INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

L2-Mandarin regional accent variability during Mandarin tone-word training facilitates English listeners' subsequent tone categorizations

Yanping Li[1], *Michael D. Tyler*[2], *Denis Burnham*[1], *Catherine T. Best*[1]

[1]The MARCS Institute, Western Sydney University, Australia
[2]Independent Researcher, Australia

yanping.li@westernsydney.edu.au, tylerspeechscience@gmail.com
denis.burnham@westernsydney.edu.au, c.best@westernsydney.edu.au

Abstract

We examined how accent variability during training on minimal-tone-contrast Mandarin words affects English listeners' subsequent generalization of tone categorization and discrimination to new talkers and accents. English listeners underwent 6 days of training on 16 pseudowords (4 tones × 4 sets) produced by 12 talkers of either Beijing (n = 24) or a mix of Beijing, Yantai, and Guangzhou (n = 24) accents. In a post-training test, they used tone contour icons to categorize the tones produced by new talkers with a familiar and an unfamiliar accent, and to discriminate all possible tone contrasts for the new talkers. While both training groups *discriminated* all six tone contrasts for both new talkers, only the multiple accent group reliably *categorized* all four tones. The single accent group failed to correctly categorize the falling tone in both generalization tests. These results suggest that accent variability during tone-word training can facilitate subsequent tone categorization.

Index Terms: talker variability; L2-Mandarin accents; lexical tone discrimination; tone contour categorization; English leaners

1. Introduction

English listeners have difficulties perceiving the four Mandarin tones (i.e., level, rising, dipping, and falling) [1], which are used to distinguish word meanings (e.g., for the consonant-vowel [CV] syllable /ma/, level = *mother*, rising = *hemp*, dipping = *horse*, and falling = *curse*). Fundamental frequency (f_0) contour and height are the primary acoustic parameters of Mandarin tones (e.g., [2]–[4]). Tone-word training, which maps a word's tone category to its meaning (and vice versa) [5], can improve perception of Mandarin f_0 patterns [6].

Importantly, using high-acoustic-variability tokens during tone-word training enhances Mandarin word learning, particularly for English learners with a high aptitude for tone perception [7]. To provide acoustic f_0 variations for each tone, the training stimuli in [6] and [7] were based on natural tokens resynthesized to vary their f_0 values using the pitch-synchronous overlap and add (PSOLA) method. Since the seed stimuli were produced by only one talker, the stimuli lacked talker variability, which facilitates second language speech training on consonants (e.g., [8], [9]) and vowels (e.g., [10], [11]). For example, Japanese learners exposed to multiple talkers (n = 5, high variability) improved in identifying English words containing /r/-/l/ contrasts and generalized their learning to new words produced by an unfamiliar talker. In contrast, learners exposed to only a single talker (low variability) during training did not [9]. Talker variability was used in [5], in which naturally produced Mandarin minimal pair tone-words were presented to English listeners with low (single talker) versus moderately high (n = 4 talkers) talker variability. Learners completed tone contour categorization and discrimination before and after tone-word training, and the testing stimuli were produced by new talkers with untrained vowels. Both training groups improved their tone categorization, but the multiple talker group did not outperform the single talker group. However, neither the single nor the multiple talker group improved in tone discrimination. These findings suggest that moderate talker variability during tone-word training does not boost tone perception, as it does in L2 segment learning (e.g., [6–9]).

Another form of variability is accent variability in Mandarin tones (see [12] for f_0 contour deviations from the norm). While effects of talker variability in L2 word learning have been examined even for tones, little is known about the effects of accent variability during tone-word training. A recent study we employed accent variability in Mandarin tone-words to train English learners' and then test their tone discrimination and categorization (using tone contour icons) with untrained words [13]. Following [5], naturally produced tone-words were used in [13] for training, yielding accent variability together with talker variability. To tease apart accent and talker variability, a control group was exposed to multiple talkers (n = 12) with one accent, whereas the experimental group learners were exposed to 12 talkers with three accents, four talkers of each accent. Results were interpreted using principles of the Perceptual Assimilation Model (PAM: [14], [15]), which indicated that learners exposed to the single accent condition (constant accent, multiple talkers) showed Uncategorized icon choices for the falling tone after training, whereas the multiple accent group (multiple accents and talkers) reliably Categorized all four Mandarin tones. Meanwhile, a single accent group showed poorer discrimination of the falling vs. level tone contrast than the other five tone contrasts (e.g., level vs. rising, level vs. dipping, rising vs. dipping, rising vs. falling, and dipping vs. falling), whereas the multiple accent group showed high sensitivity to each of the six tone contrasts. These findings indicate that accent variability in tone-word training facilitates English listeners' tone perception. However, it remains unclear whether improvements in tone perception after multiple accent training generalize beyond the talkers and accents in the training set.

The aim of this study is to extend [13] to examine the extent to which multiple accent word training can transfer to tone perception for untrained talkers and accents. Accordingly, we conducted tone-word training with accent variability, and then

10.21437/Interspeech.2023-145

conducted talker and accent generalization tests on tone categorization and discrimination. Since both the single and multiple accent training groups are exposed to high talker variability (n = 12), English learners should generalize training to new talkers. Therefore, we included new talkers in both tone perception generalization tests. However, as accent variability is a core focus of this study, we also compared generalization to a new talker with a trained (*familiar*) accent versus a new talker with an untrained (*unfamiliar*) accent. Both training groups should generalize tone perception to a new talker with a familiar accent, but only the multiple accent group should generalize well to a new talker with an unfamiliar accent. The single accent group may instead show difficulties in generalization to a new accent that are similar to the pre-training tone perception patterns found in [13], e.g., Uncategorized tone icon choices for the level and falling tones, and poorer discrimination of level vs. falling tones than that of the other five tone contrasts.

2. Experiment

2.1. Method

2.1.1. Participants

English native speakers (n = 48, see [13] for participant details) were recruited online, and randomly assigned to the single accent (n = 24, M_{age} = 24.5 years, SD = 5.8, 14 females) or multiple accent (n = 24, M_{age} = 25.5 years, SD = 5.1, 15 females) groups. They were paid using Prezzee eGift cards for their participation.

2.1.2. Stimuli

There were 16 real Mandarin words for training (four CV syllables, {/ba/, /di/, /du/, /gu/} × 4 Mandarin tones). Training stimuli were produced by female native speakers (see [12] for recording procedure), either 12 from Beijing (talker variability, constant accent) or four each from Beijing, Yantai, and Guangzhou (accent and talker variability). Each word was produced four times, yielding 768 tokens (12 talkers × 16 words × 4 tokens) in each training group. Each token was verified by four female native listeners from the same accent community as that of the talker (see [16] for word verification procedure). Both groups thus heard 12 talkers (high talker variability). Word meanings were presented using black-and-white pictures from [17], which were counterbalanced across participants, yielding 16 pseudowords for each participant.

Another set of 16 words (four CV syllables, {/ga/, /ti/, /tu/, /pu/} × 4 tones) were newly recorded for testing generalization of tone icon categorization and discrimination to new talkers with familiar Beijing or unfamiliar Shanghai accents. They were produced by female Beijing (familiar/trained accent; M_{age} = 19.0 years) and Shanghai (unfamiliar/untrained accent, M_{age} = 24.0 years) native speakers in a soundproof booth at the Speech Acquisition and Intelligent Technology Lab, Beijing Language and Culture University, Beijing, China. The procedure for word recording and word verification was the same as that for training words. This resulted in a total of 128 tone-word tokens (2 talkers × 16 words × 4 tokens).

2.1.3. Procedure

Perceptual tasks were conducted remotely using E-prime Go 1.0. Learners ran experiments on their own Windows 10 laptops/desktops. To ensure data quality, they were tested in a quiet room via a ZOOM meeting with the first author.

Participants completed tone icon categorization and discrimination of the pre-training tone stimuli (partly reported in [16]), Mandarin tone-word training and post-training word verification (reported in [18]), and generalization tests of tone perception to the new talkers and accent. This study focused on effects of L2-Mandarin accent variability on generalization of tone perception. Therefore, we briefly describe tone-word training procedure here, but see [18] for full details. English learners learned the 16 tone-pseudowords in a picture-to-word paradigm, across six training sessions that used quizzes with feedback [6], in the single or the multiple accent condition.

After completing the post-training tone perception task with the original pre-training tone stimuli (see also [16]), learners were presented with tone generalization tests with new talkers with a familiar Beijing then an unfamiliar Shanghai accent, and in each of these, discrimination then tone contour icon categorization tasks were conducted, using the same categorical AXB discrimination and icon categorization tasks as in [13]. In the AXB task, there were 384 trials (6 tone contrasts × 4 syllables × 4 AXB trial types × 4 times). In each trial, learners were told to click a button in the center of their screen then following stimulus presentations, to click on the "1" or "3" each side of and equidistant from the central button to indicate whether the X item matched category A or B. To ensure online data quality, participants were required to hold the central activation button until they heard the third stimulus. Release before then led to an automatic trial abort. This modification of the AXB was selected as winner in the E-prime Challenge 2022. Trials that were aborted (1577 occurrences, or 2.14% of all trials) were repeated at the end of each block.

Following a short break, learners completed the tone icon categorization task with 64 trials (16 words × 4 tokens) of individual test items. The response time-out was 3.5 s. The activation button design was also used for tone icon categorization. Trials that aborted automatically or lacked a response within 3.5 s (129 occurrences, or 2.10% of all trials) were repeated at the end of each block.

2.2. Results

2.2.1. Categorization of Mandarin tone contour icons

Figure 1 shows mean percent of icon choice for the four Mandarin tones in each generalization test by the single and multiple accent groups. Following [13], two criteria were used to determine whether a given tone was Categorized: (i) a given Mandarin tone icon must be selected significantly more than chance level (25%), and (ii) it must be chosen significantly more often than any of the other Mandarin tone icons. For each generalization test in each training condition, separate one-sample t-tests against chance (25%) were conducted for criterion (i) using R [19] with the *Student's t-Test* function. Multiple linear mixed-effects models were built to assess criterion (ii) with the *lmer* function from package *lme4* [20]. Percentage of choices for each tone icon category was specified as the dependent variable. Training groups, generalization tests, and Mandarin tones were specified as fixed effects, and participants as a random effect. The Kenward-Roger approximation to the degrees of freedom was used to calculate the p values for the fixed-effects factors [21] and the *Anova* function from package *car* [22] was used to calculate F. Pairwise comparisons were conducted with *lsmeans* [23] in R

whenever there were significant interactions of training group × generalization test × Mandarin tones.

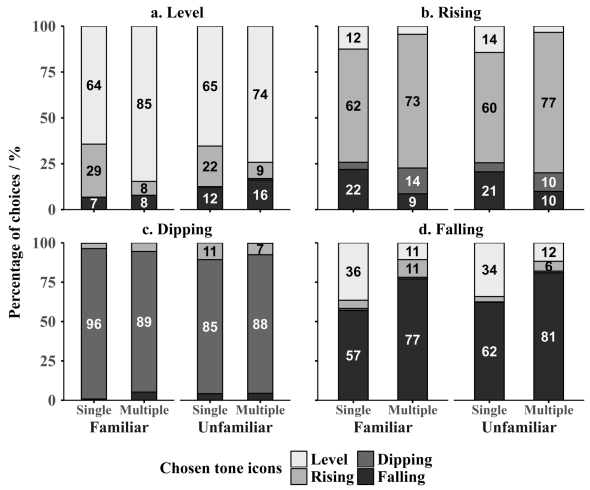

Figure 1: *Mean choice percent of the four tone icons for Mandarin tone stimuli in generalization tests. Choices < 5% were not labelled here.*

In the generalization test with the new talker with the familiar Beijing accent, when responding to the level tone stimuli, both training groups split their choices between level (M_{single} = 64.32%, *SD* = 38.90; $M_{multiple}$ = 84.64%, *SD* = 29.72) and rising icons (M_{single} = 28.91%, *SD* = 38.70; $M_{multiple}$ = 7.55%, *SD* = 19.06). However, only their level icon choices were significantly above chance (single: $t(23) = 4.95$, $p < .0001$; multiple: $t(23) = 9.83$, $p < .0001$), and they selected the level icon significantly more than the rising icon (single: Estimate = 35.42, *SE* = 6.46, $t(345) = 5.48$, $p < .0001$; multiple: Estimate = 65.36, *SE* = 6.46, $t(345) = 10.12$, $p < .0001$), suggesting that both training groups Categorized the level tone to the level icon. Nonetheless, the multiple accent group (*M* = 7.55%, *SD* = 19.06) selected marginally fewer rising icon options than the single accent group (*M* = 28.91%, *SD* = 38.70), Estimate = -21.35, *SE* = 6.46, $t(345) = -3.3$, $p = .08$. In the generalization test to the new talker with the unfamiliar Shanghai accent, the level tone was Categorized by both training groups. However, the marginal group difference for the rising icon disappeared, because the single accent group made fewer rising and more falling icon choices, whereas the multiple accent group chose fewer level and more falling icons.

For rising tone stimuli, in the Beijing (familiar accent) generalization test the single accent group split their choices between rising (*M* = 61.72%, *SD* = 36.78) and falling icons (*M* = 21.88%, *SD* = 31.71), whereas the multiple accent group split their choices between rising (*M* = 72.92%, *SD* = 31.90) and dipping icons (*M* = 14.06%, *SD* = 25.15). However, only the rising icon was selected significantly greater than chance for both groups (single: $t(23) = 4.89$, $p < .0001$; multiple: $t(23) = 7.36$, $p < .0001$), and it was chosen significantly more often than the falling icon by the single accent group (Estimate = 39.84, *SE* = 7.04, $t(345) = 5.66$, $p < .0001$), and more often than the dipping icon (Estimate = 58.85, *SE* = 7.04, $t(345) = 8.36$, $p < .0001$) by the multiple accent group, indicating that both groups Categorized the rising tone. In the Shanghai (unfamiliar accent) generalization test both groups also Categorized the rising tone to the rising icon. But only the multiple accent group also showed residual choices of the dipping icon ($M_{Gen\text{-}talker}$ = 14.06%, *SD* = 25.15; $M_{Gen\text{-}accent}$ = 10.16%, *SD* = 24.44), indicating that they perceived tones partly based on f_0 contour instead of merely on f_0 height, given that rising and dipping tones share a falling-rising contour (e.g., [24]).

Both groups Categorized the dipping tone to the dipping icon, selecting the dipping icon significantly above chance in generalization tests to new talkers of the familiar Beijing accent (M_{single} = 95.57%, *SD* = 11.87, $t(23) = 29.13$, $p < .0001$; $M_{multiple}$ = 89.32%, *SD* = 25.30, $t(23) = 12.45$, $p < .0001$), and of the unfamiliar Shanghai accent (M_{single} = 85.16%, *SD* = 16.87, $t(23) = 17.47$, $p < .0001$; $M_{multiple}$ = 88.02%, *SD* = 22.94, $t(23) = 13.46$, $p < .0001$). The dipping icon was also selected significantly more often than the other three tone icons by both training groups in each generalization test.

In the generalization test with the new talker with the familiar Beijing accent, when responding to the falling tone stimuli, the single accent group split their choices between level (*M* = 36.46%, *SD* = 30.43) and falling (*M* = 57.03%, *SD* = 34.40) icons. Both level ($t(23) = 1.84$, $p = .04$) and falling ($t(23) = 4.56$, $p < .0001$) icons were selected significantly above chance, but there was no significant difference between them (Estimate = -20.57, *SE* = 7.07, $t(345) = -2.91$, $p = .22$), suggesting that the falling tone stimuli was Uncategorized. However, the multiple accent group Categorized the falling tone stimuli as there was only the falling icon (*M* = 77.08%, *SD* = 34.95) selected significantly above chance ($t(23) = 7.3$, $p < .0001$), which was also selected significantly more often than the other three tone icons. In the generalization test on the new talker with the unfamiliar accent, both training groups retained their assimilation type, e.g., Uncategorized vs. Categorized for the single vs. multiple accent group. Specifically, the single accent group continued to split their choices between level (*M* = 34.11%, *SD* = 34.92) and falling (*M* = 62.24%, *SD* = 35.95) icons. The level icon was selected marginally above chance ($t(23) = 1.28$, $p = .10$), but the falling icon was selected significantly above chance ($t(23) = 5.07$, $p < .0001$), which was also selected significantly more often than the level icon, Estimate = -28.12, *SE* = 7.07, $t(345) = -3.98$, $p = .01$. While the assimilation type was still Uncategorized, the single accent group improved slightly as they shifted their choices from level to falling icons. In the multiple accent condition, only the falling icon (*M* = 80.99%, *SD* = 34.01) was selected significantly above chance ($t(23) = 8.07$, $p < .0001$), which was also selected more often than the other three tone icons, suggesting that the falling tone stimuli was Categorized.

2.2.2. Discrimination of Mandarin tones

Figure 2 displays tone discrimination in both training groups for each generalization test. To control for response bias, each participant's discrimination data were transformed to *A* scores [25]:

$$A = \begin{cases} \frac{3}{4} + \frac{H - F}{4} - F(1 - H) & \text{if } F \le 0.5 \le H; \\ \frac{3}{4} + \frac{H - F}{4} - \frac{F}{4H} & \text{if } F \le H < 0.5; \\ \frac{3}{4} + \frac{H - F}{4} - \frac{1 - H}{4(1 - F)} & \text{if } 0.5 < F \le H \end{cases}$$

Where *F* is false alarm rate and *H* is hit rate (see [26] for details on Signal Detection Theory). Higher *A* values indicate better discrimination.

Multiple linear mixed-effects models were built on *A* values, with training groups, generalization tests, and tone contrasts specified as fixed effects and participants as a random

effect. Calculation of p and F values was the same as in the tone categorization model. The main effects of training groups, $F(1, 549) = 18.38$, $p < .0001$; generalization, $F(1, 549) = 4.84$, $p = .03$, and tone contrasts, $F(5, 545) = 3.25$, $p = .007$, and the interaction of training groups × generalization tests × tone contrasts, $F(23, 527) = 1.91$, $p = .007$, were all significant.

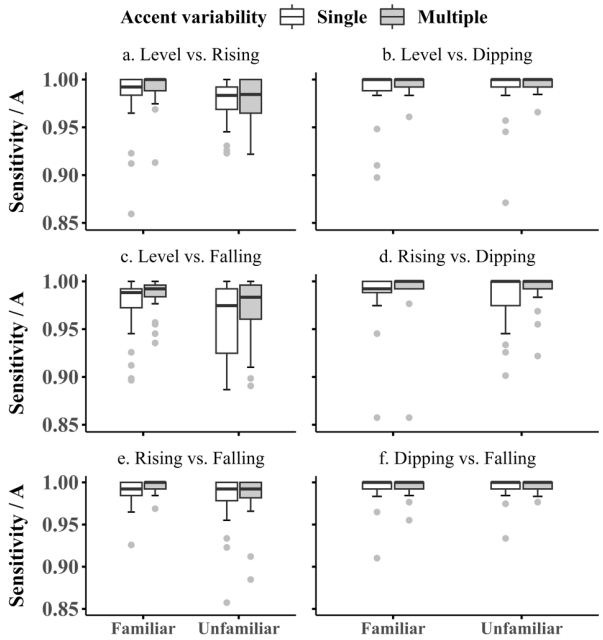

Figure 2: *AXB discrimination for the six tone contrasts in the two generalization tests, which are shown as transformed A scores [25]. Outliers are displayed as grey points.*

Pairwise comparisons showed that the single accent group was less sensitive to tone differences than the multiple accent group ($M_{single} = 0.96$ in A scores, $SD = 0.08$ vs. $M_{multiple} = 0.98$, $SD = 0.04$, $p < .0001$). Meanwhile, both groups were more sensitive to tone differences for the new talker with a Beijing (familiar) accent than the new talker with a Shanghai (unfamiliar) accent ($M_{Gen\text{-}talker} = 0.98$, $SD = 0.06$ vs. $M_{Gen\text{-}accent} = 0.97$, $SD = 0.07$; $p = 0.03$). While discrimination of the level vs. falling ($M = 0.96$, $SD = 0.07$) and the level vs. rising contrasts ($M = 0.97$, $SD = 0.06$) did not differ significantly, the A score for the level vs. falling tone contrast was significantly lower than that for rising vs. dipping ($M = 0.98$, $SD = 0.05$; $p = .02$), rising vs. falling ($M = 0.98$, $SD = 0.07$; $p = .06$), and dipping vs. falling ($M = 0.98$, $SD = 0.06$; $p = .02$) tone contrasts, and was marginally lower than that for the level vs. dipping contrast ($M = 0.98$, $SD = 0.07$; $p = .06$). There were no significant differences between any other pairs of tone contrasts. Interestingly, there were no significant differences in A scores between training groups or generalization tests for any tone contrast, suggesting that both training groups had acquired tone differences in Mandarin.

3. Discussion

This paper addressed English learners' tone perception in two generalization tests on tone icon categorization and tone discrimination after completing a 6-session training regimen on 16 Mandarin minimal-tone-set words. It follows up on [13], which showed that presenting multiple L2-Mandarin accent variations during tone-word training enhanced English learners' tone perception. The two generalization tests assessed the extent to which English learners transfer their improvements in tone perception, following tone-word training, to new talkers with familiar and unfamiliar accents.

In generalization to a new talker with a familiar (Beijing) accent, our predictions on tone perception for both training groups were upheld. Specifically, the single accent training group showed similar results as they had in the post-training test reported in [13], i.e., they categorized the level, rising, and dipping tones using tone contour icons, but they failed to categorize the falling tone as they displayed falling vs. level tone icon confusions. Conversely, and again in line with their results in [13], the multiple accent group categorized all four tones and discriminated all six tone pairs. These findings indicate that talker variability during tone-word training facilitates tone perception, which is consistent with some findings (e.g., [6–9]), but not others (e.g., [5]).

In generalization to a new talker with an unfamiliar (Shanghai) accent, however, the single accent group performed better than we predicted, that is, they categorized the level tone to a level icon, although their icon choices were Uncategorized for the falling tone stimuli, indicating that they retained their post-training patterns instead of reverting to the pre-training level of performance. While the generalization test on the unfamiliar accent should be more difficult than on the familiar one, the multiple accent group categorized the four tones without displaying differences in tone contrast discrimination, which is consistent with our predictions. Both groups retained perceptual patterns in the post-training test, but only the multiple accent condition correctly categorized the four tones, supporting the main outcome and conclusion here: that *accent variability during tone-word training facilitates subsequent generalization of tone perception to a new talker and accent*.

Both [5] and the current study investigated the effect of high talker variability during tone-word training on subsequent tone contour categorization and discrimination, but their conclusions differ. One possible explanation for the difference in conclusions may be that [5] focused on percentage of improvement in tone contour categorization, but we were more interested in Categorized vs. Uncategorized differences across training conditions and generalization test types, interpreted in relation to the framework of PAM (e.g., [14], [15]). The difference between [5] and the current study can also be caused by talker variability. There were only four talkers in [5] produced training words, but 12 talkers in the current study, which may present much more talker/accent variability during training.

4. Acknowledgements

This research was supported by a Western Sydney University (WSU) – China Scholarship Council (CSC) joint scholarship, Candidature Research Funds from the MARCS Institute, WSU, and partly by the Australian Linguistic Society Research Grants, which were all awarded to the first author. We are grateful for online recruitment assistance of CareerHubs at Australian universities, and technical support from Clarissa Montino, Johnson Chen and Dr. Chris Wang at MARCS. We sincerely thank Prof. Jingsong Zhang from the School of Computer Science, Beijing Language and Culture University (BLCU, China) for providing the recording venue, and to Dr. Jingwen Huang from Center for Studies of Chinese as a Second Language, BLCU, for recording assistance, and above all thank the participants.

5. References

[1] L. Liu *et al.*, "Learning to perceive non-native tones via distributional training: Effects of task and acoustic cue weighting," *Brain Sci.*, vol. 12, Apr. 2022, doi: 10.3390/brainsci12050559.

[2] J. M. Howie, *Acoustical studies of Mandarin vowels and tones*. in Princeton-Cambridge studies in Chinese linguistics, no. 6. New York/Cambridge: Cambridge University Press, 1976.

[3] J. J. Dreher and P.-C. Lee, "Instrumental investigation of single and paired Mandarin tonemes," *Monum. Serica*, vol. 27, no. 1, pp. 343–373, 1968, doi: 10.1080/02549948.1968.11731059.

[4] C. -K. Chuang and S. Hiki, "Acoustical features and perceptual cues of the four tones of Standard Colloquial Chinese," *J. Acoust. Soc. Am.*, vol. 52, p. 146, 1972, doi: 10.1121/1.1981919.

[5] H. Dong, M. Clayards, H. Brown, and E. Wonnacott, "The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones," *PeerJ*, vol. 7, p. e7191, 2019, doi: 10.7717/peerj.7191.

[6] P. C. M. Wong and T. K. Perrachione, "Learning pitch patterns in lexical identification by native English-speaking adults," *Appl. Psycholinguist.*, vol. 28, no. 4, pp. 565–585, 2007, doi: 10.1017/S0142716407070312.

[7] T. K. Perrachione, J. Lee, L. Y. Y. Ha, and P. C. M. Wong, "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," *J. Acoust. Soc. Am.*, pp. 461–472, 2011, doi: 10.1121/1.3593366.

[8] J. S. Logan, S. E. Lively, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/: A first report," *J. Acoust. Soc. Am.*, vol. 89, no. 2, pp. 874–886, 1991, doi: 10.1109/TMI.2012.2196707.Separate.

[9] S. E. Lively, J. S. Logan, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories," *J. Acoust. Soc. Am.*, vol. 94, no. 3, pp. 1242–1255, 1993, doi: 10.1121/1.408177.

[10] Y. Li, C. Best, C. Cao, and J. Zhang, "Hybrid perceptual training to facilitate the learning of nasal final contrasts by highly proficient Japanese learners of Mandarin," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds., Melbourne, Australia., 2019, pp. 2567–2571.

[11] P. Iverson and B. G. Evans, "Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers," *J. Acoust. Soc. Am.*, vol. 126, no. 2, pp. 866–877, 2009, doi: 10.1121/1.3148196.

[12] Y. Li, C. T. Best, M. D. Tyler, and D. Burnham, "Tone variations in regionally accented Mandarin," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association*, H. Meng, B. Xu, and T. F. Zheng, Eds., 2020, pp. 4158–4162. doi: 10.21437/interspeech.2020-1235.

[13] Y. Li, C. T. Best, M. D. Tyler, and D. Burnham, "L2-Mandarin regional accent variability during lexical tone word training facilitates naive English listeners' tone categorization and discrimination," in *Proceedings of the 18th Australasian international Conference on Speech Science and Technology*, R. Billington, Ed., Canberra, Australia, 2022, pp. 156–160.

[14] 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.

[15] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Second language speech learning: The role of language experience in speech perception and production*, M. J. Munro and O.-S. Bohn, Eds., Amsterdam: John Benjamins, 2007, pp. 13–34.

[16] Y. Li, C. Best, M. Tyler, and D. Burnham, "Native Beijing listeners' perceptual assimilation of Mandarin lexical tones produced by L2-Mandarin speakers from Yantai, Shanghai, and Guangzhou," in *Speech Prosody 2022*, ISCA, 2022, pp. 782–786. doi: 10.21437/SpeechProsody.2022-159.

[17] W. J. B. van Heuven, P. Mandera, E. Keuleers, and M. Brysbaert, "Subtlex-UK: A new and improved word frequency database for British English," *Q. J. Exp. Psychol.*, vol. 67, no. 6, pp. 1176–1190, 2014, doi: 10.1080/17470218.2013.850521.

[18] Y. Li, M. D. Tyler, D. Burnham, and C. Best, "L2-Mandarin regional accent variability facilitates Mandarin-naive English listeners' learning of Mandarin tone-words," in *Proceedings of the 20th International Congress of Phonetic Sciences, Prague, Czech Republic, 2023*, accepted.

[19] R Core Team, "R: The R Project for Statistical Computing." 2022. [Online]. Available: https://www.r-project.org/

[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *ArXiv E-Prints*, vol. arXiv:1406, 2014, doi: 10.18637/jss.v067.i01.

[21] U. Halekoh and S. Højsgaard, "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. 1–32, 2014.

[22] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd edition. SAGE Publications, Inc, 2018.

[23] R. V. Lenth, "Least-squares means: The R package lsmeans," *J. Stat. Softw.*, vol. 69, no. 1, Art. no. 1, 2016, doi: 10.18637/jss.v069.i01.

[24] C. B. Moore and A. Jongman, "Speaker normalization in the perception of Mandarin Chinese tones," *J. Acoust. Soc. Am.*, vol. 102, no. 3, pp. 1864–1877, 1997, doi: 10.1121/1.420092.

[25] J. Zhang and S. T. Mueller, "A note on ROC analysis and non-parametric estimate of sensitivity," *Psychometrika*, vol. 70, no. 1, pp. 203–212, 2005, doi: 10.1007/s11336-003-1119-8.

[26] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*, 2nd ed. Mahwah, N.J.: Lawrence Erlbaum Associates, 2005.