the syllables [jim] and [fu] produced by the other 4 talkers were used in generalization test 2 (new syllables with new talkers). For visual stimuli, five Chao Tone Letters [14] were used: ˥˧ for T2 (High-Rising), ˧ for T3 (Mid-Level), ˩ for T4 Low-Falling), ˨˧ for T5 (Low-Rising), and ˨ for T6 (Low-Level). The visual stimuli were used in training as feedback.

2.3. Procedure

2.3.1. Pre-Test

Pre-test was a high-variability AXB discrimination test. For example, a trial testing the contrast of T2 and T5 consisted of the following: [fɐn]$2_{[\text{Talker1}]}$ - [fɐn]$2_{[\text{Talker2}]}$ - [fɐn]$5_{[\text{Talker3}]}$. All possible combinations of the 5 tones were used, resulting in 10 tone pairs. Five tone pairs are considered "difficult pairs" and thus the targets: T2-T5, T3-T6, T4-T5, T4-T6, and T5-T6. The rest were "easy pairs" and they served as fillers. There were 160 trials in pre-test (2 syllable x 10 tone pairs x 8 talker combinations). Inter-stimulus interval (ISI) was set at 1500ms. The listeners were tested individually in a sound-attenuated booth (in US) or a quiet office (in Thailand). The 160 trials were randomly presented over headphones at a comfortable listening level using E−PRIME 2.0 Professional with a desktop computer (in US) or a laptop (in Thailand). The participants were asked to focus on the tones or pitch patterns of the stimuli and make their decisions as quickly as possible. No feedback was given. "A" and "B" responses, as well as reaction times on correct trials were collected. In order to familiarize listeners with the task, a 10-trial practice with feedback was conducted before the experimental trials. The practice section used syllable [jɐu] produced by three female speakers not used as talkers.

2.3.2. Training

The listeners participated in a 3-session perceptual training (1 hour per session, 1 session per day on three consecutive days), starting one day after pre-test. They were randomly assigned to two training paradigms: Auditory-only (AO) and Auditory-Visual (AV): English: AO (n=15), AV (n=15); Mandarin: AO (n=16), AV (n=16); Thai: AO (n=19), AV (n=19). The training made use of a high-variability AX "same-different" discrimination task with the same stimuli used in pre-test. "Same Trials" contained T2-T2, T3-T3, T4-T4, T5-T5, and T6-T6, and "Different Trials" contained only the "difficult pairs" in Pre-test: T2-T5, T3-T6, T4-T5, T4-T6, and T5-T6. For example, a "Same Trial" consisted of [fɐn]$2_{[\text{Talker1}]}$ - [fɐn]$2_{[\text{Talker2}]}$ and a "Different Trial" consisted of [fɐn]$2_{[\text{Talker1}]}$ - [fɐn]$5_{[\text{Talker2}]}$. ISI was set at 1500ms. Feedback was given after the listener responded to a trial. AO listeners saw the correct answer only, while AV listeners saw both the correct answer and the tone marks corresponding to the two auditory stimuli presented on the trial (Figure 2). All listeners were allowed to replay the stimuli as many times as they wanted to. "Same" and "different" responses were collected. There were 480 trials for each training session: 2 syllables x 10 pairs x 12 talker combinations (all possible talker combinations).

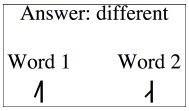

Figure 2: *An example of feedback given to AV listeners. AO listeners received the same feedback except for the tone marks.*

2.3.3. Post-test and Generalization Tests

One day after the last training session, participants took post-test, followed by two generalization tests. The procedure of post-test was identical to pre-test. The generalization tests were high-variability AXB tasks as well. Generalization test 1 consisted of two new syllables [jim] and [fu] produced by the same 4 talkers in training (80 trials: 2 syllables x 10 tone pairs x 4 talker combinations). Generalization Test 2 consisted of the new syllables produced by 4 new talkers (80 trials: 2 syllables x 10 tone pairs x 4 talker combinations).

3. Results

The results from 9 listeners were excluded. Six of them (3 English, 1 Mandarin, and 2 Thai) completed significantly fewer trials on one or more training sessions. Another 2 Thai listeners did not finish post-test due to computer errors.

For each listener at each test (pre-test, post-test, generalization test 1, and generalization test 2), percent correct response was calculated as accuracy. Raw reaction times (RTs) were log-transformed to normalize data distribution. For pre- and post-test performance, percent correct response and log-RTs data each was submitted to a 4-factor mixed-effect ANOVA with L1 (English, Mandarin, Thai) and Training Paradigm (AO, AV) as between-subject factors, and Tone Pair (T2-T5, T3-T6, T4-T5, T4-T6, T5-T6) and Test (pre-test, post-test) as within-subject factors. For performance on the two generalization tests, percent correct response and log-RTs data each was submitted to a 4-factor mixed-effect ANOVA with L1 (English, Mandarin, Thai) and Training Paradigm (AO, AV) as between-subject factors, and Tone Pair (T2-T5, T3-T6, T4-T5, T4-T6, T5-T6) and Test (Gen1, Gen2) as within-subject factors. Results related to the factor Training Paradigm are reported below in details, as it is the main research question in this paper.

3.1. Pre-test vs. Post-test

Figure 3 shows the overall accuracy by L1s, tone pairs, tests, and training paradigms. A 4-way mixed-effect ANOVA showed a main effect of Training Paradigm ($F(1,82)=5.38$, $p<.03$), suggesting that AV outperformed AO both before and after training. We concluded that it was a coincidence that AV achieved higher accuracy in pre-test because the participants were randomly assigned to the training groups. Training Paradigm did not participate in any significant interaction. AV did not improve significantly more than AO on accuracy. There were two significant 2-way interactions not related to Training Paradigms: Pair by L1 ($F(8,328)=12.99$, $p<.001$) and Pair by Test ($F(4, 328)=29.91$, $P<.001$). We found that listeners' performance on T3-T6 was worse after training than before training ($F(1,87)=15.52$, $p<.001$), but there seems to be a trend that AV had higher accuracy than AO in post-test while their performance in pre-test was the same. This suggests that visual training might have pre-

vented listeners from losing more of the level tone distinctions.

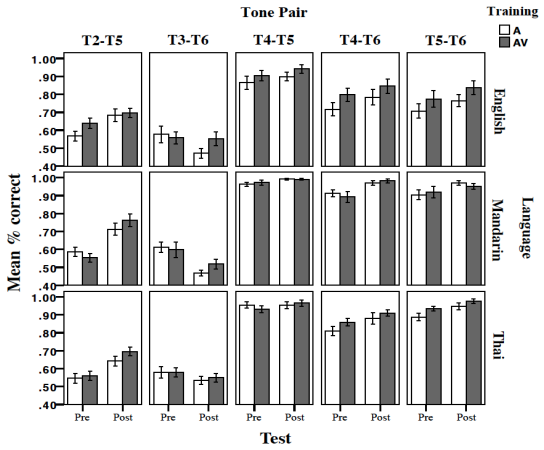

Figure 3: *Mean % correct response by L1s, tone pairs, tests, and training paradigms. Error bars indicate the standard error of the mean.*

For reaction times, Figure 4 shows the overall result of log-RTs by L1s, tone pairs, tests, and training paradigms. The factor Training Paradigm participated in a significant two-way interaction: Test by Training Paradigm ($F(1,82)=4.13$, $p<.05$). At pre-test, AV's log-RTs did not significantly differ from AO despite having higher accuracy, as mentioned earlier. At post-test, AV responded significantly faster than AO ($F(1,86)=4.63$, $p<.04$) (Figure 5). AV listeners were able to make correct decisions faster than AO listeners after training across the board. The factor L1 participated in two significant two-way interactions, not related to Training Paradigm: Test by L1 ($F(2,82)=4.48$, $p<.02$), and Pair by L1 ($F(8,328)=9.11$, $p<.001$). The interaction of Test by Pair was also significant ($F(4,328)=11.93$, $p<.001$).

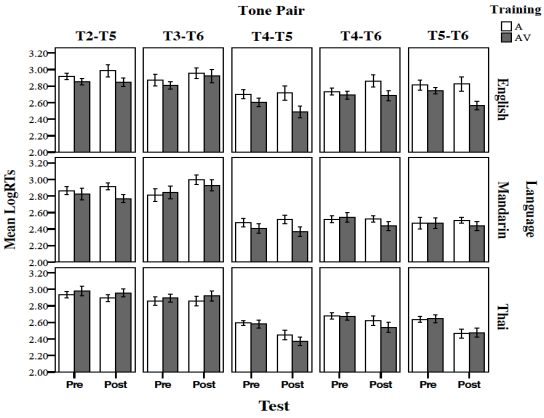

Figure 4: *Mean log-RTs by L1s, tone pairs, tests, and training paradigms. Error bars indicate the standard error of the mean.*

3.2. Generalization Tests

The overall result of accuracy is shown in Figure 6. A 4-way mixed-effect ANOVA did not show a significant main effect of Training Paradigm ($F<1$), and Training Paradigm did not participate in any significant interaction. This means that AV's performance was comparable to AO. The other three factors interacted significantly ($F(8,328)=2.84$, $p<.006$).

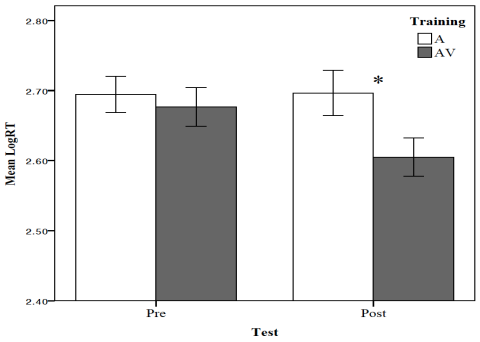

Figure 5: *Mean log-RTs by tests and training paradigms. AV listeners made correct decisions faster than A listeners in post-test. Error bars indicate the standard error of the mean.*

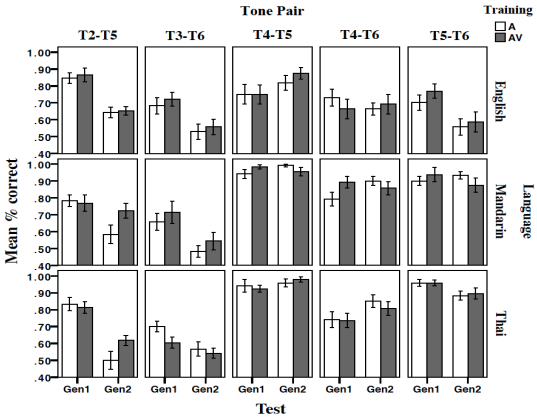

Figure 6: *Mean % correct response by L1s, tone pairs, tests, and training paradigms. Error bars indicate the standard error of the mean. Gen 1 was a generalization test with new words produced by familiar talkers. Gen 2 was a generalization test with new words produced by new talkers.*

For reaction times, the overall result is shown in Figure 7. A significant main effect of Training Paradigm was found ($F(1,82)=4.04$, $p<.05$). As shown in Figure 8, AV had significantly shorter log-RTs than AO across L1s, tone pairs, and tests, which means that AV made correct decisions faster when hearing new words from familiar talkers as well as unfamiliar talkers. The lack of Training by L1 interaction suggests that the effect of training was consistent across L1s. In addition to a significant main effect of Training Paradigm, there was a significant two-way interaction of Pair by L1 ($F(8,328)=8.67$, $p<.001$).

4. Discussion & Conclusion

The main purpose of this study is to examine whether the use of tone marks in high-variability phonetic training facilitates the perception of tones with small perceptual differences, and whether this kind of orthographic representation is more helpful for listeners from certain L1 backgrounds. Cantonese-naive listeners from three language groups (American English, Mandarin, and Standard Thai) participated in a three-day high-

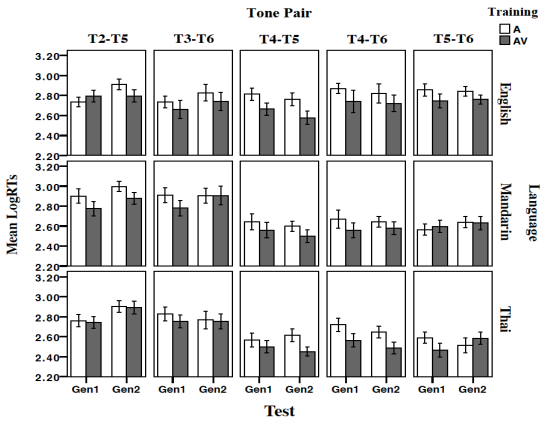

Figure 7: *Mean log-RTs by L1s, tone pairs, tests, and training paradigms. Error bars indicate the standard error of the mean. Gen 1 was a generalization test with new words produced by familiar talkers. Gen 2 was a generalization test with new words produced by new talkers.*

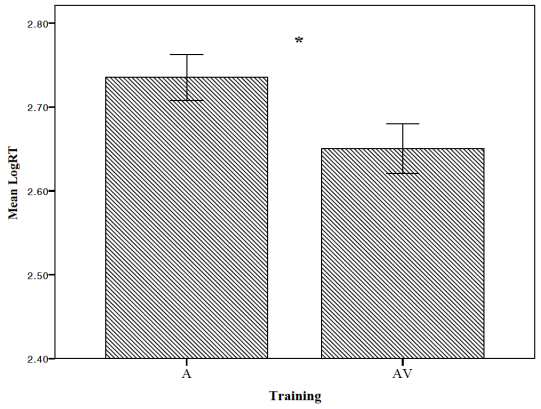

Figure 8: *Mean log-RTs by training paradigms. AV listeners made correct decisions faster than A listeners. Error bars indicate the standard error of the mean.*

variability perceptual training. The fact that training paradigms did not significantly interact with language backgrounds at all suggests that the training had more or less the same effect on both tonal language speakers and non-tonal language speakers.

Both AV and AO listeners improved on accuracy after training, and AV did not have significantly more improvement than AO despite the training with additional information. However, AV's log-RTs on correct trials were significantly shorter than AO in post-test. Tone marks did facilitate the learning of novel tones to some extent, as AV listeners were more certain about the within-category similarities and the between-category differences in the tones and thus could correctly respond faster. The results of the two generalization tests showed similar patterns. AV did not differ from AO on accuracy, but reaction times data showed that not only did AV make correct decisions significantly faster than AO when talkers' voices were familiar (generalization test 1), they also responded faster when talkers' voices were unfamiliar at all (generalization test 2).

The tone marks presented in training are abstractions for F0 contours. Learning to capture the relationship between the auditory stimuli and the visual abstractions may increase listeners' cognitive load. Nevertheless, the mapping between visual abstractions and auditory information helps create better category formation faster. There was a trend that AV had higher accuracy than AO for T3-T6, the only tone pair on which accuracy decreased after training. This indicates that while both groups were impaired by the training on the level tones with high variability, orthographic representations may have prevented AV listeners from further impairment. Although AV listeners formed better tone categories faster, the categories formed were not robust enough for them to outperform AO listeners in terms of accuracy. It should be noted that these results were obtained from 3 hours of high-variability perceptual training (1 hour per day), which is a difficult task in a very short period of time. It is possible that with longer training the facilitatory effect may be shown more clearly.

5. References

[1] Pisoni, D. B., Aslin, R. N., Perey, A. J. and Hennessy, B. L., "Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants", Journal of Experimental Psychology: Human Perception and Performance, 8:297-314, 1982.

[2] Logan, J. S., Lively, S. E. and Pisoni, D. B., "Training Japanese listeners to identify English /r/ and /l/: a first report", Journal of the Acoustical Society of America, 89:874-886, 1991.

[3] Wang, Y., Spence, M. M., Jongman, A., and Sereno, J.A., "Training American listeners to perceive Mandarin tones", Journal of the Acoustical Society of America, 106:3649-3658, 1999.

[4] Francis, A. L., Cioccaa V., Ma, L. and Fenn, K., "Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers", Journal of Phonetics, 36:268-294, 2008.

[6] Wayland, R. and Li, B., "Effects of two training procedures in cross-language perception of tones", Journal of Phonetics, 36:250-267, 2008.

[5] Lively, S. E., Logan, J. S., Pisoni, D. B., "Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories", The Journal of the Acoustical Society of America, 94(3):1242-1255, 1993.

[7] McGinnis, S., "Tonal spelling versus diacritics for teaching pronunciation of Mandarin Chinese", Modern Language Journal, 81(2):228-236, 1997.

[8] Showalter, C. E. and Hayes-Harb, R., "Unfamiliar orthographic information and second language word learning: A novel lexicon study", Second Language Research, 29(2):185-200, 2013.

[9] Mok, P. K. and Zuo, D., "The separation between music and speech: Evidence from the perception of Cantonese tones", Journal of Acoustical Society of America, 132(4): 2711-2720, 2012.

[10] Bauer, R. S, Cheung, K. H. and Cheung, P. M., "Variation and merger of the rising tones in Hong Kong Cantonese", Language Variation Change, 15:211-225, 2003.

[11] Hart, J., "Differential sensitivity to pitch distance, particularly in speech", Journal of Acoustical Society of America, 69:811-821, 1981.

[12] Mok, P. K. and Wong, P. W. Y., "Production of the merging tones in Hong Kong Cantonese: Preliminary data on monosyllables", Proceedings of Speech Prosody 2010, Chicago, IL, 2010.

[13] Boersma, P. and Weenink, D., Praat: doing phonetics by computer [Computer program]. Version 6.0.17.

[14] Chao, Y. R., "A system of tone-letters", Le Maitre Phonetique, 30:24-27, 1930.