Cross-modal perceptual learning in learning a tonal language

Xin Wang (x.wang1@mq.edu.au)
Department of Linguistics, 16 University Ave
Sydney, 2113 Australia

Luan Li (liluan@psy.ecnu.edu.cn)
Department of Psychology, Shanghai, China

Bob McMurray (bob-mcmurray@uiowa.edu)
Department of Bain and Psychological Science, Iowa City, USA

Abstract

Limited evidence shows that visual input can facilitate learning novel sound-to-meaning mappings that are crucial to learning a second language. However, the mechanisms by which visual information influences auditory learning are still unclear. Here, we investigate to what extent visual input can lead to effective learning in another domain. We trained atonal speakers with Mandarin tones in 4 conditions: Auditory Only (AO) where only auditory tones were given as input; Animated Contour (AC) where moving visual pitch contours indicating the dynamic changes of tones were given in addition to auditory tones; Static Contour (SC) where static visual pitch contours were given in addition to auditory tones; Incongruent Contour (IC) where mismatched pitch contours were given in addition to auditory tones. The results show the advantage of AC and SC over AO in learning tonal categories and that IC inhibits learning, suggesting that extracting 'compatible' properties cross modalities benefits learning most.

Keywords: cross-modal learning; L2 tone learning; lexical tones; second language acquisition

Introduction

Learning novel speech sounds poses challenges for second (L2) language learners. It is well documented that second language learners exhibit difficulty perceiving and producing L2 phonological contrasts (Cutler et al., 2006; Iverson et al., 2003; Ota et al., 2009). For example, native speakers of Japanese show difficulty distinguishing English /r/-/l/ contrasts. When learning L2 phonology, indeed, L2 adults learn L2 orthography at the same time in most instructional contexts, which was reported to support the learning of L2 speech sounds (Escudero et al., 2008; Hayes-Harb et al., 2010). However, supra-segmental information (e.g., lexical tones) is usually not encoded in orthography (Wang, 2021). Therefore, the lack of orthographic representations of lexical tones in a tonal language can add additional challenges in learning L2 (e.g., Liu et al., 2011; Wang et al., 1999). Furthermore, learning lexical tones encounters confusion as supra-segmental information which conveys different meanings in a tonal language can be intonational in an atonal language. One example is Mandarin Chinese which utilizes four distinct tones to disambiguate lexical meanings. These 4 lexical tones correspond to four distinct pitch contours, which are typically represented with numerals as in the following example: *ma1* 'mother', *ma2* 'hemp', *ma3* 'horse', *ma4* 'scold' (Chao, 1968). Phonetically, Mandarin tones 1-4 can be described as high level [55], high rising [35], low falling rising [214], and high falling [51], respectively, with the lowest pitch level assigned a value of 1 and the highest level assigned a value of 5 in phonetic transcription, as in Figure 1 (Howie, 1976). In contrast, Tone 2 and Tone 4 are comparable to intonations in English, but these intonations are not lexical in an atonal language.

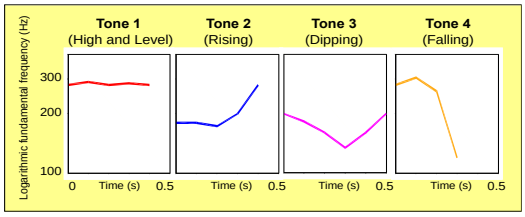

Figure 1: Mandarin tones /ma1/: mother, /ma2/: hemp, /ma3/: horse, /ma4/: scold

The majority of work done in L2 tone training and learning is under the framework of multisensory learning theory, namely, sound-to-meaning mappings benefit from information from different modalities, leading to a more distributed and robust representation (Eng et al., 2013; Godfroid et al., 2017; Liu et al, 2011). However, the results from different studies are not very consistent but somewhat contradictory. Recent work in L2 tone training appears to focus more on using hand gestures to enhance learning tonal categories (e.g., Baills et al., 2019; Morett & Chang, 2015; Morett et al., 2022; Zhen et al., 2019), which generates more consistent findings on the facilitation effect of gestures. However, it is unclear whether this effect is due to embodied cognition or cross-modal mapping. For example, Morett and Chang (2015) showed that performing and observing iconic hand gestures that mimic the pitch changes in lexical tones enhance word learning. They attributed this facilitation effect of gestures to embodiment such that language-specific features were highlighted. In contrast, Zhen et al (2019) systematically aligned the visual-motor features with the auditory tones (i.e., mappings between hand movements and

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

auditory pitch contours) and showed the facilitation effect of cross-modal alignment in learning L2 tones. Therefore, they claim that cross-modal learning is best when it is based on a common representational format of features across motor, visual and auditory domains. Note that these two different explanations and mechanisms do not contradict with each other, as Morett et al. (2022) propose that the mappings between pitch motions and positions (i.e., high pitches mapped to upward motions and positions, and low pitches mapped to downward motions and positions) could be based on embodied experience.

In light of the mechanisms of cross-modal learning articulated in Zhen et al (2019), it is crucial to understand whether non-embodied visual stimuli depicting pitch contours would provide the similar benefit in tone learning. The present study aims to investigate whether the cross-modal mappings between visual and auditory modalities can enhance tone learning, instead of using gestures. In addition, we present a novel tone training paradigm that can be easily adopted in L2 instructional contexts to facilitate lexical tone learning. To achieve this goal, we tested participants who had no tonal language background or little exposure to Mandarin with a tone training paradigm in which 4 different training conditions are compared: *Auditory Only* in which participants were trained with Mandarin tones in the auditory modality only as a baseline; *Static Contour* in which participants were trained with visual pitches in addition to auditory tones; *Animated Contour* in which participants were trained with animated pitches in addition to auditory tones; *Incongruent Animated Contour* in which participants were trained with incongruent visual pitches in addition to auditory tones. We hypothesize that auditory tones with additional visual support will benefit learning while the incongruent condition will 'destroy' the learning as between modality mappings are utilized during the learning.

Method

Participants

Three hundred and eleven undergraduate students at Macquarie University participated in the study for course credits (n=311). The study was approved by the Macquarie University Human Research Ethics Committee. Participation was voluntary and written consent was obtained from all participants. Data from 78 participants were excluded from the final analyses for the following reasons: 1) 36 participants reported that they could speak Chinese (Mandarin, Cantonese or Teochew) or other tonal languages including Thai and Vietnamese; 2) 3 participants reported having hearing loss or hearing issues; 3) 39 participants achieved over 45% accuracy in the pre-test. The final analyses were conducted based on data acquired from 233 participants.

Design

All experimental tasks were programmed and run using Gorilla and can be accessed online (https://gorilla.sc/). Participants were provided with a link and completed the experiment online on two consecutive days (approximately 20-30 minutes on Day 1 and 5 minutes on Day 2). Participants were randomly assigned to one of the 4 learning conditions: Auditory Only vs. Static Contour vs. Animated Contour vs. Incongruent Animated Contour. The numbers of participants included in the final analyses in each of the groups and conditions are: Auditory Only (n=53), Static Contour (n=62), Animated Contour (n=70) and Incongruent Contour (n=48).

Materials

Six simple vowels in Mandarin Chinese were used in this study (a e o i u ü). Each vowel was associated with 4 tones to create a total of 24 target stimuli. The auditory stimuli were recorded by a native speaker of Mandarin and were normalised. Participants were told that the stimuli were Chinese words. The 24 words were split into two sets of 12 trials each. Each participant was tested and trained on one set for tone learning, namely, 3 vowels. Additionally, twenty-four Mandarin diphthongs were selected and recorded with four tones each, to test for generalisation (ue, uo, ou, ie, ei, ao). Half of them were administered in the generalisation test immediately after the post-test, and the other half were used on the second day in a delayed generalisation test.

Procedure

The online experiment consisted of five stages: pre-test, learning (Block 1, 2, 3), post-test, generalisation test and delayed tests (including delayed tone identification and delayed generalisation test) as depicted in Figure 2. Randomly assigned to one of the 4 learning conditions, participants first completed surveys on their language background. They were then introduced to the general linguistic properties of Mandarin tones. Prior to the start of the experiment, participants were instructed to wear a headphone or earphone and adjust the volume so that they could hear the sounds clearly in a quiet room.

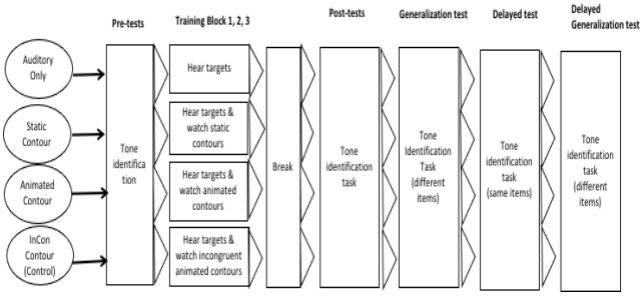

Figure 2: Flow chart depicting the learning procedure

During the pre-test, participants were given 12 auditory target stimuli in random order. They were asked to guess the

tone of the stimuli and press one of the four buttons on the screen, corresponding to Tones 1-4. This is a typical tone identification task, namely, a 4AFC task. No feedback was given. This was done to obtain a baseline of their pre-existing tonal knowledge. All participants were given the same task for the pre-test.

After the pre-test, the learning stage commenced. For the same set of 12 items, participants were trained in 3 blocks, each consisting of 24 trials with one repetition of the same item. Thus, a total of six repetitions of each item occurred during the learning blocks. Between blocks, participants were allowed to take a short break and continue when they were ready. During the learning stage, participants were trained in different conditions, as in a typical between-subject design. The visual pitches are visual lines depicting the pitch contours of 4 different Mandarin tones (see Figure 3). For example, in the Static Contour condition, participants were presented with visual pitches as in Figure 3, in addition to auditory tones. Note that in the Animated Contour condition, a video clip was presented to show the pitch movement for each tone indicating the direction of levelling (Tone 1), rising (Tone 2), dipping and rising (Tone 3) and falling (Tone 4) from left to right. The animation shows the movements of the pitch contours but also corresponds to the time duration of each auditory syllable/token. The task was tone identification, same as in the pre-test, but giving feedback. For each trial in all conditions, where both visual and auditory information was presented, the visual stimulus preceded the auditory stimulus by 500ms.

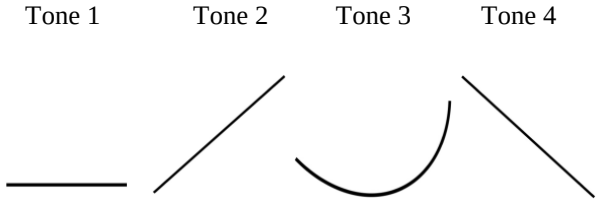

Figure 3: The visual pitches depicting Tone 1-4 in Mandarin

After the training blocks, participants completed a post-test, which is identical to the pre-test. The post-test assessed how much improvement participants made over the learning stage, by listening to the same items without any feedback or any visual information. All participants were tested with the same tone identification task.

Afterwards, a generalisation task was administered, where participants listened to 12 untrained diphthongs and performed on the same tone identification task. This purpose of this measure is to see whether the knowledge of tonal categories could be transferred to a different set of syllables that were absent in the training. Then, in 24 hours, they received an email reminder of the delayed tests. In the delayed tone identification task, participants listened to 24 auditory targets and pressed the buttons to indicate their tones. The 24 targets include 12 trained targets and 12 untrained/new targets.

Results

As the data are generated from remote online testing, we primarily rely on the accuracy as the main predictor to interpret our findings. The accuracy data was fit with Logit Mixed-effects models, using the glmer function of the lme4 package (Bates et al., 2015). We employed maximal random-effect structures in the models and included random slopes for factors of repeated measures to avoid Type I errors (Barr, et al., 2013). The fixed factors were learning condition ('condition' in the model) and learning stage ('display' in the model). We started from the maximal model which is justified by the design and them simplified it in case of convergence errors (Matuschek, et al., 2017). We performed a stepwise simplification on an overparameterized model by removing the correlation parameter, higher-order interactions, and random effect terms with least variance to address the convergence error (Singmann & Kellen, 2019). All post-hoc analyses, performed using the emmeans package, were corrected for multiple comparisons using the Holm_Bonferroni adjustment. Following standard conventions, any *p-value* smaller than .05 was deemed significant.

Table 1 presents the descriptive statistics for tone identification accuracy by learning condition and learning stage.

Table 1: Accuracy means (Standard Deviations in Parentheses) for Tone Identification by Learning Condition and Learning Stage (tests).

Condition	Pretest	Posttest	Gen	Delay_test	Delayed_gen
Animated Contour	0.23 (0.42)	0.63 (0.48)	0.58 (0.49)	0.6 (0.49)	0.61 (0.49)
Incongruent Contour	0.23 (0.42)	0.3 (0.46)	0.35 (0.48)	0.31 (0.46)	0.37 (0.48)
Auditory Only	0.23 (0.42)	0.49 (0.5)	0.47 (0.50)	0.45 (0.50)	0.44 (0.50)
Static Contour	0.24 (0.43)	0.56 (0.50)	0.57 (0.50)	0.59 (0.49)	0.58 (0.49)

For each learning stage and test, Figures 4-8 present the mean accuracy of each learning condition.

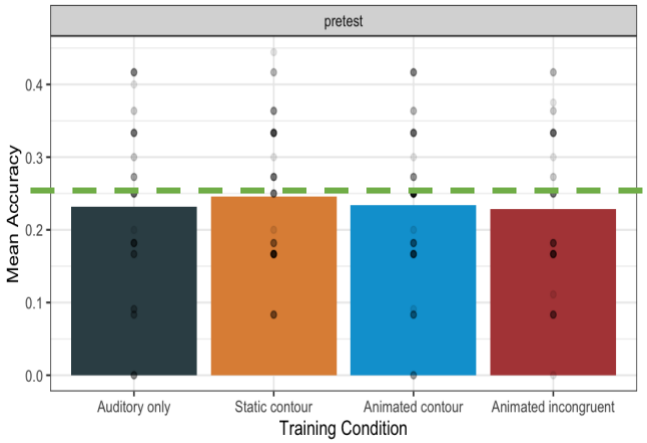

Figure 4: Accuracy per condition in the Pretest

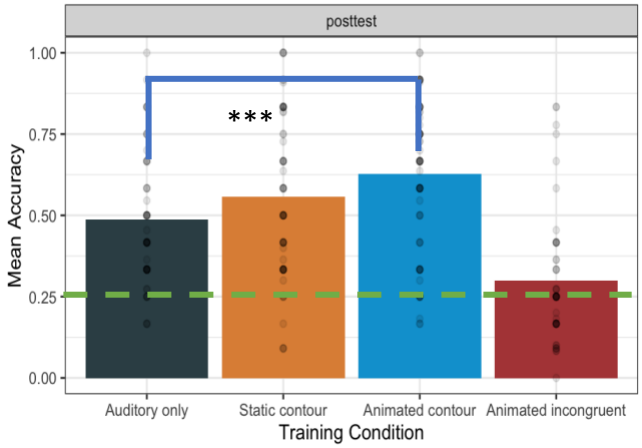

Figure 5: Accuracy per condition in the Posttest

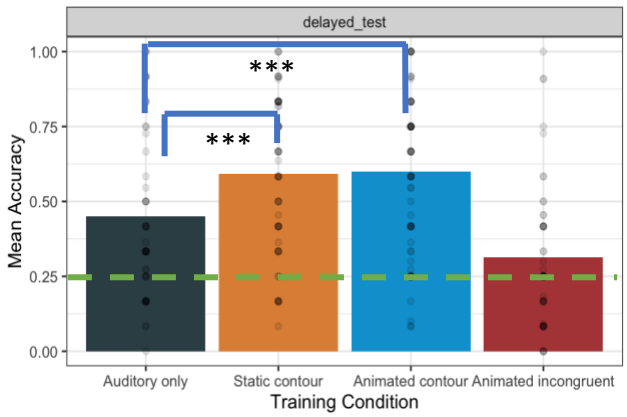

Figure 6: Accuracy per condition in the Delayed_test

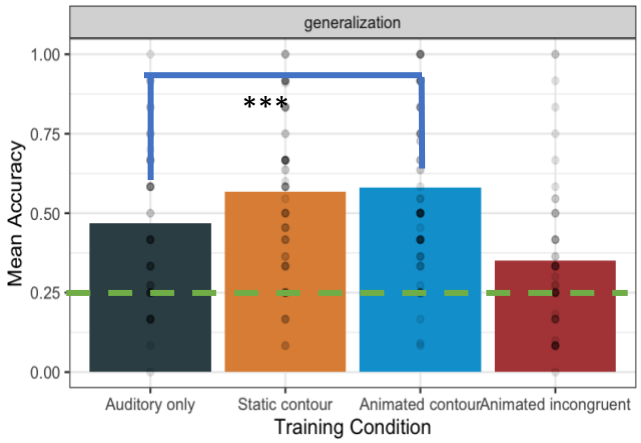

Figure 7: Accuracy per condition in the Generalization test

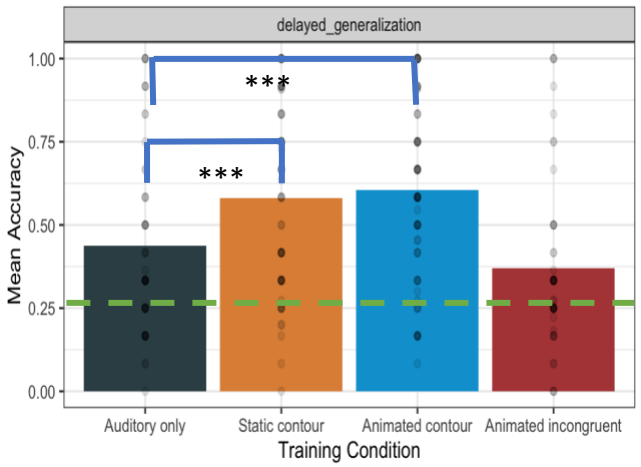

Figure 8: Accuracy per condition in the Delayed_generalization test

There are main effects of Learning condition, learning stage and interactions (all *p's* < .0001). There is no significant difference across conditions in the pre-test. In the post-test, Incongruent condition differs from all the other conditions (*p's* < .0001). Only Animated condition differs from Auditory condition significantly (*p = .0015).* In the delayed-test: Incongruent condition differs from all the other conditions (*p's* < .0001). Both Animated and Static conditions differ from Auditory Only (*p=.0004; p=.0017*). In the generalization test: Incongruent condition differs from all the other conditions (*p's < .0001)*. Only Animated Contours differs from Auditory Only (*p = .0379*). In the delayed generalization test: Incongruent condition differs from Animated and Static conditions (*p's* <.001), but not Auditory Only (*p=.9608*). Both Animated and Static conditions differ from Auditory Only (*p=.0002; p=.004*).

Taken together, these results show a clear contrast between the incongruent contour condition and the other learning

conditions, indicating incompatible visual cues appear to 'destroy' learning. Both animated and static contours show advantage in the immediate tests for both trained and untrained items, however, only the aminated condition was statistically significant compared to the auditory only condition. Both Animated and Static contours show advantage in the delayed tests for both trained and untrained items and differ from the Auditory only condition significantly. However, the difference between the Animated and Static contour conditions is non-significant.

Discussion

Our study introduces a new tone training paradigm to demonstrate healthy adults' ability to learn novel speech categories given 'compatible' visual cues. Importantly, our results show that arbitrary associations between novel visual stimuli and linguistic auditory stimuli do not benefit learning (i.e., the incongruent condition). Visual pitches that can be mapped to auditory tones in the spatial or/and temporal dimensions indeed facilitate learning (i.e., the animated and static contour conditions). In addition, our data also show individual variabilities in all the learning conditions.

Prior tone training studies have investigated the benefits of using visual cues in training auditory tones but generated mixed results (e.g., Liu et al., 2011; Godfroid et al., 2017). These mixed findings are hard to explain without systematic comparisons across different conditions as in the current study. Arguably, studies using hand gestures to mimic pitch movements present learners with familiar visual-motor information such that cross-modality mappings are easily extracted to benefit learning (e.g., Morett & Chang, 2015; Zhen et al., 2019). This might be the reason these studies show a clear benefit of using gestures in facilitating tone training.

The current study extends previous findings regarding the involvement of gestures in learning L2 tones and the possible mechanisms. It presents a novel tone training paradigm that allows us to compare different training conditions which can be easily adapted into instructional contexts. Our results show the benefit of cross-modal mappings in learning tones, but should be further replicated and extended to learning a different tonal language. Again, our results support the hypothesis proposed by Zhen et al (2019) that cross-modal learning is best when it is based on a common representational format of features across motor, visual, and auditory domains. In addition, our results are also consistent with Morett et al. (2022)'s argument that the vertical conceptual metaphor of pitch underlies effective lexical tone learning. Finally, our results have important pedagogical implications in that learners benefit from meaningful visual cues when learning novel speech sounds.

References

Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2018). Observing and Producing Pitch Gestures Facilitates the Learning of Mandarin Chinese Tones and Words. Studies in Second Language Acquisition, 41(1), 33-58. doi:10.1017/s0272263118000074

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. Journal of Statistical Software, 67(1). doi:10.18637/jss.v067.i01

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001.

Chao, Y. R. (1968). A grammar of spoken Chinese. Berkeley, CA: University of California Press.

Cutler A, Weber A and Otake T (2006) Asymmetric mapping from phonetic to lexical representa- tions in second-language listening. *Journal of Phonetics* 34: 269–84.

Eng, K., Hannah, B., Leong, L., & Wang, Y. (2013). Can co-speech hand gestures facilitate learning of non-native tones?

Escudero P, Hayes-Harb R, and Mitterer H (2008) Novel second-language words and asymmetric lexical access. *Journal of Phonetics* 36, 345–60.

Godfroid, A., Lin, C.-H., & Ryu, C. (2017). Hearing and Seeing Tone Through Color: An Efficacy Study of Web-Based, Multimodal Chinese Tone Perception Training. Language Learning, 67(4), 819-857. doi:10.1111/lang.12246

Hayes-Harb R, Nicol J and Barker J (2010) Learning the phonological forms of new words: Effects of orthographic and auditory input. *Language and Speech* 53: 367–81.

Howie, J. (1976). An acoustic study of Mandarin tones and vowels. Cambridge, UK: Cambridge University Press.

Iverson P, Kuhl PK, Akahane-Yamada R, Diesch E, Tohkura Y, Kettermann A and Siebert C (2003) A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87: B47–B57.

Liu Y, Wang M, Perfetti CA, Brubaker B, Wu S and MacWhinney B (2011) Learning a tonal language by attending to the tone: An in vivo experiment. *Language Learning* 61: 1119–41.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. Journal of Memory and Language. https://doi.org/10.1016/j.jml.2017.01.001

Morett, L. M., & Chang, L.-Y. (2014). Emphasising sound and meaning: pitch gestures enhance Mandarin lexical tone acquisition. Language, Cognition and Neuroscience, 30(3), 347-353. doi:10.1080/23273798.2014.923105

Morett, L. M., Feiler, J. B., & Getz, L. M. (2022). Elucidating the influences of embodiment and conceptual metaphor on lexical and non-speech tone learning. Cognition, 222, 105014.

Ota M, Hartsuiker RJ and Haywood SL (2009) The KEY to the ROCK: Near-homophony in non- native visual word recognition. *Cognition* 111: 263–69.

Singmann, Henrik, & Kellen, D. (2019). An Introduction to Mixed Models for ExperimentalPsychology. In New Methods in Cognitive Psychology. https://doi.org/10.4324/9780429318405

Wang Y, Spence MM, Jongman A and Sereno JA (1999) Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America* 106: 3649–58.

Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. Cognition, 187, 178-187. doi:10.1016/j.cognition.2019.03.004

Wang, X. (2021). Beyond segments. Journal of Second Language Studies, 4(2), 245-267. doi:10.1075/jsls.21011.wan