Mandarin Tone Identification in Musicians and Non-musicians: Effects of Modality and Speaking Style

Yueqiao Han, Martijn Goudbeek, Maria Mos and Marc Swerts

Tilburg Center for Cognition and Communication, Tilburg University, the Netherlands

y.han@uvt.nl

Abstract

A considerable amount of studies has shown that musical ability affects the success of second language learning. Extending the existing body of work, this study investigates the combined effects of musical ability, modality and speaking style for tone-naïve listeners in identifying Mandarin tones. In order to examine the added value of visual information and hyperarticulated speech, Mandarin tones were presented in two modalities (audio-only and audiovisual) and speaking styles (natural and teaching style) to listeners with or without musical experience. Results showed that musicians generally outperformed non-musicians, but that modality and speaking style both affected learning: both accuracy and response times were better in the audiovisual and teaching style conditions. In addition, the tones differed in learnability: the identification of tone 3 proved the easiest and all participants had more difficulty identifying tone 4. Nevertheless, musicians showed significantly greater accuracy in their identification of tones. These findings suggest that learning to perceive Mandarin tones benefits from musical expertise, visual information and hyperarticulated speaking style.

Index Terms: musical expertise, audiovisual modality, speaking style, musicians and non-musicians, Mandarin tone identification.

1. Introduction

Learning to perceive Mandarin tones is difficult for speakers of non-tonal languages. In contrast to European languages, which tend to rely completely on phonological distinctions between vowels and consonants to distinguish word meanings, tonal languages such as Mandarin Chinese use tones to distinguish meanings at the lexical level. Based on fundamental frequency (F0), pitch patterns and intrasegmental prosody, Mandarin Chinese has four main distinctive tones, numbered 1 to 4: tone 1: high level (5-5), tone 2: mid-rising (or mid-high-rising; 3-5), tone 3: low-dipping (also low-falling-rising or mid-falling-rising; 2-1-4), and tone 4: high-falling (5-1) [1]. Although tonal movement tends to co-vary with other acoustic variables, the consensus is that F0 (as the correlate of perceived pitch) is the dominant acoustic feature for Mandarin Chinese tones [2, 3]. Given the ubiquity of tonal languages and their increasing economic importance, identifying factors that promote efficient learning of Mandarin tones has attracted considerable scholarly attention (see, for example [4, 5]). In the current study, we focus on three factors which may contribute to Mandarin tone perception: musical ability (comparing musicians and non-musicians), modality (comparing audiovisual stimuli with audio-only stimuli) and speaking style (comparing a natural style with a teaching style).

Musical ability has been shown to be an important factor in many aspects of language learning. Neuropsychological as well as behavioral studies have revealed that musical expertise positively influences aspects of speech processing such as lexical pitch [6, 7, 8, 9], sentence intonation [10] and perceiving the metric structure of words [11]. Both the perception of native [12] and foreign language speech [13] have been reported to benefit from musical experience of the subjects [13, 14, 15, 16]. It is not surprising to see musical expertise facilitating speech perception, since music and speech bear several similarities [17, 18, 19]. For one thing, music and speech are complex auditory signals based on the shared acoustic parameters: both pitch and duration contribute to the melodic and rhythmic aspects of music and to the linguistic functions of speech [20]. In addition, music and speech processing both require attention, memory and sensorimotor abilities. Furthermore, it seems that processing music and language use closely related neurocognitive systems. Although the dominant view has been that language and music processing were located in different hemisphere of the brain (left for language and right for music), an increasing number of studies have found that there is a functional overlap in the brain networks that process acoustical features used in both speech and music [18, 19, 21, 22]. Musical training seems to drive adaptive plasticity in speech processing networks [23] and there is a music training transfer between music and speech [17].

Contrary to the bulk of European languages, pitch variations are linguistically relevant in tone languages (such as Mandarin Chinese) and determine the meanings of words. The unfamiliarity with tone in many Western speakers makes tone languages ideally suited to examine the influence of musical experience on language learning [10]. Previous studies have shown that musicians are more sensitive to subtle pitch variations in speech than non-musicians (e.g. [12, 24]). Results of empirical studies using behavioral methods clearly provide evidence that lexical tone perception benefits from musical expertise. A highly influential study by Gottfried and Riester [25] showed that tone-naïve English music majors identified the four Mandarin tones better than non-musicians, and that musicians were also better at producing the Mandarin tones compared to non-musicians. Furthermore, music majors performed better than non-musicians in pitch glide identification, and were more accurate in their identification of both intact and silent-center Mandarin syllable tones [26] (see also [6] for similar results). In another study, that used intact and acoustically modified (limiting the available pitch information) syllables of the four Mandarin tones produced by multiple speakers, Lee and Hung [27] assessed the difference in performance in Mandarin tone identification between English musicians (with 15 years of musical training on average, without absolute pitch abilities) and non-musicians. They found that musicians processed pitch contours better than non-musicians and concluded that (extensive) musical training facilitated lexical tone identification, although the extent to which musical ability facilitated tone perception varied as a function of the tone in question and the type of acoustic input. Taken together, these studies show that musicians consistently outperform non-musicians in the area of lexical tone processing.

Alongside the influence of musical ability of the listeners, the second factor we examined is modality, since speech perception can also be facilitated by providing visual information during articulation (e.g. [28, 29]). Visual speech information is provided by movements of the lips, the face, the head and the neck. In order to be understood well, speakers are assumed to strive to provide optimal acoustic and/or visual information to meet the demands of the target audience or the communicative situation [30]. For lexical tone perception, studies (e.g. [31, 32]) have shown that there is visual speech information that is related to lexical tone. When speakers want to convey information about tone (the pitch contour for instance), facial cues (along with gestures) are a common resource they resort to alongside the acoustic information. Our mouths and faces need to move in a certain way to produce a given tone which has consequences for the amplitude and the length of the visible articulations [33]. For instance, in Mandarin tones, vowel duration tends to be the longest for tone 3 and shortest for tone 4; amplitude tends to be lowest for tone 3 and highest for tone 4 [2]. These acoustic differences may have visual correlates, for instance in the amplitude and the length of the visible articulations [34].

However, the amount of audio-visual benefit achieved (i.e., the superiority of bimodal performance compared to unimodal performance) differs widely across individuals [35]. The added value of facial expressions for tone perception depends strongly on context, in particular on the availability of a clear and reliable acoustic signal. In situations where such a signal is available, extra visual / facial information may actually distract the listeners for the tone perception, since listeners are reluctant to use the visual information when acoustic sources are available and reliable [31].

In our study, we look into the effects of modality for musicians and non-musicians: more specifically, we presented our participants with audiovisual stimuli or stimuli that contained only audio. Through extensive musical training, musicians are particularly sensitive to the acoustic structure of sounds (i.e., frequency, duration, intensity and timbre parameters). This sensitivity has been shown to influence their perception of pitch contours in spoken language [12], but the extent to which musicians are affected by the presence of (exaggerated) visual information during speech perception has remained largely unexplored. While musicians might just benefit from the additional information like non-musicians, this is not a given. Given their extensive training to analyze the acoustic signal, they might not be as inclined to use visual cues (compared to non-musicians). Thus, they may benefit less from the added visual information. Taking into account that musicians may have developed increased abilities to focus attention on sounds and this ability may in turn help them to categorize the sounds and to make the relevant decision [17], we hypothesize that the added visual information may still benefit the Mandarin tone identification for musicians, while the contribution is likely smaller than that for non-musicians.

The third and final factor we address is that of speaking style, specifically the difference between speaking naturally and speaking in a teaching mode. Speakers show sensitivity to the characteristics of the audience they are addressing [30]. When they are addressing non-native speakers, like the participants in this study, native speakers/teachers commonly apply a so called teaching style, for instance, by speaking more slowly, more loudly and more clearly, in order to make the acoustic difference among speech units such as tones more salient for the listeners. Thus, it stands to reason that speakers' facial displays are also more exaggerated in teaching style compared to a more natural style of articulation.

When speakers employ a "teaching style", specifically geared to non-native listeners, or a more natural speaking style, geared towards fellow native speakers, the amount of auditory and visual information provided will be more pronounced in the case of a teaching style. Since speakers spend more energy to exaggerate the tone information in the teaching style, we expect more facial movements to be generated in the accompanying articulatory process, which creates potentially different information for the listeners, with musicians responding differently to a difference in speaking style than non-musicians, possibly moderated by modality as well. As a baseline hypothesis however, we start out from the fact that in teaching style speakers provide more acoustic and visual information to the listeners, which enhance their tone identification.

In sum, in this paper we investigate the effects of musical ability, modality and speaking style on Mandarin tone identification by tone-naive listeners (speakers of Dutch). Since we expect that the effects of our three independent variables will vary among tones, we individually assess the effects for each tone in our study.

2. Method

A 2 (musical ability) x 2 (modality) x 2 (speaking style) design was employed in this study. Two groups of participants (musicians and non-musicians) were divided over two modality conditions (audiovisual vs. auditory-only) and two speaking styles (natural vs. teaching). Both the accuracy (whether a response was correct or not) and the reaction time (how long a participant took to respond) for each stimulus were recorded as dependent variables.

2.1 Subjects

86 (mean age 22, 62 females) non-musicians were recruited from the Tilburg University participant pool; 84 (mean age 22, 35 females) musicians were recruited from Fontys School of Fine and Performing Arts. 83% of the participants were native speaker of Dutch, and none of them had been previously exposed to tone languages. The musician group had eight or more years of continuous music education and training up until or beyond the year 2017, while all the non-musicians had no more than three years of continuous music training. Assignment to the musical or non-musical group was based on participants' answers to a detailed music-experience questionnaire, the Goldsmith Musical Sophistication Index [36].

2.2 Material

2.1.1. Stimulus construction

A word list with 10 Mandarin monosyllables (e.g., ma, ying …) was constructed (selection based on stimulus material from [3, 37]; see the Appendix for the complete list). Each of these syllables was chosen such that the four tones would generate four different meanings resulting in 40 (10 syllables × 4 tones) different existing words in Mandarin Chinese.

2.1.2. Material recording

4 native Mandarin Chinese speakers were instructed to produce the 40 words in two different scenarios in sequence: a natural mode ("pronounce these words as if you were talking to a Chinese speaker") and a teaching mode ("as if you were talking to someone who is not a Chinese speaker"). In both conditions, there were no other instructions or constraints imposed on the way they should produce the stimuli. There was a 20-minute break for the speakers between the two recordings to avoid fatigue, with the recording of the natural stimuli preceding the recording of the teaching style stimuli.

We used Eye-catcher (version 3.5.1) and Windows Movie Maker (2012) to record the speakers' images and sounds. One of the advantages of the Eye-catcher system is that the camera is located behind the computer screen, which is convenient for unobtrusively capturing the full-frontal images of speakers' faces, similar to what listeners see in a face-to-face setting.

In total, 320 stimuli were produced; two sets of 160 video stimuli (10 syllables × 4 tones × 4 speakers), in teaching and in natural modes. These video clips were segmented into individual tokens, with each token containing one stimulus. Format Factory (version 3.9.5) was used to extract the sound from each video to generate stimuli for the audio-only conditions. This resulted in 4 types of experimental stimuli: video + teaching (VT); video + natural (VN); audio + teaching (AT); audio + natural (AN).

2.1.3. Stimuli validation

In order to validate the stimuli, 24 native Mandarin speakers were asked to identify the tones which were presented in the audio + natural condition (the supposedly most challenging condition), and their accuracy was 99.5%, indicating all stimuli were very easy to identify for native speakers.

Acoustic and visual analyses revealed differences between the two speaking styles (teaching and natural) and between the two modalities (auditory-only and audiovisual), e.g. stimuli were prolonged in teaching style compared to natural style (for a detailed acoustic and visual analysis of the stimulus material, see [34]).

2.3 Procedure

The task of the participants was to identify the tones they perceived from the audiovisual or audio-only stimuli. Participants were tested individually in a sound-attenuated booth. They wore headsets and were seated directly in front of the PC running the experiment. Three practice trials were included to allow participants to get familiar with the testing procedure and the stimuli. After the practice trials, the experiment leader checked with the participants to make sure they fully understood the concept of tones (in particular the symbol used for each tone (-, /, ᵛ, \) and the task. Finally, 160 testing stimuli were presented in randomized order for each participant (by E-Prime) in each condition (VT, VN, AT, and AN). All stimuli were presented at a comfortable hearing level. Participants were instructed to press the designated keys with the corresponding tone symbols as accurately and as quickly as possible after they had made their decisions. The responses and reaction times were recorded automatically using E-prime.

3. Results

In order to examine to what extent modality (audio-visual vs. audio-only), speaking style (natural vs. teaching mode) and the participants' musical ability affect the perception of Mandarin Chinese tones, accuracy (whether a response was correct or not) and reaction time (how long a participant took to respond) for each stimulus were analyzed. For each dependent variable, a mixed ANOVA was carried out with modality, speaking style and musical ability as between-subject factors, and speaker and tone as within-subject factors.

Figure 1 depicts the performances of musicians and non-musicians in the four experimental conditions. Overall, all the participants are able to identify Mandarin tones well above chance levels (25%), and the musician group outperforms the non-musician group in all experimental conditions, as indicated by a higher percentage of correct responses (M = 75%, SE = .02 vs. M = 48%, SE = .02) and faster reaction times for the musicians (M = 759 ms, SE = 42 vs. M = 792, SE = 41), which is in line with our hypothesis that musical ability positively affects the ability to identify Mandarin tones. However, only the difference in percentage correct between musicians and non-musicians was statistically significant (F (1, 162) = 157, p < .001, η_p^2 = .49), while the differences in reaction times were not (F (1, 162) = 0.31, p = .58, η_p^2 = .002). Moreover, musicians performed almost equally well within all four experimental conditions. Non-musicians performed similarly in VN, AT and AN (around 45%); but their accuracy increased markedly in the VT condition to 57%.

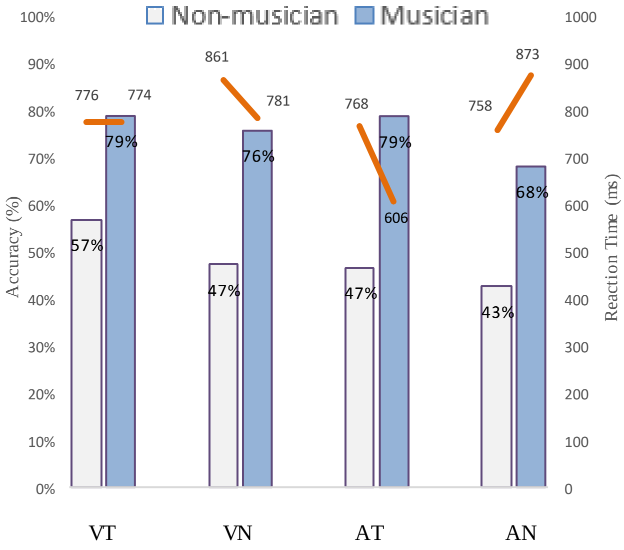

Figure 1: *Average accuracy and reaction time of Mandarin tone identification as a function of musical background, modality and speaking style.*

The statistical analyses further show that the audio-visual condition (M = 65%, SE = .02) yielded significantly higher accuracy scores than the audio-only condition (M = 59%, SE = .02); F (1, 162) = 6.82, p = .01, η_p^2 = .04. These results are in line with the hypothesis that the availability of visual cues along with auditory information is useful for people who have no previous knowledge of Mandarin Chinese tones when they need to learn to identify these tones. These differences between the two modalities were not observed for reaction times where responses in the audiovisual condition (*M* = 798 ms, *SE* = 41) were slower than in the audio-only condition (*M* = 752 ms, *SE* = 41), but not markedly so (*F* (1, 162) = .64, *p* = .42, η_p^2 = .004).

As we expected, participants that asked to identify tones produced in teaching style were significantly better at tone identification (*M* = 65%, *SE* = .02) than participants who were exposed to a more natural speaking style (*M* = 58%, *SE* = .02); *F* (1, 162) = 10.24, *p* = .002, η_p^2 = .06. However, while the direction for the effect was in line with our hypothesis and participants indeed identified Mandarin tones faster when they were produced in teaching style (*M* = 731 ms, *SE* = 42) compared to when they were produced in a natural speaking style materials (*M* = 819 ms, *SE* = 41), this effect was not significant (*F* (1, 162) = 2.25, *p* = .136, η_p^2 = .01). Notably, there are no significant interactions among musicality, modality and speaking style, which implies that they are not affected by each other on the performance of the participants.

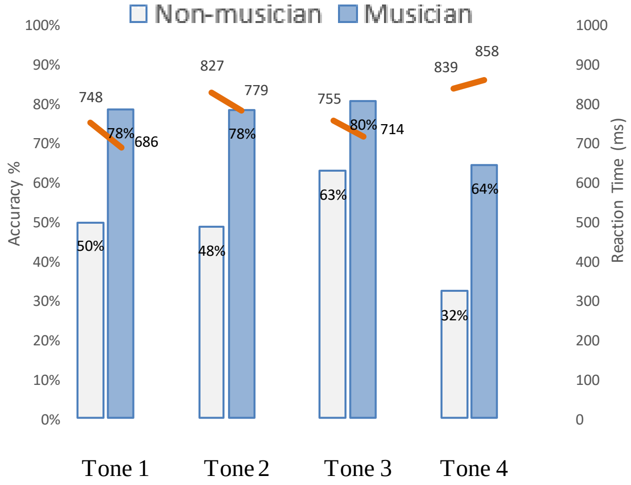

Figure 2: *Average accuracy and reaction time of Mandarin tone identification as a function of musical background and tone.*

Figure 2 shows the identification performance in terms of both accuracy and reaction times of musicians and non-musicians for each of the 4 Mandarin tones in the study. Unsurprisingly, and in line with all the previous findings we reported, musicians performed better than non-musicians for all the four Mandarin tones. For both listener types (and for all combinations of modality and speaking styles) tone had a strong effect on accuracy (*F* (3, 486) = 100.92, *p* < .001, η_p^2 = .38) and reaction time (*F* (3, 486) = 24.86, *p* < .001, η_p^2 = .133). A follow-up analysis with pairwise comparisons (with Bonferroni correction for multiple comparisons) shows that our participants in both groups generally performed best on tone 3: they gave more correct responses (*M* = 80%, *SE* = .02 for musicians; *M* = 63%, *SE* = .02 for non-musicians; *p* < .001), and were faster (*M* = 714 ms, *SE* = 40 for musicians; *M* = 754 ms, *SE* = 39 for non-musicians; *p* < .001) than when they had to identify the other tones. In contrast, tone 4 was the most challenging one for our participants to identify, both in terms of accuracy (*M* = 64%, *SE* = .02 for musicians, *M* = 32%, *SE* = .02 for non-musicians) as well as in terms of reaction time (*M* = 857 ms, *SE* = 50 for musicians; *M* = 837 ms, *SE* = 49 for non-musicians).

4. Discussion and Conclusions

In line with previous studies, we replicate the finding that musicians are at an advantage compared to non-musicians when learning to identify lexical tones in Mandarin Chinese, even if none of the participants are familiar with tonal languages [6, 8, 9, 25, 26, 27]. Based on our findings, it appears that years of musical training provide musicians with an increased sensitivity to pitch variation in lexical tone discrimination [cf 10, 26]: listeners with more musical training showed greater accuracy in their discrimination (75% vs. 48%). Importantly, although the musicians in our study performed well in the identification task (79% at the highest), they did not achieve native-like performance, even on a relatively simple task in the current experiment.

Furthermore, modality and speaking style affected tone identification. Tone-naïve listeners were better able to identify tones when they saw the speakers compared to when they only heard them, which supports the hypothesis that visual information plays a facilitating role in learning to identify Mandarin tones. The finding that musicians performed equally well in audio-only conditions as in audio-visual conditions (77% vs. 73%) indicates that for musicians the added visual information neither facilitates the speech perception nor distracts their attention. Speaking style also affected tone identification: Stimuli produced in a teaching style were identified better than those produced in a natural style. Salient acoustic information and exaggerated visual cues generated from the teaching style make tone identification easier for the listeners. However, there are no significant interactions between speaking style and musical ability. Thus, for musicians, teaching mode is not necessary superior to the natural style.

Crucially, individual tones are important contributors to differences in tone perception. In other words, it is much more important *which* tone the listeners hear than *how* they hear it. The low-dipping tone 3 is the easiest one to identify, while all listeners had more difficulty identifying the high-falling tone 4. These results indicate musical training facilitates lexical tone identification, although the facilitation varies as a function of tone and the type of acoustic input [27].

Overall, musical expertise positively influences the ability to learn to identify Mandarin tones. Learning Mandarin tones may be facilitated by being aware of the information provided by both the auditory and the visual modality, as well as by the potential benefit in clear and well-articulated speech as exemplified in teacher talk. Finally, it is clear that the individual tones differ in how easy they are to identify. We aim to investigate the contributions of these factors in future work and hope that our findings will benefit second language learners of Mandarin and will inspire further research on Mandarin tone learning.

5. Acknowledgements

This project is supported by China Scholarship Council (CSC). Our special thanks go to Fontys School of Fine and Performing Arts for their generous support in data collection.

6. References

[1] Chao, Y. R. (1948). *Mandarin Primer*. Cambridge, Mass: Harvard University Press.

[2] Tseng, C. Y., (1981). *An acoustic phonetic study on tones in Mandarin Chinese*. PhD dissertation. Brown University, Providence, RI.

[3] Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics, 36*(2), 268-294.

[4] Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics, 40*(2), 269-279.

[5] So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and speech, 53*(2), 273-293.

[6] Alexander, J. A., Wong, P. C., & Bradlow, A. R. (2005). Lexical tone perception in musicians and non-musicians. In *Ninth European Conference on Speech Communication and Technology*.

[7] Ong, J. H., Burnham, D., Escudero, P., & Stevens, C. J. (2017). Effect of Linguistic and Musical Experience on Distributional Learning of Nonnative Lexical Tones. *Journal of Speech, Language, and Hearing Research*, *60*(10), 2769-2780.

[8] Delogu, F., Lampis, G., & Belardinelli, M. O. (2006). Music-to-language transfer effect: May melodic ability improve learning of tonal languages by native nontonal speakers? *Cognitive Processing*, *7*(3), 203-207.

[9] Delogu, F., Lampis, G., & Belardinelli, M. O. (2010). From melody to lexical tone: Musical ability enhances specific aspects of foreign language perception. *European Journal of Cognitive Psychology*, *22*(1), 46-61.

[10] Marie, C.; Delogu, F.; Lampis, G.; Olivetti Belardinelli, M.; Besson, M. Influence of Musical Expertise on Segmental and Tonal Processing in Mandarin Chinese. J. Cogn. Neurosci. 2011, 23, 2701–2715.

[11] Marie, C., Magne, C., & Besson, M. (2011). Musicians and the metric structure of words. Journal of Cognitive Neuroscience, 23(2), 294-305.

[12] Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology*, *41*(3), 341-349.

[13] Marques, C., Moreno, S., Luís Castro, S., & Besson, M. (2007). Musicians detect pitch violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence. *Journal of Cognitive Neuroscience*, *19*(9), 1453-1463.

[14] Marie, C.; Delogu, F.; Lampis, G.; Olivetti Belardinelli, M.; Besson, M. Influence of Musical Expertise on Segmental and Tonal Processing in Mandarin Chinese. J. Cogn. Neurosci. 2011, 23, 2701–2715.

[15] Milovanov, R.; Huotilainen, M.; Välimäki, V.; Esquef, P.A.A.; Tervaniemi, M. Musical aptitude and second language pronunciation skills in school-aged children: Neural and behavioral evidence. Brain Res. 2008, 1194, 81–89.

[16] Milovanov, R.; Pietilä, P.; Tervaniemi, M.; Esquef, P.A.A. Foreign language pronunciation skills and musical aptitude: a study of Finnish adults with higher education. Learn. Individ. Diff. 2010, 20, 56–60.

[17] Besson, M., Chobert, J., & Marie, C. (2011). Transfer of training between music and speech: common processing, attention, and memory. *Frontiers in psychology*, *2*, 94.

[18] Patel, A.D. Music, Language, and the Brain; Oxford University Press: New York, NY, USA, 2008.

[19] Patel, A.D. Music, biological evolution, and the brain. In Emerging Disciplines; Bailar, M., Ed.; Rice University Press: Houston, TX, USA, 2010; pp. 91–144.

[20] Chobert, J., & Besson, M. (2013). Musical expertise and second language learning. *Brain Sciences*, *3*(2), 923-940.

[21] Wong, P. C., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature neuroscience*, *10*(4), 420.

[22] Besson, M., Schön, D., Moreno, S., Santos, A., & Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restorative neurology and neuroscience*, *25*(3-4), 399-410.

[23] Milovanov, R., & Tervaniemi, M. (2011). The interplay between musical and linguistic aptitudes: a review. *Frontiers in psychology*, *2*, 321.

[24] Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing research*, *219*(1-2), 36-47.

[25] Gottfried, T. L., & Riester, D. (2000). Relation of pitch glide perception and Mandarin tone identification. *Journal of the Acoustical Society of America, 108*(5), 2604.

[26] Gottfried, T. L., Staby, A. M., & Ziemer, C. J. (2004). Musical experience and Mandarin tone discrimination and imitation. *The Journal of the Acoustical Society of America*, *115*(5), 2545-2545.

[27] Lee, C. Y., & Hung, T. H. (2008). Identification of Mandarin tones by English-speaking musicians and nonmusicians. The Journal of the Acoustical Society of America, 124(5), 3235-3248.

[28] Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research, 53*(2), 298-310.

[29] Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661-699.

[30] Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*(5572), 1435-1435.

[31] Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001). *Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers*. In AVSP 2001-International Conference on Auditory-Visual Speech Processing, Aalborg, Denmark.

[32] Mixdorff, H., Hu, Y., & Burnham, D. (2005). *Visual cues in Mandarin tone perception.* In Proceedings of Eurospeech 2005 (InterSpeech-2005): Lisbon, Portugal, 405-408.

[33] Xu, Y., & Sun, X. (2002). "Maximum speed of pitch change and how it may relate to speech". *The Journal of the Acoustical Society of America, 111*(3), 1399-1413.

[34] Han, Y, Goudbeek, M, Mos, M, & Swerts, M (2018). Effects of Modality and Speaking Style on Mandarin Tone Identification by Non-native Listeners. *Phonetica*, under review.

[35] Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104*(4), 2438-2450.

[36] Müllensiefen, D., Gingras, B., Stewart, L., & Musil, J. (2014). The Goldsmiths musical sophistication index (Gold-MSI): Technical report and documentation v1.0. London: Goldsmiths, University of London.

[37] Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *The Journal of the Acoustical Society of America, 123*(4), 2356-2366.

7. Appendix

List of words used for producing the stimuli.

mā	má	mǎ	mà
yī	yí	yǐ	yì
xiē	xié	xiě	xiè
shē	shé	shě	shè
shī	shí	shǐ	shì
yōu	yóu	yǒu	yòu
fēn	fén	fěn	fèn
fū	fú	fǔ	fù
pō	pó	pǒ	pò
yīng	yíng	yǐng	yìng