Effects of auditory, visual and gestural input on the perceptual learning of tones

Katelyn Eng, Beverly Hannah, Keith Leung, Yue Wang

Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada

kse3@sfu.ca, beverlyw@sfu.ca, kwl23@sfu.ca, yuew@sfu.ca

Abstract

Research has shown that audio-visual speech information facilitates second language (L2) speech learning, yet multiple input modalities including co-speech gestures show mixed results. While L2 learners may benefit from additional channels of input for processing challenging L2 sounds, multiple resources may also be inhibitory if learners experience excessive cognitive load. The present study examines the use of metaphoric hand gestures in training English perceivers to identify Mandarin tones. Native Mandarin speakers produced tonal stimuli with simultaneous hand gestures mimicking pitch contours in space. The English participants were trained to identify Mandarin tones in one of four modalities: audio-only (AO, speaker voice only), audio-visual (AV, speaker voice and face), audio-gesture (AG, speaker voice and hand gestures) and audio-visual-gesture (AVG). Results show significant improvements in tone identification from pre- to post-training tests across all four training groups, demonstrating that gestural as well as visual articulatory information may facilitate tone perception. However, further analyses with individual tones reveal some group differences. Most noticeably, the AVG group had a slower learning curve during training compared to the other trainee groups for Tone 4, the least accurately identified tone, indicating a negative effect of multiple input modalities on the perception of difficult L2 sounds. In contrast, for Tones 2 and 3, the AG group revealed slower learning effects compared to the AV group, presumably because of the similar gestural trajectories for these two tones, which made the gestural input less distinct. Overall, the results suggest a positive role of gestures in tone identification, one that may also be constrained by phonetic and cognitive demands.

Index Terms: auditory, visual and gestural speech perception, Mandarin tone, L2 speech learning

1. Introduction

Research on multimodal speech processing has indicated that integration of auditory and visual articulatory information can enhance native and non-native speech perception [1], [2]. Additionally, co-speech hand gestures have been shown to facilitate native speech perception [3]. However, research has been inconclusive to the amount of gain from gestures in L2 speech learning [4], [5], [6]. For example, while beat gestures can aid L2 learners in parsing words into syllables [7], they are not as effective in discriminating durational differences [6]. Indeed, when integrated effectively, simultaneous auditory, visual and gestural information may have a combinatory effect in aiding speech learning [7], [8]. Conversely, the addition of gestural input may also be inhibitory as learners may experience excessive cognitive load, especially when phonetic demands are high [5], [6]. This discrepancy in the role of gestures motivates the present research.

In this study, native speakers of Canadian English were trained to perceive Mandarin Chinese lexical tones with one of four input modalities: audio-only (AO), audio-visual (AV), audio-gestural (AG), and audio-visual-gestural (AVG). The gesture used here is the metaphorical gesture, which traces an imaginary tone as it changes in pitch along the dimension of time (duration) and height (pitch), as is commonly used in Chinese tone teaching environments. This type of gestures has been shown to help pitch learning in contexts such musical training [9]. Training follows previously established high variability perceptual training procedures which involve various phonetic and speaker voice contexts to expose trainees to a variety of exemplars of the non-native speech categories [10], [11]. Trainees' performance was assessed by a pre- and a post-training tone identification test along with three intersession tests during training. Comparisons of the training effects with AG vs. AO or AV groups would determine whether and to what extent gestural information is beneficial in tone perception. On the other hand, if the addition of gestural input caused information overload, we would expect AVG training to be less effective than AG or AV training.

2. Methods

2.1. Participants

Four native Mandarin-speaking instructors (2 male, 2 female) were recorded to provide the training stimuli. They were chosen because of their familiarity with training students and knowledge of the Mandarin tones. Two additional native Mandarin speakers (1 male, 1 female) produced the pre- and post-test stimuli.

The trainees were 57 native Canadian English young adults. They had no prior experience with any tonal languages and no extensive experience with music either (with fewer than five years of musical training [12]). They were randomly assigned to one of the training groups (AO, AV, AG, or AVG), with 16 in each group (8 male, 8 female), except for the AO group, which had 9 participants (2 male, 7 female).

2.2. Stimuli

The training word list contained 80 Mandarin monosyllabic real words (20 syllables x 4 tones, Tone 1: high-level pitch, Tone 2: mid-high-rising pitch, Tone 3: low-falling-rising pitch, Tone 4: high-falling pitch), which was derived from those used in [11] and [13]. Audio-visual recordings of these words were made twice for all speakers, with and then without hand gestures. The stimuli used for the pre- and post-tests were the 60 additional Mandarin monosyllabic real words (15 tone quadruplets) used in [11].

Training stimuli speakers were instructed to simultaneously speak a word and trace an acetate graph

10.21437/SpeechProsody.2014-169

representation [14] of the corresponding tone on the feedback screen of the digital camera with their right index finger so that the gestures would be standard across speakers. The video was then mirrored so that the tone contour was presented in the correct direction for the trainees. For the AG condition, the speaker face was blacked out so that the only area visible was the arm including the hand and finger during tone tracing. Six additional native Mandarin speakers then participated in a stimuli goodness evaluation task. All the stimuli used in training were correctly identified and rated as good tokens of the Mandarin syllables by the native Mandarin speakers. Figure 1 displays sample images of (a) tone contour tracing used as gestural input in training, and (b-e) the four training conditions (AO, AV, AG, and AVG).

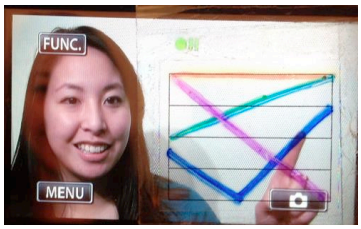

Figure 1-a: *Sample image of a speaker tracing the tone contour (pitch) on an acetate graph attached to the camera feedback screen.*

Figure 1-b: *Audio-only (AO) training presentation. No visual information is given in this condition.*

Figure 1-c: *Audio-visual (AV) training presentation with speaker face and voice.*

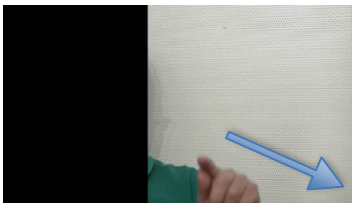

Figure 1-d: *Audio-gestural (AG) training presentation. The arrow indicates where the speaker will trace the contour of Tone 4.*

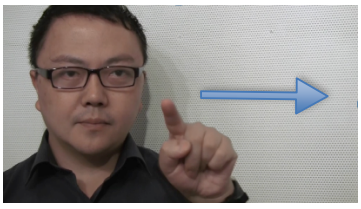

Figure 1-e: *Audio-visual-gestural (AVG) training presentation. The arrow indicates where the speaker will trace the contour of Tone 1.*

2.3. Procedures

Trainees were set up in a sound-treated booth in the Language and Brain Lab (Simon Fraser University, SFU) wearing AKG-brand circumaural headphones to hear the stimuli. They were first familiarized with the Mandarin tones by listening to a tone quadruplet and learned to associate each tone with its tonal label. They repeated this familiarization at the beginning of subsequent training sessions to review the task and hear an example before starting.

Prior to and after training, participants were tested with auditorily presented tone words described above. The identical pre- and post-tests employed a four-alternative forced choice task with no feedback provided. Identification was made using corresponding keys on the keyboard, with the labels "LEVEL" (for Tone 1), "RISING" (for Tone 2), "DIPPING" (for Tone 3) and "FALLING" (for Tone 4). Trainees were familiarized with these terms before the pre-test. After the pre-test, participant scores were calculated to determine if their percent-correct score was suitable for our inclusion criteria: If the participant scored between 25-80% correct, they were permitted to continue with the training. If the participants' scores fell outside that range, they were excluded and did not continue training. This was done to exclude those who had extreme scores due to hitting floor or ceiling.

Training took place during a two-week period with six sessions of 40 minutes each. Each session contained 40 words balanced across tones and syllables and produced by four speakers. Training stimuli were presented in AO, AV, AG or AVG, depending on the condition (as shown in Figure 1, b-e). The video was presented on a 12" (H) x 15" (W) display computer monitor, with the trainee's face roughly 18" away from the monitor screen.

Each trial started with presentation of a stimulus, followed by the trainee's task to identify the tone, and end with the feedback along with stimulus replay. Three intersession tests were administered after every second training session. These tests were used to track learning trajectory during training. They were presented in audio-only format, similar to the pre/posttest. The 80 intersession test stimuli (5 syllables × 4 tones × 4 speakers) employed selected stimuli used in training.

3. Results

3.1. Overall performance

The participants' percent correct identification scores were first analyzed using a three-way repeated measures ANOVA with Training Group (AO, AV, AG, AVG) as the between-subject factor, and Test (pre-test, intersession test 1 [Int1], intersession test 2 [Int2], intersession test 3 [Int3], post-test) and Tone (Tone1, Tone2, Tone3, Tone4) as the within-subject factors.

Significant main effects of Test [$F(4,212)=198.3$, $p<.001$] and Tone [$F(3,159)=31.6$, $p<.001$] were found. Bonferroni adjusted post hoc pairwise comparisons among tests reveal that across groups and tones, the mean pre-test score (48%) was significantly lower than all the intersession and post-test scores (Int1: 79%, Int2: 85%, Int3: 87%, post-test: 83%, $ps \leq .001$), and additionally, Int1 score was lower than Int2 and 3 scores ($ps \leq .001$), indicating significant

improvements with training. Post hoc tone comparisons show that, across tests and groups, perception of Tone 4 was significantly less accurate (68%) than Tones 1 (76%) and 2 (75%), which were in turn less accurate than Tone 3 (87%) ($ps \leq .038$), revealing Tone 4 as the most challenging tone.

The ANOVA also yielded significant interactions of Test x Tone [$F(12,636)=17.6$, $p<.001$] and Test x Tone x Group [$F(36,636)=1.6$, $p=.017$].

3.2. Individual tones and groups

Further analyses were performed based on the above interactions to identify possible Group differences as a function of Test and Tone. These involve sets of two-way Test x Group ANOVAs for each Tone, followed by further one-way ANOVAs for each Tone and Group with Test as a factor. Group comparisons of the five Tests for each Tone are displayed in Figure 2 (a-d).

First, only for Tone 4 (Figure 2-d) did the two-way ANOVAs show a significant interaction of Test and Group [$F(3,53)=4.0$, $p=.012$], along with a significant main effect of Test [$F(1,53)=337.4$, $p<.001$]. Bonferroni adjusted pairwise comparisons among the five tests for Tone 4 are consistent with the overall patterns. Across groups, the pre-test score (25%) was lower than all other test scores (Int1: 72%, Int2: 81%, Int3: 82%, posttest: 80%, $ps<.001$ for all). Furthermore, performance at Int1 was significantly poorer than that at Int2 and Int3 ($ps \leq .001$). Subsequent one-way ANOVAs with individual groups revealed that the aforementioned differences among intersession tests only occurred with the AVG group, with Int1 (67%) scoring marginally less well than in Int2 (81%, $p=.050$) and significantly lower than Int3 (86%, $p=.005$) and post-test (83%, $p=.007$), indicating a slower learning curve (during training) and a lack of generalization (to new stimuli at posttest) when training involved all three input modalities.

Consistently, for Tones 1-3, the two-way Test x Group ANOVAs also revealed significant main effects of Test. Though the ANOVAs yielded no significant interactions of Test and Group for these tones, the multiple levels of variables may obscure any possible difference in Test scores for each Tone and Group. Therefore, one-way ANOVAs were still run for each Group and Tone.

For Tone 1 (Figure 2-a), the two-way ANOVA yielded a significant effect of Test [$F(1,53)=125.5$, $p<.001$], with the scores being significantly lower at pre-test (45%) than at all the other tests which scored equally high (Int1: 81%, Int2: 85%, Int3: 85%; post-test: 82%, $ps<.001$ for all). Further one-way ANOVAs for each Group did not reveal any different patterns either. Thus all four training groups improved equally and retained the level of performance observed at Int1.

For Tone 2 (Figure 2-b), there was also a significant effect of Test [$F(1,53)=102.3$, $p<.001$], with pre-test (52%) being significantly lower than the other tests across groups (Int1: 75%, Int2: 83%, Int3: 86%, post-test: 76%, $ps<.001$ for all). Additionally, Int1 score was significantly lower than Int2 and Int3 scores ($p \leq .014$), both of which scored higher than post-test ($ps \leq .005$). Subsequent analyses with each trainee group revealed some group-specific patterns. First, the overall result of a higher-intersession-than-post-test performance was only exhibited in AO and AG groups, whose Int3 (AO=92%, AG=88%) score was higher than the post-test (AO=72%, AG=77%, $ps \leq .008$). Moreover, the overall gradual learning pattern was particularly true for AG, whose Int1 (75%) score did not differ significantly from the pre-test (51%, $p=.163$). Finally, only for AO, the pre- (50%) and post-test (72%) scores did not show significant improvement ($p=.074$). These results show different learning trajectories across groups for Tone 2.

For Tone 3 (Figure 2-c), a significant effect of Test was again found [$F(1,53)= 61.0$, $p<.001$], with the pre-test (69%) being lower than the subsequent tests across groups (Int1: 89%, Int2: 90%, Int3: 92%, posttest: 93%, $ps<.001$ for all). Subsequent individual group analyses revealed that only the AG group showed exceptions to this general pattern, in that their pre-test score (65%) was not significantly different from their Int1 (82%, $p=.068$) and Int2 (81%, $p=.138$) scores, showing a slower learning effect.

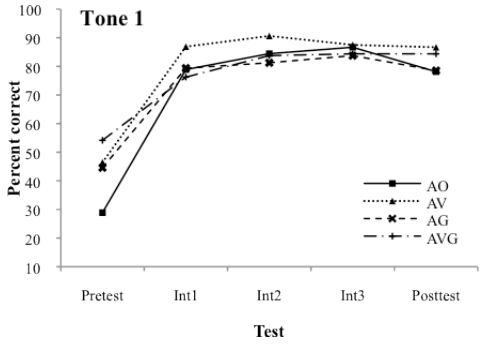

Figure 2-a: *Percent correct identification scores for each group in five tests for Tone 1.*

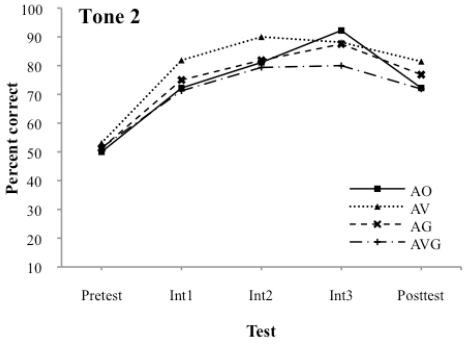

Figure 2-b: *Percent correct identification scores for each group in five tests for Tone 2.*

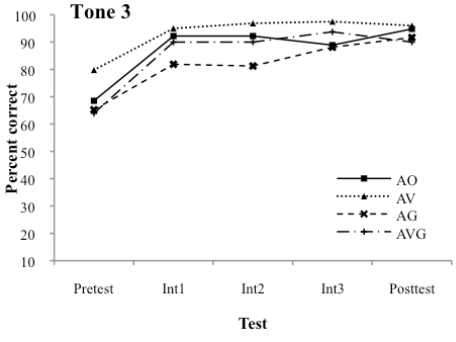

Figure 2-c: *Percent correct identification scores for each group in five tests for Tone 3.*

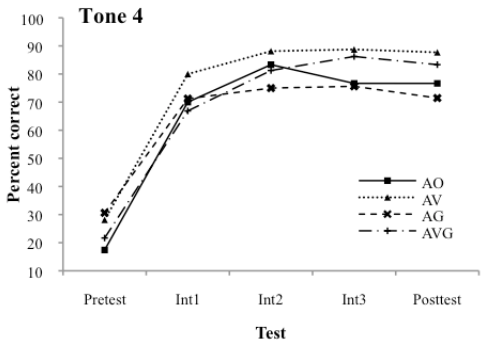

Figure 2-d: *Percent correct identification scores for each group in five tests for Tone 4.*

3.3. Summary

The results demonstrate that the performance for all four groups improved through training. The perception for each tone was slightly different, with Tone 4 being the most difficult tone to identify. Moreover, the results for individual tone and group showed robust differences in Tone 4 learning, where only the AVG group exhibited a delay in improvement during training and a failure in generalization at post-test. Similar gradual learning patterns were also observed for Tone 2 and Tone 3 with the AG group. Furthermore, for Tone 2, the AO group did not improve significantly from the pre-test to the post-test.

4. Discussion and concluding remarks

Overall, the results showed substantial improvement after training for all groups, which is in keeping with previous research indicating that speech perception from multimodal presentation benefits L2 speech learning [2], [7], [8], [15].

However, further analysis with individual tones did reveal differences in the extent of improvement among the four training groups. Most noticeably, compared to the other three trainee groups, the AVG group exhibited a slower learning curve during training and a lack of generalization after training for Tone 4 (with high-falling pitch), also shown as the most difficult tone to identify. It is possible that the cognitive resources for the identification of Tone 4, the most challenging tone, were overtaxed by too many channels of input, which in turn resulted in lower performance of the AVG group compared to the other trainees. This supports the "information overload" hypothesis, being in line with previous claims that gestural input in addition to auditory and visual input may increase the cognitive load, resulting in an inhibitory effect in learning [6], particularly when phonetic demands are high [5].

Results for the learning trajectory of Tone 2 (with rising pitch) showed an increased identification from the pre- to post-training tests for AV and AG but not for AO, suggesting that the addition of either visual or gestural input could result in a significant improvement in tone identification. This is in line with the results for Tone 4, where similar levels of improvement were found in both AV and AG groups. The positive effects of visual input support the previous findings that visual articulatory movements involving the head, mouth, and eyebrows provide robust correlates to tone perception [16], [17], [18]. Unlike visual speech information present in a speaker's face, which has anticipated and fixed articulatory configurations for the resultant speech sounds, the gestures used in the current study involve spatial changes which are not directly bound to speech. However, the AG group's improvements demonstrate that the trainees were able to make the spatial-auditory association embodied by the hand gestures tracing pitch trajectories and utilize that to aid their learning, just as how similar gestures could guide musical pitch processing [9] and aid L2 prosodic perception [7].

Nonetheless, it is also worth noting that visual and gestural modalities are not always equal in terms of degree of improvement. For both Tone 2 and Tone 3, the increase in tone identification accuracy during training for the AG group was more gradual than that for AV. One possible reason for the slower improvement of the AG group may be due to the similar gestural trajectories of the two tones (both involving a long rising pitch contour), which may have made the two gestures less distinct. Thus it may have taken the AG group longer to effectively integrate the gestural information as compared to the AV group whose input involved more predictable articulatory-auditory correspondence.

Taken together, the differing patterns observed within these results demonstrate the complex role of multimodal input in L2 speech perception. The most promising finding is that co-speech gestural as well as visual articulatory information can aid L2 speech learning. However, facilitation from multiple input domains may not be additive and may be constrained by phonetic and cognitive demands. When perceiving phonetically challenging sounds, learners (particularly those at the elementary level) may find multiple input resources distracting, as they may not be able to simultaneously focus on all these input domains and effectively integrate them into a single percept. As such, training with fewer input domains may reduce the cognitive load required to attend to multiple channels [6], [19]. Moreover, the different patterns for individual tones and training groups suggest that multimodal facilitative effects may take place in a complimentary manner. Learners could be trained to selectively focus on those domains that can most readily and reliably aid their learning. Further research may delve into these avenues to better understand the complex relationship between L2 prosody learning and gestures.

5. Acknowledgements

All authors made equal contributions to this study. We thank Drs. Yukari Hirata and Spencer Kelly (Colgate University) and Drs. Allard Jongman and Joan Sereno (University of Kansas) for their valuable comments. We also thank Anthony Chor, Mathieu Dovan, Courtney Lawrence, and Lindsay Leong (SFU) for their assistance in data collection and analysis. This research has been funded by research grants from SFU Vice President Academic and the Social Sciences and Humanities Research Council of Canada to YW.

6. References

[1] Davis, C., & Kim, J., "Audio-visual interactions with intact clearly audible speech," Q. J. Exp. Psychol. 57A, 1103–1121, 2004.

[2] Hazan, V., Sennema, A., Faulkner, A., & Ortega-Llebaria, M., "The use of visual cues in the perception of non-native consonant contrasts," J. Acoust. Soc. Am. 119, 1740–1751, 2006.

[3] Krahmer, E. & Swerts, M., "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," J. Memory Lang. 57, 396–414, 2007.

[4] Kelly, S. D., Manning, S. M., & Rodak, S., “Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education,” Lang. Linguistics Compass 2, 569-588, 2004.
[5] Kelly, S., & Lee, A., “When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high,” Lang. Cognitive Proc. 2, 793-807, 2012.
[6] Hirata, Y., & Kelly, S. D., “Effects of lips and hands on auditory learning of second-language speech sounds,” J. Speech Hear. Res. 53, 298-310, 2010.
[7] McCafferty, S., “Gesture and the materialization of second language prosody,” IRAL 44, 197–209, 2006.
[8] Kelly, S. D., McDevitt, T., & Esch, M., “Brief training with co-speech gesture lends a hand to word learning in a foreign language,” Lang. Cognitive Proc. 24, 313-334, 2009.
[9] Connell, L., Cai, Z. G., & Holler, J., “Do you see what I'm singing? Visuospatial movement biases pitch perception,” Brain Cognition 81, 124-130, 2013.
[10] Lively, S. E., Logan, J. S., and Pisoni, D. B. ~1993!. ‘‘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. 94, 1242–1255
[11] Wang, Y., Spence, M., Jongman, A., & Sereno, J.A., “Training American listeners to perceive mandarin tones,” J. Acoust. Soc. Am. 106, 3649-3658, 1999.
[12] Cooper, A. & Wang, Y., “The influence of linguistic and musical experience on Cantonese word learning,” J. Acoust. Soc. Am. 131, 4756-69, 2012.
[13] Liu, S. & Samuel, A. G., “Perception of mandarin lexical tones when F0 information is neutralized,” Lang. Speech 47, 109-138, 2004.
[14] Chao, Y. R., “Mandarin Primer: An Intensive Course in Spoken Chinese,” Harvard University Press, 1948.
[15] Wang, Y., Behne, D.M., & Jiang, H. “Linguistic experience and audio-visual perception of non-native fricatives,” J. Acoust. Soc. Am. 124, 1716- 1726.
[16] Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., Ciocca, V., Morris, R. H., Hill, H.,Vignali, G., Bollwerk, S., Tam, H., & Jones, C., “The perception and production of phones and tones: The role of rigid and non-rigid face and head motion,” Proc. 7th Int’l Seminar on Speech Production, pp. 185–192, 2006.
[17] Chen, T.H. & Massaro, D.W., “Seeing pitch: Visual information for lexical tones of Mandarin-Chinese,” J. Acoust. Soc. Am. 123, 2356-2366, 2008.
[18] Mixdorff, H., Hu, Y., & Burnham, D., “Visual cues in Mandarin tone perception,” Proc. InterSpeech, pp. 405–408, 2005.
[19] Hardison, D. A., “Acquisition of second-language speech: Effects of visual cues, context, and talker variability,” Appl. Psycholing. 24, 495-522, 2003.