frontiers in Communication

ORIGINAL RESEARCH
published: 22 November 2019
doi: 10.3389/fcomm.2019.00070

Mandarin Tone Identification by Tone-Naïve Musicians and Non-musicians in Auditory-Visual and Auditory-Only Conditions

Yueqiao Han, Martijn Goudbeek, Maria Mos and Marc Swerts*

Department of Communication and Cognition, Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg, Netherlands

A considerable number of studies have shown that musical ability has a positive effect on language processing. Extending this body of work, this study investigates the effects of musicality and modality on Mandarin tone identification in tone-naïve participants. To examine the effects of visual information in speech, Mandarin tones were presented in auditory-only or auditory-visual modalities to participants with or without musical experience. The Goldsmith Musicality Index was used to assess the musical aptitude of the participants. Overall, musicians outperformed non-musicians in the tone identification task in both auditory-visual and auditory-only conditions. Both groups identified tones more accurately in the auditory-visual condition than in the auditory-only condition. In addition, performance differed by tone: musicality holds its main effect on each level of tone; while the influence of modality differs for individual tones; the identification of tone 3 (a low-falling-rising) proved to be the easiest, while tone 4 (a high-falling tone) was the most difficult to identify for all participants. Out of all the musical skills measured by the Goldsmith Musicality Index, the amount of musical training was the only predictor that had an impact on the accuracy of Mandarin tone perception. These findings suggest that learning to perceive Mandarin tones benefits from musical expertise, and visual information can facilitate Mandarin tone identification, but mainly for tone-naïve non-musicians.

Keywords: Mandarin tone identification, audiovisual modality, musicians and non-musicians, the Goldsmith Musicality Index, musicality, visual information

OPEN ACCESS

Edited by:
Marc F. Joanisse,
University of Western Ontario, Canada

Reviewed by:
Sarah Creel,
University of California, San Diego, United States
Caicai Zhang,
Hong Kong Polytechnic University, Hong Kong

****Correspondence:***
Yueqiao Han
y.han@uvt.nl

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *31 May 2019*
Accepted: *04 November 2019*
Published: *22 November 2019*

Citation:
Han Y, Goudbeek M, Mos M and Swerts M (2019) Mandarin Tone Identification by Tone-Naïve Musicians and Non-musicians in Auditory-Visual and Auditory-Only Conditions. Front. Commun. 4:70. doi: 10.3389/fcomm.2019.00070

INTRODUCTION

More than half of the languages (60–70%) spoken in the world are so-called tone languages (Yip, 2002). Of these, Mandarin Chinese is spoken by the largest population by far (total users in all countries in 2015: 1,107,162,230[1]. Learning to identify Mandarin tones is difficult for speakers of non-tonal languages. Unlike most European languages, which rely primarily on phonological distinctions between consonants and vowels to distinguish word meanings, tone languages, such as

[1]https://www.ethnologue.com/language/cmn

Mandarin Chinese, additionally use tones to distinguish meanings at the lexical level. Marked by fundamental frequency (F0), pitch patterns and intrasegmental prosody, Mandarin Chinese has four main distinctive tones, conventionally numbered 1–4: tone 1: high-level (5-5[2]); 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) (Chao, 1948). Although tonal movement tends to correlate with other acoustic variables, the consensus is that F0 (as the correlate of perceived pitch) is the dominant acoustic feature for Mandarin Chinese tones (Tseng, 1981). Given the ubiquity of tonal languages and their increasing economic importance (Maddieson et al., 2013), identifying factors that promote efficient learning of Mandarin tones has attracted considerable scholarly attention (for example, So and Best, 2010; Hao, 2012). In the current study, we focus on two factors which may contribute to Mandarin tone perception: musical ability (comparing musicians and non-musicians), and modality (comparing auditory-visual stimuli with auditory-only stimuli).

Tone Perception and Musical Ability

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 (Alexander et al., 2005; Delogu et al., 2006, 2010; Ong et al., 2017), sentence intonation, and perceiving the metric structure of words (Marie et al., 2011). Both the perception of native (Schön et al., 2004) and foreign language speech (Marques et al., 2007) have been reported to benefit from musical experience (Milovanov et al., 2008, 2010; Marie et al., 2011). The current study aims to explore whether musical expertise also helps tone naïve listeners to correctly identify Mandarin Chinese tones.

It is not surprising that musical expertise facilitates speech perception, since music and speech are similar in several ways (Patel, 2010; Besson et al., 2011). For one thing, music and speech are both complex auditory signals based on similar acoustic parameters: both pitch and duration contribute to the melodic and rhythmic aspects of music and to the linguistic functions of speech (Chobert and Besson, 2013). In addition, music and speech processing both require attention, memory, and similar sensorimotor abilities. Furthermore, recent insights suggest 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 hemispheres 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 (Besson et al., 2007; Wong et al., 2007; Patel, 2010; Mok and Zuo, 2012). Besides, musical training appears to drive adaptive plasticity in speech processing networks (Milovanov and Tervaniemi, 2011) and there is a music training transfer between music and acoustic processing in speech, such as frequency and duration (Besson et al., 2011). In line with the findings above, one would expect musicians to exhibit superior performance on pitch processing, and, as a result, being better at learning to discriminate tones.

The unfamiliarity with tone in many Western speakers makes tone languages ideally suited to examine the influence of musical experience on language acquisition (Marie et al., 2011). Previous studies have shown that musicians are more sensitive to subtle pitch variations in speech than non-musicians (e.g., Schön et al., 2004; Micheyl et al., 2006). Behavioral studies clearly provided evidence that lexical tone perception benefits from musical expertise. For example, a relevant study by Gottfried and Riester (2000) 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 as 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 (Gottfried et al., 2004; see also Alexander et al., 2005 for similar results). In another study, by using intact and acoustically modified syllables (silence-center syllables and onset-only syllables) of the four Mandarin tones produced by multiple speakers, Lee and Hung (2008) 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. Specifically, the advantage of the musicians (in accuracy and reaction time) in identifying Mandarin tones decreased when the acoustic information was reduced (from intact syllables to silent-center and onset-only syllables); and musical background mainly benefited the identification of tones 1, 2, and 4. Taken together, these studies show that musicians consistently outperform non-musicians in the area of lexical tone processing of non-tone language speakers.

Much previous related research, such as the studies mentioned above, has focused on comparing musicians and non-musicians with regards to cognition, behavior and brain structure/function (Hassler and Gupta, 1993; Koelsch et al., 2000; Gaser and Schlaug, 2003; Aheadi et al., 2010). These earlier studies have usually compared two groups of participants, musicians and non-musicians, based on musical abilities conferred by musical training/education or based on the skill/level of playing musical instruments. While these criteria suffice to distinguish two different groups in general, they fail to provide insights into which aspects of musical ability contribute to the improved tone perception in musicians, since an individual's musicianship status is not a unitary construct, but comprises multiple abilities, such as singing ability, perceptual abilities, and duration of training. In the current study, we use the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014) as a tool to provide a more fine-grained analysis of the

[2]A numerical substitute has been commonly used for tone contours, with a numerical value assigned to the beginning, end, and sometimes middle of the contour. The numbers 1–5 refer to relative pitch differences; they are not absolute values, and will vary from speaker to speaker.

reported musical abilities of participants. With this, we aim to better understand the differences between musicians and non-musicians and relate these differences in specific musical abilities to different performance in tone perception.

Tone Perception and Visual Information

The extent to which musicians outperform non-musicians in tone perception could be mediated by other factors, such as the presence of visual information, which has been shown to facilitate speech perception (e.g., Sueyoshi and Hardison, 2005; Hirata and Kelly, 2010). Visual speech information is provided by movements in the facial area: specifically, movements created by the face, the head and neck, and the lips. In order to be understood, speakers are assumed to strive to provide optimal acoustic and visual information to meet the demands of the target audience or the communicative situation (Burnham et al., 2002). Several studies (e.g., Burnham et al., 2000, 2001; Mixdorff and Charnvivit, 2004; Mixdorff et al., 2005a,b) have shown that visual speech information is related to the production of lexical tones. When speakers want to convey information about tone (the pitch contour for instance), facial cues (along with gestures) are a common visual resource they resort to alongside the acoustic information (Swerts and Krahmer, 2008; Zheng et al., 2018). Because our mouth, face, and head needs to move in a certain way to produce a given tone, the amplitude (range) and the length (duration) of the visible articulations change. For example, in Mandarin Chinese tones, there are clear differences in the duration of the vowels and the amplitude across tones: tone 3 usually has the longest vowel duration, while tone 4 tends to be the shortest; the amplitude for tone 3 is usually the lowest one, whereas tone 4 normally has the highest amplitude (Tseng, 1981). It makes much sense that these acoustic differences, for instance in the amplitude and the length of the articulation, have correlating visual characteristics (Han et al., 2018). Physiological studies (e.g., Xu and Sun, 2002) suggest certain restrictions with respect to the coordination of the laryngeal and articulatory systems, which may lead to visual cues for tones (Mixdorff et al., 2005a). In addition, regarding prosodic features, a significant correlation has been found between the motion of the head and fundamental frequency during the production of speech (Yehia et al., 2002). For example, in the case of the Mandarin tone 3 (a low-dipping tone in terms of height and contour), the correlated direction of head/neck motion during tone production is usually signaled by a low-falling-rising head movement.

However, the extent to which auditory-visual information facilitates or improves tone identification compared to auditory-only information (i.e., the superiority of bimodal performance compared to unimodal performance) differs widely across individuals (Grant and Seitz, 1998). Furthermore, the benefit of visual/facial information for tone perception depends strongly on context, and in particular on the availability of a clear and reliable acoustic signal. In situations where such a signal is available, extra visual information may actually distract the perceivers instead of facilitating their tone perception, since they are reluctant to use the visual information when acoustic sources are available and reliable. For example, Burnham et al. (2001) have found that in an experiment using clean speech, Australian English speakers performed better in a task of identifying Cantonese words that differed only in tone in the auditory-only (AO) condition than in the auditory-visual (AV) condition (where they had access to lip and face movements).

In our study, we look into the effects of modality and musicianship on Mandarin tone perception. More specifically, we presented musicians and non-musicians with auditory-visual or auditory-only tone stimuli. Because of 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 (Schön et al., 2004), but the extent to which musicians are affected by the presence of (exaggerated) visual information during speech perception has remained largely unexplored. Besides, while they are obviously related, pitch perception is not the same as the identification of lexical tone. While musicians might benefit from the additional information just 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. Musicians may have developed the ability to focus on specific properties of sounds and that this superior ability may in turn help them categorize the sounds and make the relevant decision (Besson et al., 2011). We hypothesize that musicians may still benefit from the added visual information for the Mandarin tone identification, but that this contribution is likely *smaller* than that for non-musicians. While performing tone identification, our participants are exposed to different stimuli to which they have to respond. Over the course of the task, we expect participant's performance to improve progressively. In order to investigate the learning process and to see whether the two participant groups differ with respect to their learning rate, e.g., whether the musicians learn faster, or display superior performance from the beginning, we will look at the performance over time. In general, we assume that performance improving with training. Whether musicians outperform non-musicians from the beginning, or show a steeper learning curve, is an open question.

In sum, we investigate the effects of musical ability on Mandarin tone identification by tone-naive listeners (speakers of Dutch), with a specific interest in how their performance is mediated by differences in modality. The Gold-MSI was used to measure the musical sophistication of each participant. We conducted a linear regression analysis to find out whether a specific musical ability/skill as measured by the subscales of the Gold-MSI is related to successful tone identification. Since the effects of our two independent variables might vary among tones, we subsequently assess the effects for each tone individually in our study.

MATERIALS AND METHODS

A 2 (musical ability) × 2 (modality) between-participant design was employed in this study. Two groups of participants (musicians and non-musicians) were randomly divided over two

modality conditions (auditory-visual vs. auditory-only). Given the likelihood of learning effects, it was not possible to include modality as a within-participant factor. Accuracy, defined as the percentage correct identification of a tone based on its auditory realization, was the dependent variable.

Participants

There were 170 participants comprising two groups that differed in musical ability: 86 non-musicians (mean age 22, 62 females) were recruited from the Tilburg University participant pool; 84 musicians (mean age 22, 35 females) were recruited from the Fontys School of Fine and Performing Arts (located in Tilburg). Eighty-three percent of the participants were native speakers of Dutch, with the remaining participants reporting German, French, Greek, English, Portuguese, Spanish, Italian, Russian, Indonesian, Bengali, and Arabic as their native language. None of them were native speakers of tone languages, and none had had formal training to learn a tone language. The musician group consisted of participants who had eight or more years of intensive music training and practice up until 2017, while none of the non-musicians had received continuous musical training[3]. A self-reported musical sophistication questionnaire, the Gold-MSI, was used to assess the musical skills and behaviors of the participants.

Materials and Stimuli

Gold-MSI

Individuals differ in their repertoire of musical behaviors and in the level of skill they display for particular musical behavior (Müllensiefen et al., 2014). The Gold-MSI is an attested self-assessment instrument that measures individual differences on multiple dimensions toward musical skills and behaviors. Thirty-eight items in total measure individual differences in musical sophistication. Among them, 31 items are rated on a seven-point scale (1 = *completely disagree* and 7 = *completely agree*); for the remaining 7 items, participants choose one answer from 7 options (the first option yields 1 point; the seventh option yields 7 points (e.g., I can play 0/1/2/3/4/5/6 or more musical instruments).

The Gold-MSI is a multi-faceted instrument that measures different aspects of musical sophistication. It has five sub-scales and one general score for the following facets: *active engagement* comprised of nine items covering a range of active musical engagement behaviors (e.g., "I spend a lot of my free time doing music-related activities."); *perceptual abilities* also with nine items, most of them related to musical listening skills (e.g., "I am able to judge whether someone is a good singer or not."); *musical training* combines seven items about the extent of musical training and practice (e.g., "I have had formal training in music theory for __ years."); *singing abilities* consists of seven items that reflected different skills and activities related to singing (e.g., "I am able to hit the right notes when I sing along with a recording."); *emotions* included six items describing active behaviors in response to music (e.g., "I sometimes choose music that can trigger shivers down my spine."); the *general musical sophistication* had 18 items which incorporated representative questions from all the five sub-scales.

The Gold-MSI is used in this study to measure the individual's musicality. The factor structure and internal reliability of the Gold-MSI have previously been tested with a German sample (Schaal et al., 2014), and validated for use with secondary school pupils in a large German sample of 11–19 years old (Fiedler and Müllensiefen, 2015); it has also been used in a study with young and older Dutch adults (Vromans and Postma-Nilsenová, 2016).

Stimulus Construction

We constructed a word list with 10 Mandarin monosyllables (e.g., ma, ying …, based on stimulus material from Francis et al., 2008 and from Chen and Massaro, 2008). Each of these syllables was chosen in such a way that the four tones would generate four different meanings, resulting in 40 (10 syllables × 4 tones) different existing words in Mandarin Chinese [see Appendix (**Table A1**)] for a complete list of the stimuli, previously used by Han et al., 2018).

Material Recording

Four (2F, 2M) 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 ("pronounce these words as if you were talking to someone who is not a Chinese speaker"), with the recording of the natural stimuli preceding the recording of the teaching style stimuli. In both conditions, there were no other instructions or constraints imposed on the way the stimuli should be produced. A 20-min break was given to the speakers between the two recordings to counter fatigue. Speaking style as a factor is not reported on in the current paper[4].

The images and sounds from the speakers were recorded by Eye-catcher (version 3.5.1) and Windows Movie Maker (2012). One of the advantages of the Eye-catcher system is that the camera is located behind the computer screen and thus records people "through" the 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, two sets of 160 video stimuli (10 syllables × 4 tones × 4 speakers) were produced in teaching and in natural modes. These video clips were segmented into individual tokens, with each token containing one stimulus. We used Format Factory (version 3.9.5) to extract the sound from each video to generate stimuli for the auditory-only conditions. This resulted in four types of experimental stimuli: video + teaching (VT); video + natural (VN); audio + teaching (AT); audio + natural (AN), with each set containing 160 testing stimuli. Therefore, the auditory-visual (Video) conditions include VT and VN, and the auditory-only (Audio) conditions are AT + AN.

[3]Some of the non-musicians periodically had had some musical education, for example in their middle school.

[4]An analysis with speaking style as one factor has been reported elsewhere (Han et al., 2018). This paper focused on the effects of musicality and modality on tone identification for tone-naïve participants (experimental stimuli were the same as the previous paper, but there was a different group of participants). Including speaking style as a factor in the analyses did not meaningfully alter the effects of the other independent variables.

Procedure

This study was carried out in accordance with the recommendations of the Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences, Tilburg University. The protocol was approved by the Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences, Tilburg University. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

The task of the participants was to (learn to) identify the tones they perceived from auditory-visual or auditory-only stimuli. We used E-prime (Version 2.0; Zuccolotto et al., 2012) to set up and run the experiment. Upon their arrival, participants signed an informed consent form that contained information about the nature of the experiment and their voluntary participation in it, agreeing for the data to be used for scientific research. They then filled out the Gold-MSI questionnaire, assessing their musical background. Next, they received a brief instruction about Mandarin Chinese tones. This instruction was displayed on the screen (see **Figure 1** for a screenshot): "there are four tones in Mandarin Chinese: the first tone is a High-Level tone, symbolized as "–", the second tone is a Mid-Rising tone, symbolized as "/", the third tone is a Low-Dipping tone, symbolized as "ˇ", and the fourth tone is a High-Falling tone, symbolized as "\".

The introduction was followed by exposure to three practice trials, either auditory-only or auditory-visual, depending on the condition they were randomly assigned to. After those, the experiment leader checked whether they had fully understood the concept of tones (in particular the symbols) and the task was clear. The main experiment consisted of 160 testing stimuli (video/audio), which were presented in an individually randomized order (operated by E-Prime). Participants received feedback in both the practice and testing trials in the form of a "good job" or "incorrect" message on the screen after their response. If no response was recorded within 10 s after the end of the stimulus, "no response" was shown. This registered as a missing response.

Participants were seated in a sound-attenuated room, wearing headsets, directly in front of the PC running the experiment. All stimuli were presented at a comfortable hearing level. They were told to press the designated keys with the corresponding tone symbols ("–", "/", "ˇ", "\", see **Figure 2**) on them as accurately and as quickly as possible after they made their decisions. Their responses (and reaction times) were recorded automatically by E-prime.

RESULTS

Overall Tone Perception

In order to examine to what extent modality (auditory-visual vs. auditory-only) and musical ability (musicians vs. non-musicians) affect the perception of Mandarin Chinese tones, a mixed ANOVA was carried out with modality and musical ability as between-subject factors, and speaker and tone as within-subject factors. The percentage of correct responses (accuracy) was analyzed as the dependent variable.

Figure 3 depicts the performance of musicians and non-musicians in the two experimental conditions. Overall, participants were able to identify Mandarin tones well-above chance (25%) (a histogram of each participant's accuracy

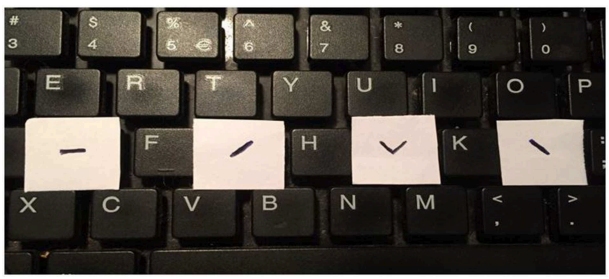

FIGURE 2 | Picture of the designated keys with tone symbols.

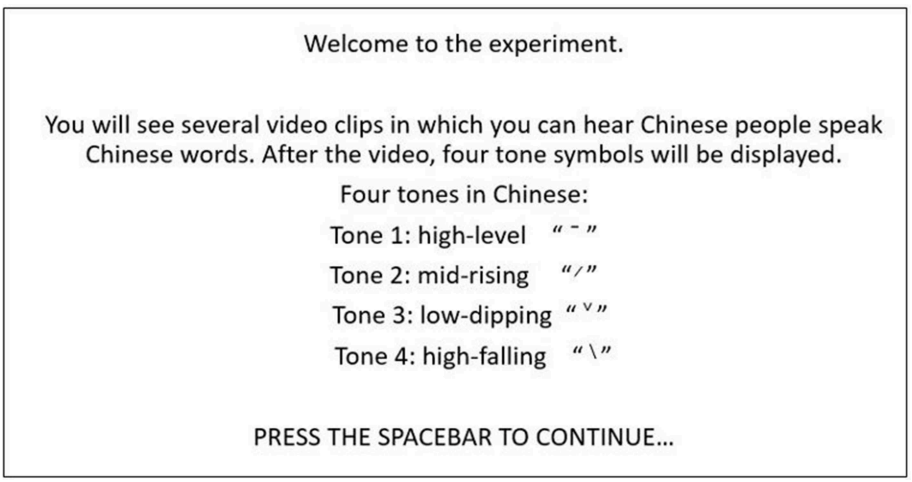

FIGURE 1 | Screenshot for a brief introduction for Mandarin Chinese tones (in auditory-visual conditions).

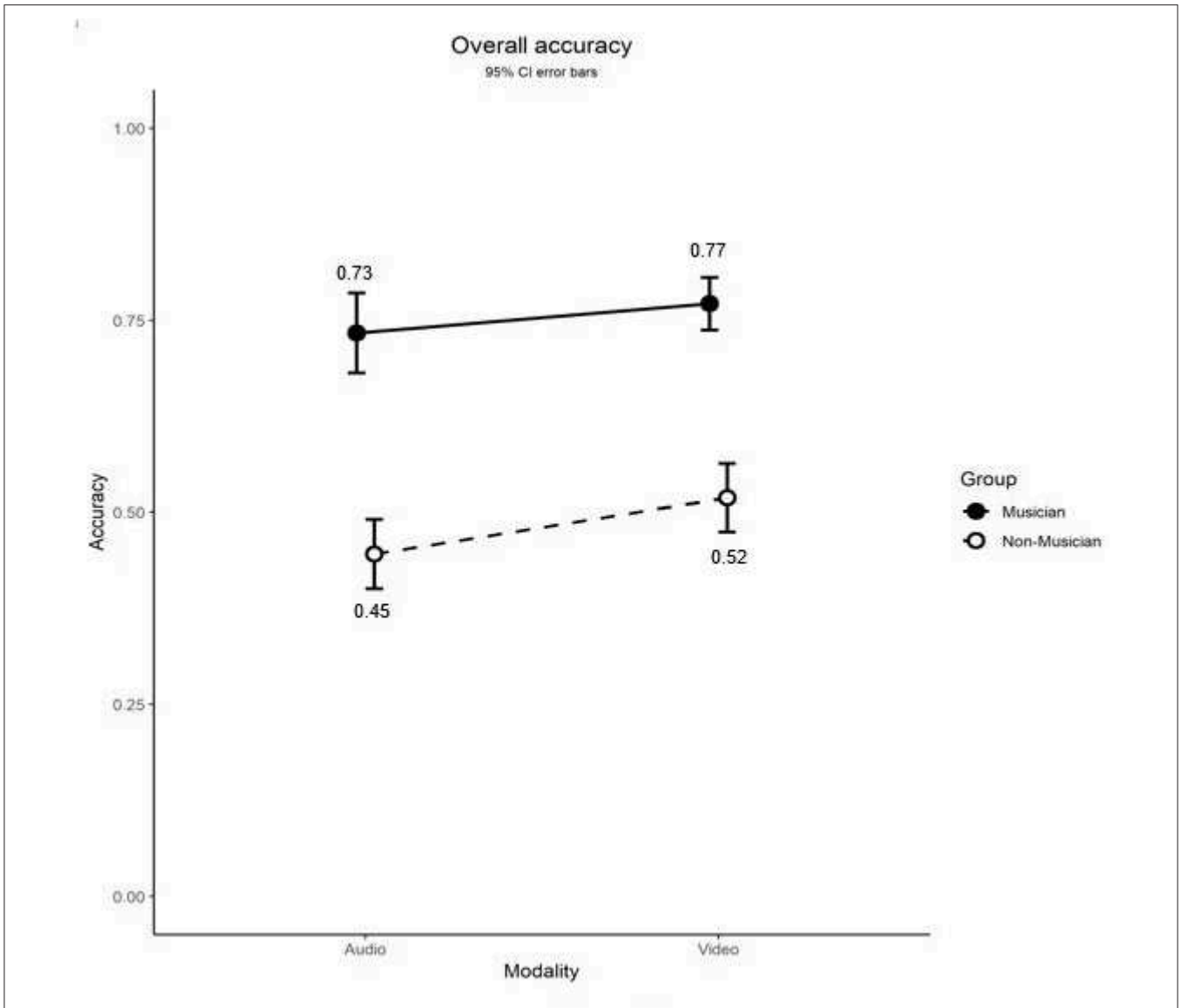

FIGURE 3 | Average accuracy in percentage correct of Mandarin tone identification as a function of musical background and modality. Video represents auditory-visual and Audio represents auditory-only. Error bars represent standard errors.

in Appendix (**Figure A1**) shows that only about 1.8%, i.e., 3 individuals, of the participants score below chance over all tones), and the musician group outperformed the non-musician group in both experimental conditions, as indicated by a higher percentage of correct responses ($M = 75\%$, $SE = 0.02$ vs. $M = 48\%$, $SE = 0.02$). The difference in percentage correct between musicians and non-musicians was statistically significant [$F_{(1, 166)} = 150$, $p < 0.001$, $\eta_p{}^2 = 0.48$], which was in line with our hypothesis that musical ability positively affects the ability to identify Mandarin tones.

The statistical analyses further showed that the auditory-visual condition ($M = 65\%$, $SE = 0.02$) yielded significantly higher accuracy scores than the auditory-only condition ($M = 59\%$, $SE = 0.02$); $F_{(1, 166)} = 6.39$, $p = 0.012$, $\eta_p{}^2 = 0.037$. 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. For musicians, seeing the speaker (video condition) helped them to identify more tones correctly as compared to only listening to the speaker (audio condition): 77 vs. 73%, $t_{(13, 438)} = 5.12$, $p < 0.001$. For non-musicians, the effect of visual information was even greater: 52% in auditory-visual condition and 45% in auditory-only condition, respectively; $t_{(13, 758)} = 8.60$, $p < 0.001$. Notably, there was no significant interaction between musicality and modality: $F_{(1, 166)} = 0.66$, $p = 0.42$, $\eta_p{}^2 = 0.004$, which indicates that the effects of musicality or modality on tone perception are not dependent on each other. Overall, the lack of two-way interaction indicates

that both groups identified tones more accurately in the video condition than in the audio condition.

Musicians indeed outperformed non-musicians in terms of average accuracy in Mandarin tone identification. However, as we mentioned in the first section of this study, the learning patterns over time for both participant groups are of interest in this regard as well. We included stimulus number (1–160) as a predictor in the regression model. In the regression with stimulus number as a predictor (both separately and together with musicianship and modality) and the average accuracy as the dependent variable, stimulus number significantly predicts performance: $F_{(1, 638)} = 32.82$, $p < 0.001$, with an R^2 of 0.049 and $F_{(3, 636)} = 829$, $p < 0.001$, with an R^2 of 0.796, respectively. We also did a regression analysis with the interaction between stimulus number and musicianship (to see whether the learning rate of musicians and non-musicians differed), but that effect was small and non-significant ($p = 0.057$). The scatter plots in **Figure 4** show the learning curves for musicians and non-musicians in audio (auditory-only) and video (auditory-visual) conditions. In general, both groups of participants showed improvement in performance over time, which shows both groups are learning. Although musicians performed better than non-musicians across the board (they had a higher accuracy), they did not learn faster than non-musicians.

Individual Tone Perception

Figure 5 shows the identification performance in terms of accuracy of musicians and non-musicians in the two experimental conditions for each of the four Mandarin tones. Musicians again performed better than non-musicians for all four Mandarin tones. While musicality continued to play a significant role in each tone's identification: $F_{(3, 498)} = 10.76$, $p < 0.001$, $\eta_p^2 = 0.061$, modality did not affect the accuracy of identifying each tone: $F_{(3, 498)} = 1.79$, $p = 0.149$, $\eta_p^2 = 0.011$. For both participant types (and for all combinations of modality) tone had a strong effect on accuracy [$F_{(3, 498)} = 102$, $p < 0.001$, $\eta_p^2 = 0.38$]. Furthermore, there was a significant two-way interaction between tone and musicality and a significant three-way interaction between tone, musicality, and modality [$F_{(3, 498)} = 3.33$, $p = 0.022$, $\eta_p^2 = 0.02$]. Given the significant three-way interaction we conducted four separate (one for each tone) two-by-two analyses of variance with musicality and modality as independent variables and accuracy as the dependent variable. For tone 1, this resulted in a main

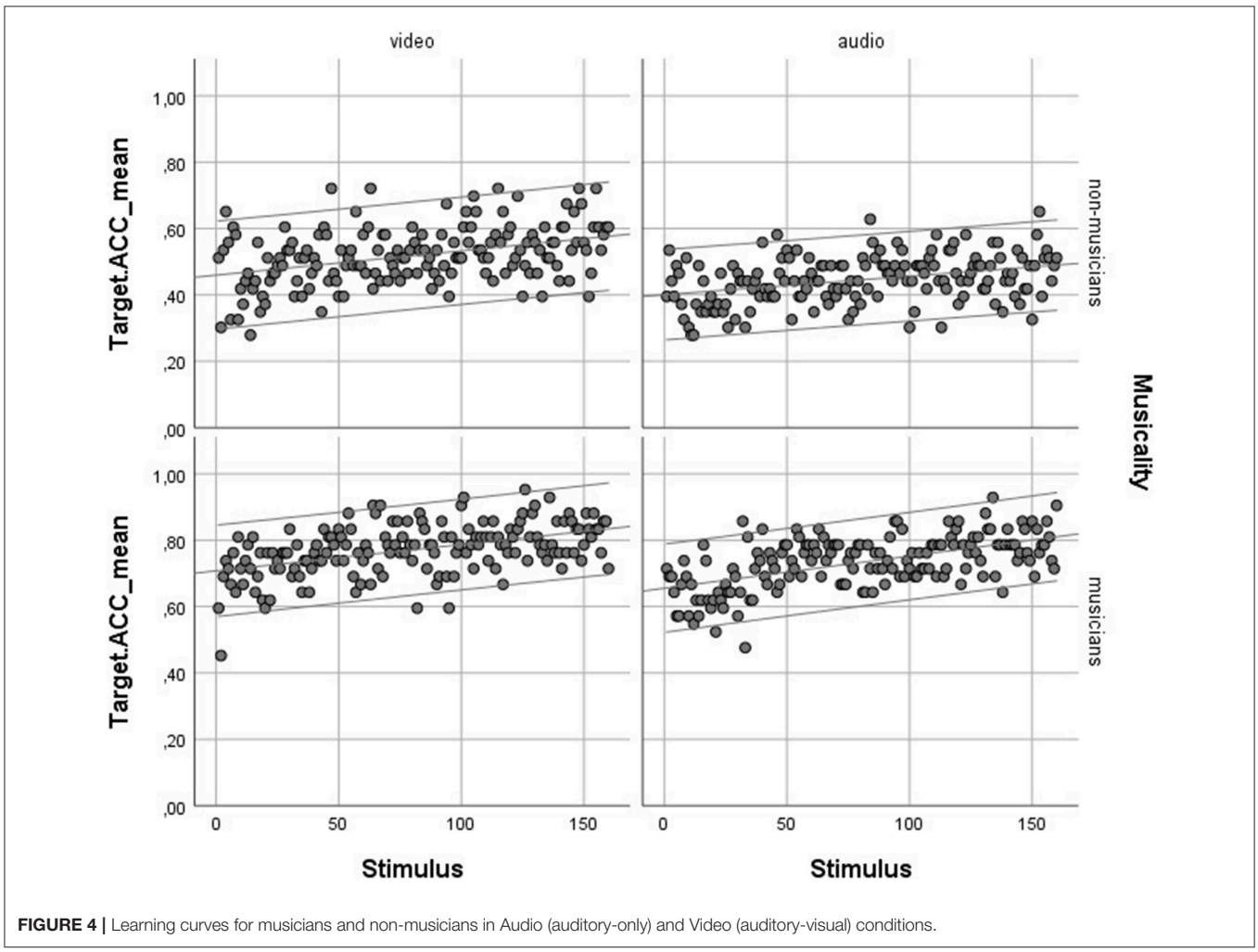

FIGURE 4 | Learning curves for musicians and non-musicians in Audio (auditory-only) and Video (auditory-visual) conditions.

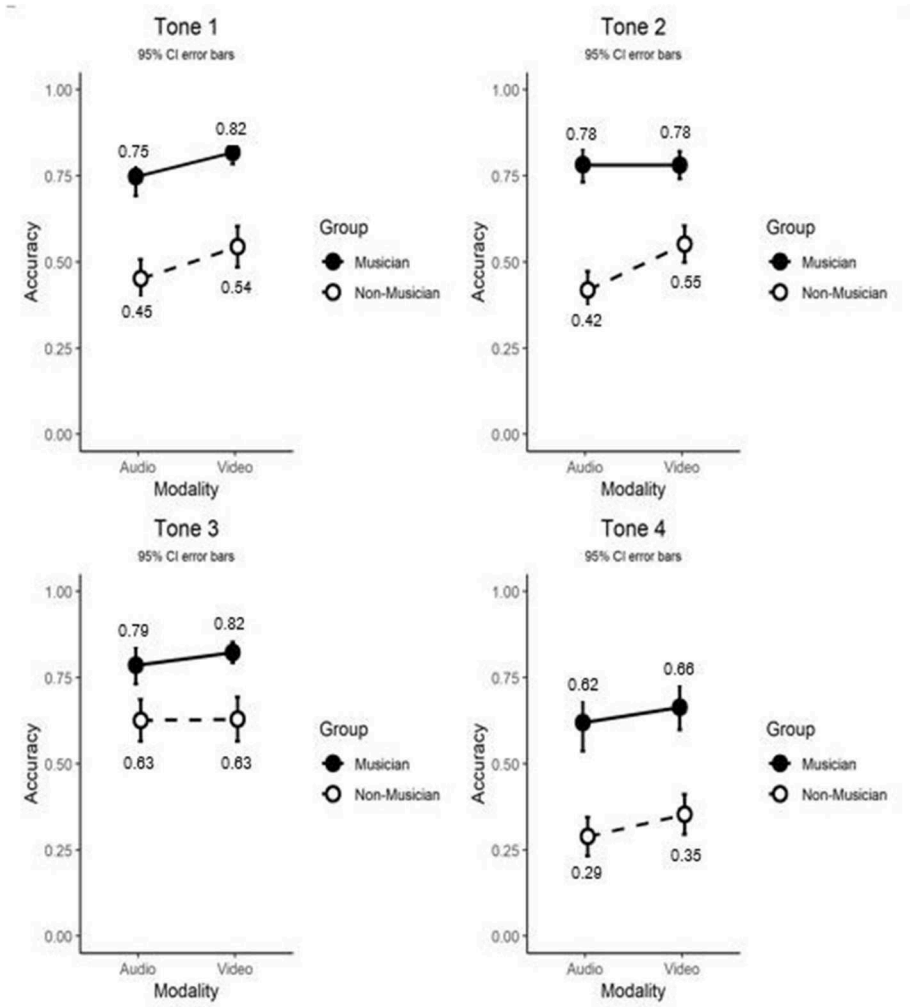

FIGURE 5 | Average accuracy in percentage correct of Mandarin tone identification as a function of modality, musicality and tone. Video represents auditory-visual and Audio represents auditory-only. Error bars represent standard errors.

effect of modality [$F_{(1, 169)} = 9.99$, $p = 0.002$], with accuracy being higher in the video condition, a main effect of musicality [$F_{(1, 169)} = 122.18$, $p < 0.001$], with accuracy being higher for musicians, but no significant interaction between the two main effects [$F_{(1, 169)} = 0.19$, $p = 0.66$]. For tone 2, we found a main effect of modality [$F_{(1, 169)} = 7.15$, $p = 0.008$], with accuracy being higher in the video condition, a main effect of musicality [$F_{(1, 169)} = 144.82$, $p < 0.001$], with musicians being more accurate than non-musicians, but this effect was qualified by a significant interaction between musicality and modality [$F_{(1, 169)} = 7.45$, $p = 0.007$]. This significant interaction was further analyzed in two independent *t*-tests contrasting the auditory-visual and the auditory-only modality separately for each participant group. For musicians, this analysis showed no significant difference between the two modalities; $t_{(82)} = -0.04$, $p = 0.97$. Non-musicians however, were significantly more accurate in the video condition; $t_{(84)} = 3.55$, $p = 0.001$. For tone 3, there was no main effect of modality [$F_{(1, 169)} = 0.576$, $p = 0.450$], but the main effect of musicality [$F_{(1, 169)} = 43.46$, $p \leq 0.001$] was present, again with musicians being more accurate. As with tone 1, there was no significant interaction between musicality and modality [$F_{(1, 169)} = 0.40$, $p = 0.556$]. Similarly, for tone 4, there was no main effect of modality [$F_{(1, 169)} = 2.79$, $p = 0.097$], but musicians scored better than non-musicians [$F_{(1, 169)} = 95.79$, $p < 0.001$], with accuracy being higher for musicians, but no significant interaction between the two main effects [$F_{(1, 169)} = 0.092$, $p = 0.762$].

In sum, the main effect of musicality holds for each tone: musicians performed better than non-musicians with all tones, the significant two-way interaction points toward differences of degree. In contrast, the effect of modality is present only for tone 1, where performance in the video condition is consistently superior. There is a main effect for tone 2, but this is driven by the significant two-way interaction where musicians are unaffected by modality, but non-musicians do perform better in the auditory-visual condition compared to the auditory-only condition.

The tonal confusion matrices below (**Table 1A** for musicians and **Table 1B** for non-musicians) give more insight in the way our participants perceived individual Mandarin tones. The data show that, regardless of the fact that musicians performed much better than non-musicians on the identification of each tone, both groups of participants can identify the tones above chance (25%). For both groups, the low-dipping tone 3 was the easiest to recognize (80.4 and 62.7% for musicians and non-musicians), while the high-falling tone 4 was the most difficult to identify (64.2 and 32.1% respectively). In general, when hearing tone 4, participants often confused it with the high-level tone 1, though the confusions were not necessarily symmetrical: tone 1 was mostly confused with mid-rising tone 2, rather than with tone 4. The participants seemed to be able to identify equally well-tone 1 and tone 2. These confusions are similar to the ones we find in our previous study (Han et al., 2018).

A More Fine-Grained Look at Musicality

In the above analyses, we grouped the participants according to their affiliated institutions: musicians (participants from the Fontys School of Fine and Performing Arts) and non-musicians (participants from Tilburg University). However, while the sample of musicians was clearly more musical than the non-musicians, we did not quantify the extent of the difference, neither for musicality as a whole or for different aspects of musicality, nor did we take into account the possibility that, at least in some areas, there might be university students with considerable musical experience. To get a better handle on the musical abilities of both the musicians and non-musicians in our study, we employed the Gold-SMI questionnaire. With the five dimensions/sub-scales included in the questionnaire (active engagement, perceptual abilities, musical training, singing abilities, and emotions) as dependent variables, and group membership (musicians and non-musicians) as the independent variable, we conducted a multivariate analysis of variance (MANOVA) in order to get a more detailed picture of the differences and similarities in musicality between these two groups. **Table 2** contains a summary of the outputs for the five dependent variables for musicians and non-musicians. On average, musicians attained a higher value when compared to non-musicians in each category.

Using Pillai's trace, there was a significant multivariate effect of group membership (musician vs. non-musician) on the five musical dimensions of the subject, V =0.67, $F_{(5,164)} = 67.6$, $p < 0.001$, $\eta_p{}^2 = 0.67$, with musicians scoring higher than non-musicians on all five subscales. As shown in **Table 1**, significant univariate effects were also found for the five dimensions. The effect sizes (partial eta-squared) of the five subscales differ considerably, ranging from 0.24 to 0.66, indicating marked differences in the importance of the subscales. The most prominent difference between musicians and non-musicians is in (reported) musical training ($\eta_p{}^2 = 0.66$).

Musicality and Tone Perception

In order to obtain a comprehensive view of the relation between musical experience of the participants and their tone perceptual ability, we first constructed a (Pearson) correlation matrix. **Table 3** shows the correlations among the five sub-scales (active engagement, perceptual abilities, musical training, singing abilities, and emotions) and the accuracy of the participant's tone identification. The results indicated that there is a significant positive association between all sub-scales and the performance (accuracy).

TABLE 1A | Confusion matrix for tone (percentage correct) in musicians.

		Responded tone			
		1	**2**	**3**	**4**
Presented tone	1	**78.3**	10.8	3.0	7.9
	2	3.6	**78.2**	14.9	3.4
	3	0.9	12.2	**80.4**	6.5
	4	23.1	9.0	3.6	**64.2**

Numbers in bold face represent the category in which the responded tone corresponds with the produced tone.

TABLE 1B | Confusion matrix for tone (percentage correct) in non-musicians.

		Responded tone			
		1	**2**	**3**	**4**
Presented tone	1	**49.8**	26.3	8.9	15.0
	2	16.8	**48.5**	20.7	14.1
	3	5.2	15.1	**62.7**	17.0
	4	34.7	23.0	10.3	**32.1**

Numbers in bold face represent the category in which the responded tone corresponds with the produced tone.

TABLE 2 | MANOVA results for non-musicians (*N*= 86) and musicians (*N* = 84).

	Non-musicians		Musicians		*F* (1, 168)	$\eta_p{}^2$
Variable	**Score**	***SD***	**Score**	***SD***		
Active engagement	3.95	1.06	5.34	0.63	106.17**	0.39
Perceptual abilities	4.88	0.87	5.97	0.55	94.86**	0.36
Musical training	2.70	1.38	5.64	0.57	325.29**	0.66
Singing abilities	3.80	1.13	5.46	0.77	125.49**	0.43
Emotions	5.02	0.86	5.86	0.63	52.03**	0.24

*Scores range from 1 to 7. The maximum score is 7; **p < 0.001.*

TABLE 3 | Table of Correlations for Gold-MSI variables and accuracy of tone perception.

	1	2	3	4	5	6
1	Active engagement					
2	Perceptual abilities	0.686***				
3	Musical training	0.727***	0.716***			
4	Singing abilities	0.630***	0.807***	0.717***		
5	Emotion	0.700***	0.678***	0.561***	0.560***	
6	Accuracy	0.484***	0.547***	0.653***	0.525***	0.441***

****p < 0.001, one-tailed. N = 170.*

We also conducted two linear regression analyses to see if the more fine-grained Gold-MSI scales predict anything above the binary classification between musicians and non-musicians. Specifically, we compared a linear regression with groups (musicians vs. non-musicians) as the predictor (Model 1) with a regression that also includes the five sub-scales of Gold-MSI as predictors (Model 2), using the overall accuracy as outcome variable. **Table 4** contains the summary for the two models. The data in the table show that adding Gold-MSI predictors significantly improves the model ($R^2 = 0.47$ in Model 1 and $R^2 = 0.51$ in Model 2; $F_{change} = 2.99$, $p = 0.013$); and that musical training is the only Gold-MSI variable that predicts additional variance in identification accuracy: $b = 0.37$, $\beta = 0.24$, $t_{(163)} = 2.08$, $p = 0.039$, although weakly so. None of the other predictors were significantly related to accuracy.

From the analysis above, the amount of musical training received emerges as the only predictor of accuracy of tone perception. To investigate whether the other predictors add anything to the effects of training considering both groups simultaneously, we repeated our regression analysis comparing a model with musical training as the predictor (Model 3) and a model with musical training plus the other four sub-scales of Gold-MSI as predictors (Model 4). As before, the data in **Table 5** show that the other four predictors did not significantly improve the accuracy of the tone perception ($R^2 = 0.43$ in Model 3 and $R^2 = 0.44$ in Model 4; $F_{change} = 1.29$, $p = 0.28$).

The above analyses provide a general picture of the relationships between musicality (and the five subscales) and the accuracy of tone perception. As a final step we zoomed in on the individual tones to see whether these particular factors predicted the perception of specific tones. Multiple regressions were conducted for musicians and non-musicians combined for each of the individual tones. The results showed that the musicality significantly predicted accuracy for each of the individual tones, and out of the five individual factors, musical training was the only constant factor of predicting the accuracy for each individual tone, while the other factors had no consistent effect on accuracy.

TABLE 4 | Multiple linear regressions for accuracy of tone identification in Model 1 (Groups as the predictor) and Model 2 (Groups + five sub-scales as the predictors).

	Model	*B*	*SE*	β	*t*	*p*
1	Groups	27.02	2.24	0.68	12.09	0
	$R^2 = 0.47$					
	$F_{change} = 146.09$					<0.001
2	Groups	17.84	3.80	0.45	4.70	0.00
	Active engagement	−0.21	0.19	−0.11	−1.15	0.25
	Perceptual abilities	0.41	0.26	0.17	1.55	0.12
	Musical training	0.37	0.18	0.24	2.08	0.04
	Singing abilities	−0.12	0.22	−0.05	−0.53	0.60
	Emotion	0.31	0.32	0.08	0.97	0.33
	$R^2 = 0.51$, $R_{change}^2 = 0.05$					
	$F_{change} = 2.99$					0.013

DISCUSSION

We set out to investigate two factors that influence Mandarin tone perception in tone-naïve listeners: the musicality of the participants (comparing musicians and non-musicians) and the stimulus modality (comparing audio-visual and auditory-only stimuli). The findings of the study were:

(1) All participants were able to identify Mandarin tones well-above chance level;
(2) Musicians outperformed non-musicians in both auditory-visual and auditory-only presenting conditions;
(3) The amount of musical training is the only factor that relates to successful tone identification;
(4) The auditory-visual condition yielded significantly better results than the auditory-only condition;
(5) The effect of musicality and modality on tone identification varies among individual tones.

We will discuss these findings one by one.

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 for non-native listeners (Gottfried and Riester, 2000; Gottfried et al., 2004; Alexander et al., 2005; Delogu et al., 2006, 2010; Lee and Hung, 2008). Based on our findings, we would argue that the length of musical training led (musicians) listeners' to a better performance in Mandarin Chinese tone identification: listeners with more musical training showed considerably greater accuracy in their identification (75 vs. 48%). Importantly, although the musicians in our study performed well in the identification task (79% at the highest for the dipping tone 3), they did not achieve native-like performance (as reported in Han et al., 2018), and the learning patterns tell us that musicians did not learn faster than non-musicians. Musicians showed their superior performance at the beginning of the task. Interestingly, the increase in performance follows a linear for both musicians and non-musicians, and does not seem to plateau, indicating that more exposure leads to better performance, and potentially (in the case of a longer learning period) may lead to still higher final accuracy scores.

TABLE 5 | Multiple linear regressions for accuracy of tone identification in Model 3 (musical training as the predictor) and Model 4 (musical training + four other sub-scales as the predictors).

	Model	*B*	*SE*	*B*	*t*	*p*
3	Musical training	1.02	0.09	0.65	11.17	0.00
	$R^2 = 0.43$					
	$F_{change} = 124.81$					<0.001
4	Musical training	0.87	0.15	0.56	5.62	0.00
	Active engagement	−0.17	0.20	−0.09	−0.87	0.38
	Perceptual abilities	0.29	0.28	0.12	1.05	0.29
	Singing abilities	0.07	0.23	0.03	0.32	0.75
	Emotion	0.35	0.34	0.09	1.02	0.31
	$R^2 = 0.44$, $R_{change}^2 = 0.02$					
	$F_{change} = 1.29$					0.28

Although musical training has been identified as the only factor that predicts tone identification, it is not a foregone conclusion that the other aspects of musicality do not affect the learning of Mandarin tones. Because our study uses natural groups of musicians and non-musicians, musical training is confounded with group membership. Importantly, if we analyse both groups separately, there is no relationship between musical training and tone identification performance among the musicians and non-musicians. The absence of the relationship between musical training and tone identification in musicians might be due to a lack of variation in training among musicians (a restriction of range effect). Alternatively, since the Gold-SMI is originally intended for using in the general population, it may not be able to capture the more subtle differences among musicians in as much detail as is required to differentiate among musicians. In addition, the parts of the Gold-MSI we used all relied on self-report, which might not be able to capture important differences in factors such as perceptual abilities. However, and importantly, Müllensiefen et al. (2014) reported high correlations (ranging from 0.30 to 0.51) for the relation between self-report and objective listening performance (see page 9, for the AMMA listening test) and similarly in an online listening test (correlations ranging from 0.11 to 0.52). Future studies could include behavioral tests (also present in the Gold-MSI) to be able to better characterize the differences in musical skills and relate them to tone identification.

Nevertheless, our findings point to the interesting possibility of aiding language learning by providing learners additional musical training. Since musical training is the only consistent predictor for performance on the tone identification task, and training is something that potentially anyone can do—it is not a talent or innate ability—our results are promising for educational purposes. For example, second/foreign language learners could get some musical training to facilitate their language learning; schools can enrich students' curriculum with musical lessons; teachers may consider blending musical training into their language materials.

With respect to modality, tone-naïve listeners were able to identify tones better when they saw *and* heard the speakers compared to when they only heard them. This supports the hypothesis that visual information plays a facilitating role in learning to identify Mandarin tones for tone-naive listeners, although the effect was not that large, with participants' accuracy increasing by 6% in the auditory-visual condition. Rather than distracting the listeners (as suggested by Burnham et al., 2001), the presence of facial expressions appears to facilitate Mandarin tone perception in clean speech. Both participant groups benefited from visual information, but numerically the non-musicians did so more than the musicians. This could be because musicians are trained to be particularly sensitive toward acoustic information, and they are already so good at identifying tones that the additional contribution of visual information is limited. This explanation is in line with our earlier assumption that musicians would benefit less from the added visual information compared to non-musicians. In our data, the modality effect is restricted to tones 1 and 2 (for non-musicians). This may be related to the intrinsic properties of individual tones, as tones differ in how easy they are to identify, and in the amount of auditory contour information they provide. For example, the auditory contour of tone 1 (high-level tone) and 2 (mid-rising tone) is much less pronounced than that of tone 3 (low-dipping tone) and 4 (high-falling tone). As the auditory information is often mimicked in facial expressions (see for example, Swerts and Krahmer, 2008), there is simply less auditory information to transfer to the visual domain. Regardless of their specific contour, the tones differ in their overall difficulty, with tone 3 being the easiest and tone 4 being the hardest (as shown in the confusion matrices). It might be that our non-musicians ignored visual information in tone 3 because auditory information was sufficient for them, and, also ignored visual information for tone 4, because combining the auditory and visual information is too challenging. In contrast, tone 1 and 2 present the sweet spot where perceivers are able to take both auditory and visual information into account.

Our findings imply that (non-native) listeners learning Mandarin tones might benefit from pointing out the information that visual cues can contribute. Although we do not really know yet what the exact visual cues are, or in other words, what the listeners should look at, our finding is a good starting point for further exploration. For instance, in a teaching context, teachers should consider using their facial expressions while talking to students/learners, while learners/students could be trained to consciously attend to visual information. Similarly, in online learning environments, using video is likely to be more effective than using audio-only material.

The extent to which the listeners can benefit from visual cues also depends on individual speaker characteristics. There are substantial differences in the degree to which the speakers' faces exhibit relevant characteristics (Gagné et al., 1994; Bradlow and Bent, 2002; Han et al., 2018). Most previous studies have concluded that female speakers in general are better than male speakers are at displaying salient articulation, such as expanding their overall vowel space and increasing their F0 mean (Kricos and Lesner, 1982; Cox et al., 1987; Ferguson, 2004; Ferguson and Kewley-Port, 2007). However, due to the limited number of speakers, our study does not allow us to draw conclusions about gender differences, or speaker differences for that matter. Nevertheless, further research should take into account the variations between speakers' realizations of visual information.

Crucially, individual tones are important contributors to the observed differences in tone identification. In other words, it is more important which tone the listeners hear than the modality in which it is presented. The low-dipping tone 3 is the easiest one to identify, while all listeners had more difficulty identifying the high-falling tone 4, and this holds for both musicians and non-musicians in both experimental conditions (auditory-visual and auditory-only). This is possibly due to their specific temporal characteristics—tone 3 has the longest duration and two intensity peaks, while tone 4 has the shortest duration, and only one intensity peak. Our findings with respect to the accuracy differences between tones differ somewhat from previous studies: while tone 3 has indeed consistently been found the easiest to identify, due to the longest vowel duration, tone 4 has not always been found to be the most

difficult one to recognize (Blicher et al., 1990; Fu and Zeng, 2000; Mixdorff et al., 2005b). Nevertheless, it is clear that individual tones differ in learnability, which, too, is relevant when considering teaching Mandarin tones (for example when designing a curriculum).

CONCLUSION

In sum, the present study contributes to the literature on the relationship between musicality and tone identification, and the roles played by auditory and visual speech information. The results showed that musical training in particular facilitates Mandarin tone perception. Furthermore, learning Mandarin tones can be facilitated by being aware of the information provided by both the auditory and the visual modality. 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.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Dr. J. M. N. E. Jans. Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YH, MM, and MS conceived the original idea. YH carried out the experiment and wrote the full manuscript. MG, MM, and MS contributed to manuscript revision and supervised the whole process. All authors were involved in the statistical analysis and contributed to the final manuscript.

FUNDING

This project was sponsored by China Scholarship Council (CSC).

ACKNOWLEDGMENTS

Special thanks go to Marlon Titre and Fontys School of Fine and Performing Arts for their participation and assistance during the data collection.

REFERENCES

Aheadi, A., Dixon, P., and Glover, S. (2010). A limiting feature of the Mozart effect: listening enhances mental rotation abilities in non-musicians but not musicians. *Psychol. Music* 38, 107–117. doi: 10.1177/0305735609336057

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

Besson, M., Chobert, J., and Marie, C. (2011). Transfer of training between music and speech: common processing, attention, and memory. *Front. Psychol.* 2:94. doi: 10.3389/fpsyg.2011.00094

Besson, M., Schön, D., Moreno, S., Santos, A., and Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restor. Neurol. Neurosci.* 25, 399–410.

Blicher, D. L., Diehl, R. L., and Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin tone 2/tone 3 distinction: evidence of auditory enhancement. *J. Phon.* 18, 37–49.

Bradlow, A. R., and Bent, T. (2002). The clear speech effect for non-native listeners. *J. Acoust. Soc. Am.* 112, 272–284. doi: 10.1121/1.1487837

Burnham, D., Ciocca, V., Lauw, C., Lau, S., and Stokes, S. (2000). "Perception of visual information for Cantonese tones," in *Proceedings of the Eighth Australian International Conference on Speech Science and Technology* (Canberra, ACT: Australian Speech Science and Technology Association), 86–91.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science* 296, 1435–1435. doi: 10.1126/science.1069587

Burnham, D., Lau, S., Tam, H., and 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).

Chao, Y. R. (1948). *Mandarin Primer*. Cambridge, MA: Harvard University Press.

Chen, T. H., and Massaro, D. W. (2008). Seeing pitch: visual information for lexical tones of Mandarin-Chinese. *J. Acoust. Soc. Am.* 123, 2356–2366. doi: 10.1121/1.2839004

Chobert, J., and Besson, M. (2013). Musical expertise and second language learning. *Brain Sci.* 3, 923–940. doi: 10.3390/brainsci3020923

Cox, R. M., Alexander, G. C., and Gilmore, C. (1987). Intelligibility of average talkers in typical listening environments. *J. Acoust. Soc. Am.* 81, 1598–1608. doi: 10.1121/1.394512

Delogu, F., Lampis, G., and Belardinelli, M. O. (2006). Music-to-language transfer effect: may melodic ability improve learning of tonal languages by native nontonal speakers? *Cogn. Process.* 7, 203–207. doi: 10.1007/s10339-006-0146-7

Delogu, F., Lampis, G., and Belardinelli, M. O. (2010). From melody to lexical tone: musical ability enhances specific aspects of foreign language perception. *Euro. J. Cogn. Psychol.* 22, 46–61. doi: 10.1080/09541440802708136

Ferguson, S. H. (2004). Talker differences in clear and conversational speech: vowel intelligibility for normal-hearing listeners. *J. Acoust. Soc. Am.* 116, 2365–2373. doi: 10.1121/1.1788730

Ferguson, S. H., and Kewley-Port, D. (2007). Talker differences in clear and conversational speech: acoustic characteristics of vowels.? *J. Speech Lang. Hear. Res.* 50, 1241–1255. doi: 10.1044/1092-4388(2007/087)

Fiedler, D., and Müllensiefen, D. (2015). Validation of the Gold-MSI questionnaire to measure musical sophistication of German students at secondary education schools. *Res. Music Educ.* 36, 199–219. [in German]

Francis, A. L., Ciocca, V., Ma, L., and Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *J. Phon.* 36, 268–294. doi: 10.1016/j.wocn.2007.06.005

Fu, Q. J., and Zeng, F. G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pac. J. Speech Lang. Hear.* 5, 45–57. doi: 10.1179/136132800807547582

Gagné, J. P., Masterson, V., Munhall, K. G., Bilida, N., and Querengesser, C. (1994). Across talker variability in auditory, visual, and audio-visual speech intelligibility for conversational and clear speech. *J. Acad. Rehabil. Audiol.* 27, 135–158.

Gaser, C., and Schlaug, G. (2003). Brain structures differ between musicians and non-musicians. *J. Neurosci.* 23, 9240–9245. doi: 10.1523/JNEUROSCI.23-27-09240.2003

Gottfried, T. L., and Riester, D. (2000). Relation of pitch glide perception and Mandarin tone identification. *J. Acoust. Soc. Am.* 108, 2604. doi: 10.1121/1.4743698

Gottfried, T. L., Staby, A. M., and Ziemer, C. J. (2004). Musical experience and Mandarin tone discrimination and imitation. *J. Acoust. Soc. Am.* 115, 2545–2545. doi: 10.1121/1.4783674

Grant, K. W., and Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *J. Acoust. Soc. Am.* 104, 2438–2450. doi: 10.1121/1.423751

Han, Y., Goudbeek, M., Mos, M., and Swerts, M. (2018). Effects of modality and speaking style on Mandarin tone identification by non-native listeners. *Phonetica* 76, 263–286. doi: 10.1159/000489174

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *J. Phon.* 40, 269–279. doi: 10.1016/j.wocn.2011.11.001

Hassler, M., and Gupta, D. (1993). Functional brain organization, handedness, and immune vulnerability in musicians and non-musicians. *Neuropsychologia* 31, 655–660. doi: 10.1016/0028-3932(93)90137-O

Hirata, Y., and Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *J. Speech Lang. Hear. Res.* 53, 298–310. doi: 10.1044/1092-4388(2009/08-0243)

Koelsch, S., Gunter, T., Friederici, A. D., and Schröger, E. (2000). Brain indices of music processing: "nonmusicians" are musical. *J. Cogn. Neurosci.* 12, 520–541. doi: 10.1162/089892900562183

Kricos, P. B., and Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *Volta Rev.* 84, 219–225.

Lee, C. Y., and Hung, T. H. (2008). Identification of Mandarin tones by English-speaking musicians and nonmusicians. *J. Acoust. Soc. Am.* 124, 3235–3248. doi: 10.1121/1.2990713

Maddieson, I., Dryer, M. S., and Haspelmath, M. (2013). *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Marie, C., Delogu, F., Lampis, G., Belardinelli, O., and Besson, M. (2011). Influence of musical expertise on segmental and tonal processing in Mandarin Chinese. *J. Cogn. Neurosci.* 23, 2701–2715. doi: 10.1162/jocn.2010.21585

Marques, C., Moreno, S., Luís Castro, S., and Besson, M. (2007). Musicians detect pitch violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence. *J. Cogn. Neurosci.* 19, 1453–1463. doi: 10.1162/jocn.2007.19.9.1453

Micheyl, C., Delhommeau, K., Perrot, X., and Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hear. Res.* 219, 36–47. doi: 10.1016/j.heares.2006.05.004

Milovanov, R., Huotilainen, M., Välimäki, V., Esquef, P. A. A., and Tervaniemi, M. (2008). Musical aptitude and second language pronunciation skills in school-aged children: neural and behavioral evidence. *Brain Res.* 1194, 81–89. doi: 10.1016/j.brainres.2007.11.042

Milovanov, R., Pietilä, P., Tervaniemi, M., and Esquef, P. A. A. (2010). Foreign language pronunciation skills and musical aptitude: a study of Finnish adults with higher education. *Learn. Individ. Diff.* 20, 56–60. doi: 10.1016/j.lindif.2009.11.003

Milovanov, R., and Tervaniemi, M. (2011). The interplay between musical and linguistic aptitudes: a review. *Front. Psychol.* 2:321. doi: 10.3389/fpsyg.2011.00321

Mixdorff, H., and Charnvivit, P. (2004). "Visual cues in Thai tone recognition," in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages* (Beijing).

Mixdorff, H., Charnvivit, P., and Burnham, D. (2005a). "Auditory–visual perception of syllabic tones in Thai," in *Proceedings of AVSP 2005, International Conference on Auditory–Visual Speech Processing,* eds E. Vatikiotis-Bateson, D. Burnham, and S. Fels (Adelaide, SA: Causal Productions), 3–8.

Mixdorff, H., Hu, Y., and Burnham, D. (2005b). "Visual cues in Mandarin tone perception," in *Proceedings of Eurospeech 2005 (InterSpeech-2005)* (Lisbon), 405–408.

Mok, P. P., and Zuo, D. (2012). The separation between music and speech: evidence from the perception of Cantonese tones. *J. Acoust. Soc. Am.* 132, 2711–2720. doi: 10.1121/1.4747010

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS ONE* 9:e89642. doi: 10.1371/journal.pone.0089642

Ong, J. H., Burnham, D., Escudero, P., and Stevens, C. J. (2017). Effect of linguistic and musical experience on distributional learning of nonnative lexical tones. *J. Speech Lang. Hear. Res.* 60, 2769–2780. doi: 10.1044/2016_JSLHR-S-16-0080

Patel, A. D. (2010). "Music, biological evolution, and the brain," in *Emerging Disciplines*, eds C. Levander and C. Henry (Houston, TX: Rice University Press), 91–144.

Schaal, N., Bauer, A.-K., and Müllensiefen, D. (2014). Der Gold-MSI: replikation und validierung eines fragebogeninstrumentes zur messung musikalischer erfahrenheit anhand einer deutschen stichprobe. [The Gold-MSI: replication and validation of a survey instrument for measuring musical sophistication with a German sample]. *Music. Sci.* 18, 423–447. doi: 10.1177/1029864914541851

Schön, D., Magne, C., and Besson, M. (2004). The music of speech: music training facilitates pitch processing in both music and language. *Psychophysiology* 41, 341–349. doi: 10.1111/1469-8986.00172.x

So, C. K., and Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: effects of native phonological and phonetic influences. *Lang. Speech* 53, 273–293. doi: 10.1177/0023830909357156

Sueyoshi, A., and Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Lang. Learn.* 55, 661–699. doi: 10.1111/j.0023-8333.2005.00320.x

Swerts, M., and Krahmer, E. (2008). Facial expression and prosodic prominence: effects of modality and facial area. *J. Phon.* 36, 219–238. doi: 10.1016/j.wocn.2007.05.001

Tseng, C. Y. (1981). *An acoustic phonetic study on tones in Mandarin Chinese* (Ph.D. dissertation). Brown University Press, Providence, RI, United States.

Vromans, R., and Postma-Nilsenová, M. (2016). "Can musical engagement alleviate age-related decline in inhibitory control?" in *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (Austin, TX).

Wong, P. C., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* 10, 420–422. doi: 10.1038/nn1872

Xu, Y., and Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *J. Acoust. Soc. Am.* 111, 1399–1413. doi: 10.1121/1.1445789

Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *J. Phon.* 30, 555–568. doi: 10.1006/jpho.2002.0165

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Zheng, A., Hirata, Y., and Kelly, S. D. (2018). Exploring the effects of imitating hand gestures and head nods on L1 and L2 Mandarin tone production. *J. Speech Lang. Hear. Res.* 1–17. doi: 10.1044/2018_JSLHR-S-17-0481

Zuccolotto, A. P., Roush, R. E., Eschma, A., and Schneide, W. (2012). *E-Prime [Computer Software]*. Pittsburgh, PA: University of Pittsburgh.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

TABLE A1 | List of words used for producing the stimuli.

mā	má	mǎ	mà
yi	yí	yi	yì
xiē	xié	xiě	xiè
shē	shé	shě	shè
shi	shí	shi	shì
you	yóu	you	yòu
fēn	fén	fěn	fèn
fu	fú	fu	fù
po	pó	po	pò
ying	yíng	ying	yìng

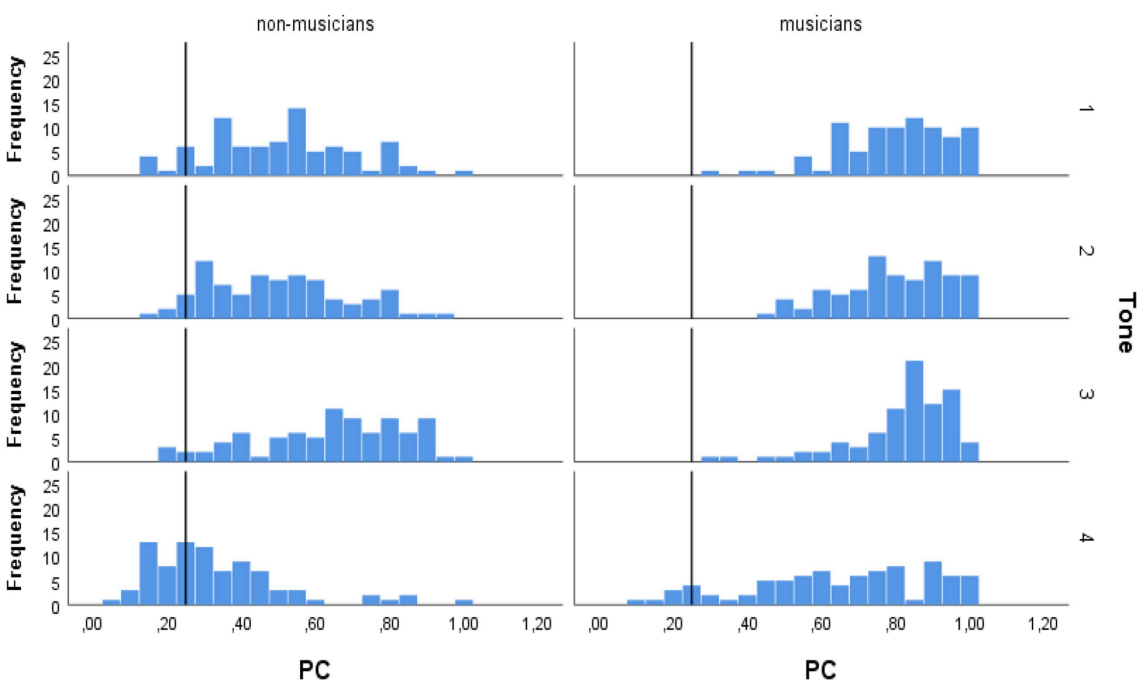

FIGURE A1 | A histogram of each participant's accuracy on individual tone level. PC, Proportion Correct. The chance level of the accuracy is 0.25.