The Perception of Lexical Tone and Intonation in Whispered Speech by Mandarin-speaking Congenital Amusics

Gaoyuan Zhang[1], Jing Shao[2], Caicai Zhang[3*], Lan Wang[4]

[1]Department of Chinese Language and Literature, Peking University

[2]Department of English Language and Literature, Hong Kong Baptist University, Hong Kong, SAR, China

[3]Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China

[4]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

*Corresponding author:

Caicai Zhang: Room EF741, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, 11 Yuk Choi Rd, Hung Hom, Hong Kong SAR, China.

Tel: (+852) 34008465. Email address: caicai.zhang@polyu.edu.hk.

Running head: Whispered speech perception in amusia

This work was supported by grants from the National Natural Science Foundation of China (NSFC: 11504400), the Research Grants Council of Hong Kong (ECS: 25603916), the PolyU Start-up Fund for New Recruits, and Shenzhen Fundamental Research Program (JCYJ20160429184226930; JCYJ20170413161611534).

Abstract

Purpose: A fundamental feature of human speech is variation, including the manner of phonation, as exemplified in the case of whispered speech. In the current study, we employed whispered speech to examine an unresolved issue about congenital amusia, a neurodevelopmental disorder of musical pitch processing, which also affects speech pitch processing like lexical tone and intonation perception. The controversy concerns whether amusia is a pitch-processing disorder or can affect speech processing beyond pitch.

Method: We examined lexical tone and intonation recognition in 19 Mandarin-speaking amusics and 19 matched controls in phonated and whispered speech, where fundamental frequency (F0) information is either present or absent.

Results: The results revealed that the performance of congenital amusics was inferior to that of controls in lexical tone identification in both phonated and whispered speech. These impairments were also detected in identifying intonation (statements/questions) in phonated and whispered modes. Across the experiments, regression models revealed that F0 and non-F0 (duration, intensity and formant frequency) acoustic cues predicted tone and intonation recognition in phonated speech, whereas non-F0 cues predicted tone and intonation recognition in whispered speech. There were significant differences between amusics and controls in the use of both F0 and non-F0 cues.

Conclusions: The results provided the first evidence that the impairments of amusics in lexical tone and intonation identification prevail into whispered speech, and support the hypothesis that the deficits of amusia extend beyond pitch processing.

Keywords: congenital amusia, lexical tone perception, intonation perception, whispered speech, Mandarin Chinese.

Introduction

Deficits of congenital amusia

Congenital amusia (amusia hereafter) is an innate neurodevelopmental disorder that affects fine-grained pitch (pitch is the perceptual correlate of F0) processing throughout the lifetime (Hyde & Peretz, 2003; Peretz et al., 2002, 2003, 2007). It is believed that individuals with amusia (amusics hereafter) have difficulties in detecting out-of-tune melodies and singing in tune although they have normal hearing, intelligence, sufficient exposure to music and no brain injury (Ayotte et al., 2002; Mignault Goulet et al., 2012; Peretz et al., 2002). It has been reported that amusics are not only impaired in musical pitch processing, but also in the processing of several dimensions in speech that rely on pitch, such as the perception of lexical tone, intonation (statement/question), and emotional prosody(Cheung et al., 2021; Huang et al., 2015a; Liu et al., 2010; Nan et al., 2010; Patel et al., 2008; Thompson et al., 2012) These results indicate that the pitch-processing deficit in amusia is not restricted to music, but transfers to the language domain (Vuvan et al., 2015).

However, an unresolved issue is whether amusics' inferior performance is restricted to pitch processing or not, and relatedly whether amusics are impaired in non-pitch processing in speech. Amusia, in its most common form, is thought to reflect impaired pitch processing, and therefore sometimes also called tone-deafness (Cousineau et al., 2015). Previous research has mainly focused on pitch processing(Albouy et al., 2013, 2015; Ayotte et al., 2002; Foxton, 2004; Peretz et al., 2002, 2005), revealing that amusics have reduced sensitivity to fine-grained pitch differences(Foxton, 2004; Liu et al., 2012), and inferior pitch memory (Gosselin et al., 2009; Tillmann et al., 2009; Williamson & Stewart, 2010). However, it is unclear whether amusics are deficient at processing other, non-pitch auditory cues, and if yes what cues are affected. A few studies have reported that amusics might have inferior durational, frequency, or amplitude processing abilities beyond pitch processing (Jones, Zalewski, et al., 2009; Lehmann et al., 2015;

Peretz & Vuvan, 2017; Phillips-Silver et al., 2011; Whiteford & Oxenham, 2017). Jones, Zalewski et al. (2009) reported that compared with the musically intact control group, the amusic group had poor performance in pitch, duration pattern discrimination and auditory gap detection tasks. Phillips-Silver et al. (2011) proposed a new form of congenial amusia, namely beat-deafness, based a group of individuals who had deficiencies in perceiving and producing the musical beat. Peretz & Vuvan, (2017) divided individuals into pitch-based amusics and time-based amusics (i.e., beat-deaf amusics) based on the criteria of performing below the cutoff score (2 SD below the mean) on two pitch-related tests (Scale and Off-key) and above the cutoff in the Off-beat test for the former, and performing above the cutoff score on the pitch-related tests and below the cutoff in the Off-beat test for the latter. The authors reported a prevalence rate of 1.5% for pitch-based amusics and 3.1% for time-based amusics. An important difference between these two subtypes of amusia is that time-based amusics are often associated with other developmental disorders such as dyscalculia and dyslexia, whereas the pitch-based form of amusia is believed to emerge in isolation, and relatively free of language problems (Peretz & Vuvan, 2017). But two recent studies reported association between pitch processing deficits and dyslexia (Couvignou et al., 2019; Couvignou & Kolinsky, 2021). Whiteford & Oxenham, (2017) reported that amusics were impaired in both low frequency (500Hz and 2000Hz) detection and high-frequency (8000Hz) detection that is beyond the frequency range of musical pitch, indicating that amusia is not a deficit selective to the musical attributes of pitch. They also assessed the participants' sensitivity to frequency modulation detection and amplitude modulation detection in a one-interval yes/no task and a standard two-alternative forced choice task. They found that the ability of amusics to detect frequency modulation and amplitude modulation was significantly poorer than that of controls in both tasks. Although the amplitude detection task did not involve any pitch-related changes, the amusics were still at a disadvantage, which questioned the general assumption that amusia is a pitch-specific deficit.

Regarding speech processing, some evidence indicates that amusics' impairments extend beyond pitch processing, affecting phonological awareness, segmental processing, or speech comprehension (Jiang et al., 2012; Jones, Lucker, et al., 2009; Liu et al., 2015; Sun et al., 2017; Zhang et al., 2017). It has been found that amusia may affect phonological processing such as phonological awareness (Jones, Lucker, et al., 2009; Sun et al., 2017), and phonological representations of lexical tones as revealed by categorical perception studies (Huang et al., 2015a; Jiang et al., 2012; Zhang et al., 2017). The results of Zhang et al. (2017) showed that Cantonese-speaking amusics had impairment in the discrimination of vowels in addition to pure tones and lexical tones, but their ability of voice-onset time (VOT) processing remained largely intact. Liu et al. (2015) examined the speech intelligibility of Mandarin-speaking amusics in perceiving Mandarin sentences with natural and flattened F0 curves in quiet and in noise. The authors found that Mandarin-speaking amusics had deficits in speech intelligibility for both natural-F0 and flat-F0 sentences regardless of noise, and their deficit in speech intelligibility was not associated with their pitch perception deficit. These results led the authors to argue that segmental processing might be impaired in amusics, independent of pitch processing. However, a problem with the previous studies is that these additional speech-processing deficits in amusia are likely to stem from their low-level auditory pitch deficit. For instance, a phonological deficit in categorical perception of lexical tones might originate from the low-level pitch-processing deficit in Chinese speakers with amusia from birth. The transfer of a low-level auditory deficit to a higher-level phonological deficit has been reported in other types of developmental disorders such as developmental dyslexia (Goswami et al., 2010, 2011). Even in the case of Liu et al. (2015), where pitch was flattened and neutralized, some pitch cues were still present. To investigate whether amusics are impaired in other areas of speech processing that do not involve pitch, it is thus necessary to employ speech conditions where pitch is absent, like in the case of whispered speech.

Perception of lexical tones in whispered speech

Compared with phonated speech, the most obvious characteristic of whispered speech is that the dominant perceptual cue – fundamental frequency (F0 hereafter) is absent. Previous studies have shown that compared with phonated tones, listeners have difficulty in identifying lexical tones in whispered speech, but the accuracy is above chance level, meaning that lexical tones can still be recognized to some extent (Gao, 2002; Jensen, 1958; Jiao et al., 2015; Jiao & Xu, 2019; Yang et al., 2005). For instance, Gao (2002) examined tone recognition in monosyllables in whispered Mandarin speech, and found that the rank of recognition accuracy for the four tones in the order of increasing difficulty is T3 (low dipping tone), T4 (high falling tone), T1 (high level tone) and T2 (high rising tone). Jiao et al. (2015) showed that the identification rates for phonated tones in Mandarin were: T1 (98.9%), T2 (98.9%), T4 (94.7%) and T3 (94.2%); in contrast, the rates for whispered tones were: T3 (85%), T4 (66.9%), T2 (35.8%) and T1 (21.7%).

Most research have found that there in fact is tone perception in whispered speech rather than contextually determined interpretations (Heeren, 2015). Therefore, there must be non-F0 cues in the speech signal that can compensate for the absence of F0 in whispered tone recognition (Jiao & Xu, 2019). Some researchers proposed some secondary cues, such as duration (Gao, 2002; Li & Guo, 2012; Li & Rong, 2012; Liu & Samuel, 2004; Yang et al., 2005), intensity contour (Gao, 2002; Li & Guo, 2012), and formant frequency (Eklund & Traunmüller, 1997; Higashikawa et al., 1996; Kallail & Emanuel, 1984; Li & Xu, 2005; Matsuda & Kasuya, 1999), that were exaggerated by the speakers to the needs of the listeners. For example, the duration of tones in whispered speech was much longer than that in phonated speech(Li & Guo, 2012; Li & Rong, 2012; Yang et al., 2005), and the duration differences across the four lexical tones were exaggerated in whispered Mandarin(Li & Guo, 2012). With regard to the intensity contour, compared with phonated speech, Gao (2002) found that speakers exaggerated the intensity contour when they were whispering, and concluded that intensity contour played a crucial role in whispered tone perception especially for Tone 3 and Tone 4. Higashikawa et al. (1996) did the formant analysis of the vowel /a/ with low, mid and high pitch in whispered and phonated

Japanese and found that the formants were higher overall in whispered speech than in phonated speech. However, there is no consensus on the question of whether non-F0 acoustic are exaggerated or not. Indeed, a few papers reported that non-F0 acoustic cues were not enhanced in whispered utterances. For example, Chang & Yao (2007) found that there were similar differences of duration and average intensity across the four tones in whispered and normal speech. Jiao & Xu (2019) examined duration, intensity, spectral tilt and formants of phonated tones and whispered tones respectively, and showed that there was no special articulatory manoeuver enhancement in whispered tones. Even if the non-F0 acoustic cues were not exaggerated in whispered utterances, they might carry certain acoustic distinctions to support tone identification.

To summarize, the majority of previous studies have focused on examining the enhancement of acoustic cues in whispered speech. Although there is some controversy regarding whether non-F0 acoustic cues are enhanced or not, such cues are believed to compensate for the lack of F0 and support tone recognition in whispered speech to some extent. However, few studies have directly examined the relationship between tone recognition performance and the distribution of non-F0 acoustic cues across the four tones. Furthermore, it remains unclear whether non-F0 acoustic cues are employed differentially by individuals with and without amusia among native Chinese speakers.

Perception of intonation in whispered speech

It is generally agreed that statements are characterized by a falling intonation, whereas questions are associated with a rising F0 contour, which conveys the meaning of non-finality and inquiry (Ohala, 1983). Studies on different languages have shown that questions are marked by acoustic characteristics such as a final increase in F0, higher F0 level, wider F0 range, longer syllable duration on the final position, and higher intensity level (Gussenhoven & Chen, 2000; Hirschberg & Ward, 1992; Ho, 1977; Ma et al., 2006).

As in the case of whispered tone perception, previous studies have shown that compared with phonated speech, listeners have difficulty in identifying boundary tone and intonation in whispered speech, but the accuracy is above chance level, meaning that boundary tone and intonation can still be recognized to some extent (Heeren & Heuven, 2009; Jiao & Xu, 2019). Some studies suggested that there is a compensatory strategy beneficial for perceiving boundary tone and intonation in whispered speech (Heeren, 2015; Heeren & Heuven, 2009). But it is not entirely clear what non-F0 acoustic cues facilitated intonation perception in whispered speech, and whether they are enhanced in the whispered mode (Jiao & Xu, 2019). Heeren & Heuven (2009) found that statements and questions could still be identified well above chance in whispered Dutch, and formant frequency and intensity differences that correlated with high and low boundary tones might be possible perceptual cues for linguistic intonation. Jiao & Xu (2019) investigated the production and perception of questions and statements in whispered Chinese, which showed that compared to whispered statements, whispered questions had flattened spectral slope. Yet they speculated that this acoustic cue did not assist the identification of questions since whispered questions were recognized much less accurately than statements. They concluded that there was no evidence of effective production enhancement for intonation identification in whispered Mandarin.

Again, a limitation of previous studies on whispered intonation is that few studies have directly examined the relationship between intonation perception performance and the distribution of non-F0 acoustic cues. Instead, the potential contribution of a certain acoustic cue to perception was inferred indirectly by examining whether there is exaggeration of that cue in production or whether it differs significantly between whispered statements and questions. While these analyses are certainly informative, it is necessary to examine whether the statement-question distance in one or a set of non-F0 acoustic cues can explain the perceptual performance of Mandarin listeners, and whether there are individual differences between listeners with high and low musical aptitude (i.e., musically intact controls and congenital amusics).

The current study

As mentioned above, whispered speech provides an ideal case to examine whether amusics have impairment in other aspects of the linguistic domain other than pitch processing. In addition, it remains unexplored to what extent the distribution of non-F0 acoustic cues in whispered speech can predict listeners' recognition performance of lexical tones and speech intonation respectively. Relatedly, no studies have examined before whether and if so how acoustic cues are differentially employed by listeners with amusia and intact musical abilities in the recognition of phonated and whispered tones and intonation.

To address the aforementioned questions, the current study compared the performance of Mandarin-speaking amusics and controls in recognizing lexical tones and speech intonation (statements/questions) in phonated and whispered speech through a series of identification tests. We then examined to what extent the participants' musical pitch, rhythm and memory abilities, as assessed by the Montreal Battery of Evaluation of Amusia (MBEA) (Peretz et al., 2003; Vuvan et al., 2018), can predict their tone and intonation recognition performance in phonated and whispered speech respectively using regression analyses. Regarding the contribution of acoustic cues to tone and intonation recognition in phonated and whispered speech, the following set of acoustic cues were selected and measured based on previous studies(Jiao et al., 2015; Jiao & Xu, 2019; Lima et al., 2016; Liu et al., 2012): phonated tone – F0, duration, intensity and formant frequency; whispered tone – duration, intensity and formant frequency; phonated intonation – F0, duration and intensity; whispered intonation – duration and intensity. We then examined which set of acoustic cues can predict the recognition performance of lexical tones and intonation in phonated and whispered speech in amusics and controls respectively using regression analyses.

There are two possible predictions. According to the hypothesis that amusia is primarily a pitch-processing disorder and relatively free of language problems(Ayotte et al., 2002; Hyde & Peretz, 2003; Peretz & Vuvan, 2017), the amusics would perform comparably to the controls in the recognition of lexical tones and intonation in whispered speech, where pitch is absent, and yet inferiorly in the case of phonated speech. Furthermore, amusics are less likely to rely on pitch cues and may tend to employ other acoustic cues on which they have no or less severe auditory deficits for lexical tone and intonation recognition in phonated and whispered speech. On the other hand, given the hypothesis that the amusics are impaired in speech processing beyond pitch (Jiang et al., 2012; Jones, Lucker, et al., 2009; Liu et al., 2015; Sun et al., 2017; Zhang et al., 2017), we expect amusics to perform worse than controls even in whispered speech, and the presence of non-F0 acoustic cues may not sufficiently enable them to compensate for their deficient pitch processing. Furthermore, fewer (F0 or non-F0) acoustic cues may explain the recognition performance in amusics.

Materials and Methods

Participants

Nineteen Mandarin-speaking amusics and 19 matched musically intact controls were recruited in this study. All participants were native speakers of Mandarin Chinese from northern China, right-handed and reported no previous history of speech, hearing, neurological or psychiatric impairments. No participants had any formal musical training. All the participants were selected by the MBEA, which consists of six subtests: scale, contour, interval, rhythm, meter and memory (Peretz et al., 2003). Among the six subtests, scale, contour and interval are pitch-based tests that concern melodic organization, rhythm and meter are duration-based tests that concern temporal organization, and memory concerns incidental memory of previous heard melodies. To separate amusics from controls, the pitch composite scores (sum of the number of correct trials in the three pitch-based subtests) were calculated, and those participants who scored at or below 65 were classified as amusics (Liu et al., 2012, Vuvan et al., 2018). Results of independent-samples t-tests confirmed that the amusics performed significantly worse than the controls on the pitch composite scores and the global score (percent of correct responses averaged across the six subtests), as well as on the percent of correct responses of each subtest (*p*s $< .001$). The demographic characteristics of the participants are shown in Table 1. The experimental procedures were approved by the Human Subjects Ethics committee of the Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. Informed written consent was obtained from the participants in compliance with the experiment protocols.

Stimuli

The speech stimuli were recorded by a 27-year-old female Mandarin speaker and a 28-year-old male Mandarin speaker, who are native speakers of Mandarin Chinese from northern China. All the word (tone) and sentence (intonation) stimuli are provided in supplementary Table 1. We recruited two speakers to record the speech stimuli in order to avoid the influence of idiosyncrasies of individual speakers, for the reason that the stimuli produced by some speakers might be easier to identify than those by other speakers according to a previous study (Gao, 2002). The two speakers were asked to produce the isolated words and sentences both in whispered and phonated mode as naturally as possible. The sound recording was done in a soundproof room using Praat (Boersma & Weenink, 2001), with 44.1 kHz sampling rate. We used ProsodyPro (Xu, 2013) to extract the acoustic parameters (F0, duration and intensity) of the stimuli. The measurements of F1 and F2 in both phonated and whispered speech followed the method described in (Sharifzadeh et al., 2012). The F1 and F2 were measured from the average of the steady portion where the formants were relatively clear and steady by a trained phonetician.

To assess tone identification in Mandarin Chinese, 36 words with nine base (C)V syllables (/ta/, /ti/, /tu/, /pa/, /pi/, /tʃu/, /a/, /i/, /u/) contrasting the four lexical tones were selected as the stimuli (supplementary Table 1). Each word was produced three times by each speaker under phonated and whispered conditions. Overall, there were a total of 432 tokens (9 syllables × 4 tones × 2 phonation modes × 3 repetitions × 2 speakers). From the three repetitions, one token with accurate and clear pronunciation was selected and used as the stimuli in the ensuing perception tasks, totalling 144 tokens. Figure 1 shows the time-normalized F0 contours of the four Mandarin tones from the two speakers, averaged across all the syllables which contained the same tones in the phonated stimulus set. We measured F0 mean, F0 SD, duration, mean intensity, F1 and F2 of the phonated stimuli produced by the two speakers. A series of one-way ANOVAs with the factor of *lexical tone* were conducted on each acoustic cue, with the *p*-value corrected for multiple comparisons (.05/6 = .008). There were significant differences between the four phonated tones in F0 mean, F0 SD, duration, and intensity ($ps \leq .05$), but no significant effect in F1 or F2 (see supplementary Table 2 for the acoustic cues of the four tones and statistical results). The post hoc analyses found that every tone was significantly different from each other in F0 mean (T1 > T4 > T2 > T3, $ps < .001$) and F0 SD (T4 > T2 > T3 >T1, $ps \leq .003$); for duration, the results fell into the pattern of T3 > T1 ≈ T2 > T4 ($ps \leq 0.003$); for intensity, T3 was significantly lower than T1 and T4 ($ps \leq .03$). We also measured duration, mean intensity, F1 and F2 of the whispered stimuli produced by the two speakers. The results of one-way ANOVA with the factor of *lexical tone* with correction for multiple comparisons (.05/4 = .0125) revealed that there was a significant difference between the four whispered tones in duration ($p < .001$), but not in intensity, F1 or F2 (see supplementary Table 3 for the acoustic cues of the four whispered tones and statistical results). The post hoc analyses found that the duration of T3 was significantly longer than the other three tones, while the duration of T4 was significantly shorter than the other three tones ($ps \leq .001$).

To assess intonation identification in Mandarin Chinese, 25 statement-question pairs sharing the same words were constructed as the stimuli (see supplementary Table 1 for the sentences). Five sentence lengths were included (4, 5, 6, 7 and 10 syllables), with five sentences for each sentence length. For each length, four out of the five sentences contained words with identical tones on every position (e.g., the sentence ‘张薇开车’ consisted of only T1) and the last sentence contained words with varied tones (e.g., the sentence ‘李刚讲课’ consisted of varied tones), which ensured that all four tones appeared on every position for each sentence length to avoid any potential influence of lexical tones on the intonation F0 patterns (Yan, 2016). Each sentence was produced twice by each speaker under the phonated and whispered conditions. Overall there were a total of 400 tokens (5 sentences × 2 intonations × 5 length types × 2 phonation modes × 2 repetitions × 2 speakers). From the two repetitions, one token with accurate and clear pronunciation was selected and used as stimuli in the ensuing perception tasks, totalling 200 tokens. Figure 2 displays the real-time F0 contours of one pair of statement and question produced by the male speaker. As can be seen, the differences between statements and questions were not only present in the F0, but also in the total sentence duration. For phonated speech, we tested the acoustic characteristics (F0, duration and intensity) of the whole sentences and their final syllables produced by the two speakers which followed the method in Liu et al. (2012) and Lima et al. (2016). Paired-samples *t*-tests with the factor of *intonation* (corrected *p*-value at .05/9 = .006) indicated that significant intonation differences were detected on all acoustic cues of the whole sentences, and on the F0 mean and intensity of the final syllables ($ps < .001$) (see supplementary Table 4 for the acoustic cues of statements and questions and statistical results). For whispered speech, we measured the acoustic cues (mean intensity and duration) of the whole sentences and their final syllables produced by the two speakers, which echoes acoustic characteristics in the phonated mode. Paired-samples *t*-tests with the factor of *intonation* (corrected *p*-value at .05/4 = .0125) indicated that statements had significantly longer overall duration than questions ($p < .001$), but significantly shorter duration on the final syllable ($p = .001$) (see supplementary Table 5 for the acoustic cues of statements and questions and statistical results).

Procedure

The study included two tasks: lexical tone identification and intonation identification, both of which were implemented using E-prime 2.0. In both tasks, the two phonation modes (phonated and whispered speech) were presented in separate blocks. The stimuli from the two speakers were also presented in two separate sub-blocks in order to avoid the effect of talker variation. The order of these two tasks (the lexical tone identification task and the intonation identification task) was counterbalanced across the participants. Furthermore, half of the participants completed the phonated speech block first, and the other half completed the whispered speech block first. The presentation order of the two sub-blocks (two speakers) within each task was also counterbalanced across the participants.

In the lexical tone identification task, the stimuli were presented three times, resulting in a total of 216 trials (9 syllables × 4 tones × 3 repetitions × 2 speakers) in each of the two phonation modes. Within each sub-block, all the trials were presented randomly. In each trial, a fixation occurred at first for 500ms, followed by the presentation of a spoken stimulus via the headphones. The participants were asked to identify the tone of the stimulus by pressing buttons 1-4 referring to the four lexical tones in Mandarin on a computer keyboard. The experiment only proceeded to the next trial when a response was received. Participants were instructed to respond as quickly as possible. There were practices before each task to familiarize the participants with the experimental procedure. The practice contained the three syllables /a, i, u/ with the four tones in both phonation modes presented only once in random order. Half of the participants practiced on the stimuli produced by the female speaker in the phonation mode and the stimuli produced by the male speaker in the whispered mode; this was reversed in the other half of the participants.

As for the intonation identification task, the stimuli were presented twice, resulting in a total of 200 trials (5 sentences × 2 intonations × 5 length types × 2 repetition × 2 speakers) in each of the two phonation modes. The buttons “Q” and “S” which refer to questions and statements respectively were response buttons on a computer keyboard. The other aspects were the same as those in the tone identification task. There were also practices before each task to familiarize the participants with the experimental procedure. The practice contained ten trials comprising the 10-sllyable sentences with the two intonation patterns in both phonation modes produced by the two speakers, which were randomly presented once.

Data analysis

For the tone identification task, accuracy was recorded and analysed. Accuracy was the percentage of trials correctly identified for each tone per subject. For the intonation identification task, performance was scored as the sensitivity index d' (Macmillan & Creelman, 2005). We used the sensitivity index d' to analyse the intonation identification data for the reason that Jiao & Xu (2019) indicated that statement is likely to be treated as a default choice when identification was challenging and as a result the signal detection method allows us to avoid any response bias (Irwin et al., 1992). The d' was computed as the z-score of the hit rate minus that of the false alarm rate for each phonation mode per subject. Specifically, the hit rate was the rate of "question" responses to the "question" test items, while the false alarm rate was the rate of "question" responses to the "statement" test items. *Group* × *lexical tone* × *phonation mode* repeated-measures ANOVA was conducted on the accuracy of the tone identification task. *Group* × *phonation mode* repeated-measures ANOVA was conducted on the d' of the intonation identification task.

Two sets of regression analyses were conducted to examine (1) to what extent the participants' musical pitch, duration and memory abilities can explain their recognition performance in phonated and whispered speech respectively, and (2) what acoustic cues can explain the listeners' recognition performance in phonated and whispered speech respectively, and whether different cues were employed by amusics and controls. For the first set of analyses, multiple linear regression models were constructed on the average accuracy of tone identification (averaged across four tones) for the phonated and whispered mode separately, collapsing amusics and controls, with melodic organization, temporal organization and melodic memory as three predictors. Similar regression analysis was conducted on the d' score of intonation identification for the phonated and whispered mode separately. To keep the set of predictors small and to avoid collinearity, we combined the six MBEA subtests into three sets: melodic organization (which is the average accuracy of the three subtests: scale, contour and interval), temporal organization (which is the average accuracy of the two subtests: rhyme and meter), and melodic memory (Peretz et al., 2003). Prior to the regression models, we conducted bivariate Pearson correlations (two-tailed) to estimate the degree of association between the three sets of musical abilities and tone/intonation identification performance, and between the three sets of musical abilities themselves. Only musical abilities that showed significant correlation with tone/intonation identification performance were then included in the linear regression models in a stepwise manner to determine their relative contribution[1].

For the second set of analysis, we employed logistic regression to examine the relationship between the acoustic cues for tone and intonation types and the two groups' responses separately for the phonated and whispered mode. Multinomial logistic regression analyses were conducted for tones and binominal logistic regression analyses were conducted for intonations. For the phonated mode, the following acoustic cues were included: F0 mean, F0 SD, duration, intensity, F1 and F2. For the whispered mode, the following acoustic cues were included: duration, intensity, F1 and F2. Accuracy instead of d' was used as the dependent variable in the regression analyses on intonation perception because the difference of statements and questions was collapsed in the d' score. For each regression model, we ensured that the VIF value was below 5 (Zhang & Dong, 2004) to avoid collinearity. In the phonated mode, we excluded the F0 mean and intensity of the final syllables since they were highly correlated with the F0 mean and intensity of the whole sentences respectively (r > 0.8), in order to reduce the VIF to be below 5. The details of the correlation analyses are presented in supplementary Table 6.

Results

Lexical tone identification task

Figure 3 shows the tone identification accuracy under phonated and whispered conditions for the two groups of participants. There was a significant main effect of *group* (F (1, 36) = 8.79, p = .005, $\eta^2_{partial} = 0.20$), where the score of the control group (M = 0.81, SD = 0.016) was significantly higher than that of the amusic group (M = 0.75, SD = 0.016). The *group* factor did not interact with the other two factors (*lexical tone* and *phonation mode*). Significant main effects of *phonation mode* (F (1, 36) = 308.68, p < .001, $\eta^2_{partial} = 0.90$), *lexical tone* (F (2.49, 89.93) = 58.09, p < .001, $\eta^2_{partial} = 0.62$), and a significant interaction between *lexical tone* and *phonation mode* (F (2.63, 94.57) = 67.98, p < .001, $\eta^2_{partial} = 0.65$) were also detected. To explore the two-way interaction, we first conducted independent-samples t-tests to examine the effect of *phonation mode* within each lexical tone. A significant effect of *phonation mode* was observed in all lexical tones (ps ≤ .01), where the accuracy of the phonated mode was always higher than that of the whispered mode. Then one-way ANOVAs with the factor of *lexical tone* within each phonated mode were conducted, revealing a significant effect of *lexical tone* in both phonation modes (ps ≤ .02). The post hoc analyses revealed that the rank of identification accuracy of the four lexical tones were different in the two phonation modes. For the phonated mode, the accuracy rank of the four tones was T4 ≈ T1 ≈ T3 > T2, where the accuracy of T2 was significantly lower than the other three tones (ps < .05). Nonetheless, the accuracy rank of the four tones was T3 > T4 > T1 ≈ T2 in the whispered mode. The accuracy for T3 was significantly higher than the other three tones (ps < .001) and the accuracy of T4 was also significantly higher than T2 and T1 (ps < .001) in the whispered mode.

Confusion matrixes across the tones were constructed for each phonation mode. The details of confusion matrixes are presented in supplementary Table 7 and 8. All the tones were recognized above chance level. In the phonated mode, the confusion pattern of controls was roughly similar to that of amusics, that is, T1 and T2 were to some extent confused with each other, and T3 was more often confused with T2, whereas T4 exhibited no clear confusion bias. In the whispered mode, the confusion pattern differed from that of the phonated mode, but controls and amusics exhibited roughly similar confusion patterns except for T1. For both controls and amusics, they were likely to confuse T2 with T3, and to a less extent also confuse T3 with T2, and T4 was more often confused with T1.

Intonation identification task

Figure 4 shows the d' of intonation identification under phonated and whispered conditions for the two groups. There was a significant main effect of *group* (F (1, 36) = 13.11, p = .001, $\eta^2_{partial} = 0.27$), where the d' score of the control group (M = 1.91, SD = 1.1) was significantly higher than that of the amusic group (M = 1.48, SD = 1.05). There was a significant main effect of *phonation mode* (F (1, 36) = 855.84, p < .001, $\eta^2_{partial} = 0.96$)), where the d' score of the phonated mode (M = 2.67, SD = 0.53) was significantly higher than that of the whispered mode (M = 0.72, SD = 0.39). The interaction between *group* and *phonation mode* was not significant.

As for the confusion matrix in the phonated mode (see supplementary Table 9), the controls showed comparable identification accuracy for statements and questions, whereas there was a clear decline in the identification accuracy for questions in amusics. In whispered utterances (see supplementary Table 9), statements were identified with over 80% accuracy in both groups, whereas questions were recognized at around chance-level in the control group and below chance-level in the amusic group. These observations further confirm that statement is likely to be treated as a default choice by listeners (Jiao & Xu, 2019).

Regression analyses

Two sets of regression analyses were conducted. The first linear analysis aimed to test whether the participants' musical pitch (melodic organization), rhythm (temporal organization), and melodic memory abilities can account for their performance in the two phonation modes. The bivariate correlations showed that all three musical abilities were significantly correlated with the participants' performance on the four identification tests (ps $\leq .02$), and thus were all included in the regression models (see Table 2 and supplementary Figure 1). The stepwise regression models were significant in tone and intonation identification for both phonation modes (ps $\leq .001$) (see Table 3). In the phonated mode, temporal organization was a significant predictor for tone identification, and melodic memory was a significant predictor for intonation identification (ps $< .001$), with increased scores of the two predictors contributing to higher accuracy of tone and intonation identification. The pattern was different in the whispered mode. Melodic memory was a significant predictor for both whispered tone and intonation identification (ps $\leq .001$), with increased scores of melodic memory contributing to higher accuracy of whispered tone and intonation identification.

The second analysis concerns whether the acoustic cues can account for the participants' responses, and whether amusics and controls employed these acoustic cues in a different manner in their perception. We conducted multinomial logistical regression for tone identification and binominal regression for intonation identification. Tables 4-7 show the main findings. The figures are displayed in supplementary materials (supplementary Figure 2-5). For phonated tone (Table 4) and intonation (Table 5), the models were significant in both cases (ps $< .001$). The Nagelkerke R^2 of the estimate of tone and intonation identification was 0.91 and 0.81 respectively. Note that in the tone model, the second tone in each pair was used as the baseline for the contrast (e.g., T1 was the baseline in the T2 vs. T1 pair), and controls were used as the baseline for comparison with amusics. We found that F0 mean, F0 SD, duration, intensity, F1 and F2 significantly predicted the identification of almost all tone pairs (ps ≤ .003). Importantly, non-F0 cues were also significant predictors, which indicated that F0 is not the only acoustic cue that can differentiate the four tones. The group factor was significant only in the tone pairs including T1 (ps ≤ .04). There were multiple significant interactions between group and acoustic cues, but the specific patterns varied across the tone pairs. For instance, the interaction between group and F0 mean was significant in all tone pairs except for T4-T1 (ps ≤ .003). Note that the F0 mean of T2 (vs. T1) and T3 (vs. T1 and T2) is *lower* than the baseline tone in each of these contrasts (see supplementary Table 2). The *positive* coefficients in these contrasts suggested that an increase in F0 mean was *more likely* to lead to the identification of T2 and T3 (i.e., the *wrong* tones) in these contrasts by the amusics compared to the controls, which implied that the amusics may have used F0 mean less efficiently in the identification of these tones. Similarly, the F0 mean of T4 (vs. T2 and T3) is *higher* than the baseline tone in each of these contrasts (see supplementary Table 2). The *negative* coefficients in these contrasts suggested that an increase in F0 mean is *less likely* to lead to the identification of T4 (i.e., the *correct* tone) in these contrasts by the amusics, which also implied worse usage of F0 mean in these tone contrasts by the amusics compared to the controls. For easy reference, cases where a certain acoustic cue led to worse or better performance in amusics were marked differentially in Table 4. Overall, the results showed that amusics employed F0 mean, F0 SD, duration and intensity worse than controls, but employed F1 and F2 better than controls in various tone contrasts.

In the phonated intonation model, we found that sentence F0 mean, sentence F0 SD, sentence F0 direction, sentence intensity, final syllable F0 SD, final syllable duration and group significantly predicted intonation identification accuracy (ps ≤ .01). Again, several non-F0 cues were significant predictors in the model, which implied that non-F0 acoustic cues are likely to contribute to phonated intonation identification. The interactions between group and sentence F0 mean, sentence F0 SD, sentence F0 direction, sentence intensity and final syllable duration were also significant ($ps < .001$). Note that statement was used as the baseline for contrast with question, and controls were used as the baseline for contrast with amusics. The value of question on each of the aforementioned acoustic cue was *greater* than that of statement (see Supplementary Table 4). The negative coefficients in these significant interactions indicate that as the acoustic value increases, it is less likely for the amusics to choose questions (i.e., the correct intonation) over statements compared to the controls.

As for the whispered mode, the models were significant for tone (Table 6) and intonation (Table 7) ($ps < .001$). The Nagelkerke R^2 of the estimate of tone and intonation identification was 0.38 and 0.09 respectively. In the whispered tone model, we found that duration, intensity, F1 and F2 significantly predicted the identification of almost all tone pairs ($ps \leq .02$). The interactions between group and duration, F1 and F2 were significant in several tone pairs ($ps \leq .04$). Overall, the amusics employed duration worse than the controls in the T2-T1, T4-T1, T3-T2, T4-T3 pairs; they also employed F1 worse in the T3-T2 pair, and F2 worse in the T4-T1 pair. In contrast, the amusics employed F2 better than the controls in the T2-T1 and T3-T1 pairs.

In the whispered intonation model, we found that sentence duration and final syllable duration significantly predicted intonation identification ($ps < .001$). The interaction between group and sentence duration was also significant in the model ($p = .02$). Since the sentence duration was longer in statements than in questions, the significant interaction (with positive coefficient) indicates that when the sentence duration increases, it is more likely for the amusics to identify the intonation pattern as question (i.e., the wrong intonation) compared to the controls.

Discussion

The present study examined the identification of lexical tone and intonation by Mandarin-speaking amusics and controls in the phonated and whispered modes. In the lexical tone identification task, the results showed that Mandarin-speaking amusics demonstrated an overall lower accuracy compared with the controls regardless of the phonated and whispered modes. Likewise, the amusics performed inferiorly with respect to the controls in terms of the d' scores in the intonation identification task in both modes. These results indicated that the impairment of amusics extends to tone and intonation identification in whispered speech, where the F0 is absent, implying that amusics are likely to be impaired in other aspects of speech processing other than pitch. In the text below, we discussed the results of the current study in relation to the two research questions raised in the introduction: (1) whether amusia is a pitch-processing disorder or whether it affects other aspects of the linguistic domain beyond pitch processing; (2) what acoustic cues other than the F0 can predict the listeners' recognition performance of lexical tones and speech intonation in the phonated and whispered mode respectively, and whether different acoustic cues were employed by amusics and controls.

Impairment of amusics in tone and intonation identification in the phonated and whispered modes

In the phonated mode, the accuracy of the amusic group was significantly lower than the control group, for both lexical tone identification and intonation identification. The inferior performance of amusics in tone identification is in line with the results of several studies on tonal language speakers with amusia (Nan et al., 2010; Shao et al., 2019, 2016; Shao & Zhang, 2020; Wang & Peng, 2014; Zhang et al., 2018), which have suggested that amusia is a domain-general disorder rather than being restricted to the musical domain (Douglas & Bilkey, 2007; Patel et al., 2008; Thompson, 2007; Vuvan et al., 2015). Furthermore, we found that the accuracy of most tones was above 90% except for T2 in the amusic group, which echoed with the results in (Nan et al., 2010) that T2 was the most difficult tone to identify for Mandarin-speaking amusics, especially for those with lexical tone agnosia, who were markedly impaired in lexical tone perception (i.e., with scores below 3 SD of the controls' scores in tone identification and discrimination). A plausible explanation is that similar acoustical characteristics shared by T2 and T3 may exacerbate the confusion (Nan et al., 2010), which is evidenced by the high confusion rate of T2 with T3 (and to some extent with T1) in the confusion matrix in the current study (see supplementary Table 7). On the other hand, this result differed from Liu et al. (2012), who did not find differences between the Mandarin-speaking amusics and controls in the tone identification task, but the discrepancy can be attributed to task differences. Since Liu and colleagues asked the participants to recognize the lexical tone stimuli as words using Chinese characters, instead of as tonal categories, it is likely that their task involved less abstract phonological processing, and as a result did not reveal the group difference.

With regard to intonation identification in the phonated mode, compared with controls, the significantly lower accuracy of the amusic group is consistent with the results in Jiang et al. (2010) and Liu et al. (2010), but not with Liu et al. (2012). Liu et al. (2012) have found that Mandarin-speaking amusics performed as well as controls on intonation identification in natural speech. Material differences may explain the somewhat different results of these two studies. The average pith range of statements and questions was 10.49 and 9.59 semitones respectively in the current study (statement-question difference: 0.9 semitones), whereas that of statements and questions was 11.35 and 7.75 semitones in Liu et al. (2012) (statement-question difference: 3.6 semitones). The average pitch excursion of the final syllable of statements and questions was 6.88 and 5.83 semitones respectively in the current study (statement-question difference: 1.05 semitones), whereas that of statements and questions was 6.82 and 3.59 semitones in Liu et al. (2012) (statement-question difference: 3.23 semitones). It is clear from the comparison above that the phonated stimuli contained smaller differences between statements and questions in the current study than those in Liu et al. (2012). Larger differences may enable amusics to distinguish and identify statements and questions, therefore showing comparable performance with controls.

In support of this argument, previous studies have revealed that amusics' average threshold for discriminating pitch direction is around two semitones (Foxton, 2004; Liu et al., 2010). Thus naturally produced smaller pitch contrasts in the current study may be more sensitive in revealing the amusics' intonation processing deficit.

In the whispered mode, amusics again had worse performance than controls in both lexical tone and intonation identification tasks, which is consistent with the previous findings that the amusics' impairments extend beyond pitch processing, affecting phonological awareness, segmental processing or speech comprehension (Jiang et al., 2012; Jones, Lucker, et al., 2009; Liu et al., 2015; Sun et al., 2017; Zhang et al., 2017). It also complements the previous findings by further providing evidence that amusics are impaired in lexical tone and intonation identification even when the F0 is absent. Altogether, there is convergent evidence for the notion that the deficits of amusia exist outside of pitch processing. That being said, these results do not necessarily negate the hypothesis that amusia is a pitch-processing disorder, because amusics are indeed impaired in lexical tone and intonation identification in the phonated mode. We will return to the discussion of the deficits of amusics in non-pitch processing in speech after discussing the contribution of acoustic cues to lexical tone and intonation identification in phonated and whispered speech first below.

Contribution of acoustic cues to phonated and whispered tone and intonation identification

It remains controversial which non-F0 acoustic cues facilitate tone and intonation identification where the F0 is absent. Previous studies have probed this question from the perspective of the enhancement of other acoustic cues (e.g., duration, intensity and formant frequency) in whispered speech compared to phonated speech, but no consensus has been reached on this issue. Furthermore, few previous studies have directly examined the relationship between acoustic cues and the participants' identification performance.

The current study filled this gap and generated new results from the regression analyses. We found that duration, intensity, F1 and F2 were significant predictors for whispered tone identification. These results corroborated with the proposals that duration (Jiao & Xu, 2019; Yang et al., 2005), mean intensity (Jiao & Xu, 2019; Li & Guo, 2012), and formant frequency (Heeren & Heuven, 2009; Li & Xu, 2005) facilitate whispered tone recognition. Intriguingly, duration, intensity, F1 and F2 were also significant predictors in *phonated* tone identification, which indicates that F0 is not the only acoustic cue that contributes to tone identification, although it is a dominant one, as evidenced by the drop in tone identification accuracy from the phonated mode to the whispered mode. As for whispered intonation identification, the results of regression analyses indicated that sentence duration and final syllable duration were significant acoustic predictors.

Crucially, we found significant interactions between group and several acoustic cues in the regression models on tone and intonation identification in phonated and whispered speech, which indicates that there were relative weaknesses in the usage of various acoustic cues by the amusics. In phonated tone identification, the amusics not only employed F0 cues (F0 mean and SD) less efficiently than the controls, but also exhibited worse usage of duration and intensity in almost all tone contrasts. On the other hand, the amusics appeared to have employed F1 and F2 cues better than the controls in almost all tone contrasts. But it is worth noting that the acoustic distinction in F1 and F2 among the four tones was relatively small and not significant (whereas there were significant tone differences in F0 mean, F0 SD, duration and intensity; see supplementary Table 2), which implies that F1 and F2 may not be the most optimal cues to employ in phonated tone identification. In whispered tone identification, where the F0 was absent, non-F0 acoustic cues including duration, intensity, F1 and F2 presumably played a greater role. Here, the amusics not only employed duration cues less efficiently compared to the controls, but also demonstrated worse performance in the usage of F1 and F2 in several tone pairs; they only used F2 more efficiently than the controls in the T2-T1 and T3-T1 pairs. Likewise, in phonated intonation identification, the amusics not only employed F0 cues less efficiently compared to the controls, but also exhibited a weakness in the usage of sentence intensity and final syllable duration cues. Where the F0 cues were absent, duration cues (sentence and final syllable duration) significantly predicted intonation identification, and the amusics continued to show inferior performance in the usage of sentence duration relative to the controls. Taken together, these observations appeared to suggest that in speech signals (phonated or whispered) with rich acoustic redundancies where multiple acoustic cues index a functional contrast (e.g., statement vs. question or the four tones), the amusics may not employ the most optimal acoustic cues for the contrast or use them less efficiently compared to the controls.

How to explain the worse performance of amusics in whispered tone and intonation identification in the current study? As the regression analyses revealed weaknesses in the amusics' usage of duration and to some extent formant frequency cues in whispered speech, a most straightforward explanation is that the amusic participants recruited in this study may have impaired duration or formant frequency processing. This explanation is compatible with the various findings that amusics have inferior durational, frequency, or intensity processing abilities beyond pitch processing (Jones et al., 2009; Lehmann et al., 2015; Peretz & Vuvan, 2017; Phillips-Silver et al., 2011; Whiteford & Oxenham, 2017). Future studies should directly examine fine-grained duration and formant frequency processing (e.g., using threshold tasks) together with whispered speech perception in amusics in a single study, so as to further reveal which sub-domain of acoustic processing best explains the performance of amusics in whispered speech perception.

An alternative explanation is that the phonological representations of lexical tone and intonation are impaired in amusics. Several previous studies have indicated that Chinese speakers with amusia are impaired in the phonological representation of lexical tones (Huang et al., 2015a; Jiang et al., 2012; Zhang et al., 2017) . For instance, Jiang et al. (2012) examined the performance of amusics and controls in the categorical perception of lexical tone, and found that amusics performed less categorically, exhibiting less between-category benefit than the controls, which suggested that there was a deficit of higher-level phonological processing of lexical tones of the amusic group. According to this view, regardless of whether amusics are deficient in earlier auditory processing of pitch and non-pitch cues, when the acoustic cues are mapped onto the phonological representation of lexical tones in the categorization process, an impairment in amusics is detected, even in the case of whispered speech.

Finally, although it is not our primary interest, the finding that temporal organization significantly predicted phonated tone identification, and melodic memory significantly predicted whispered tone, phonated intonation and whispered intonation identification requires an explanation. It is unexpected that melodic organization, which is related to pitch processing, was not a significant predictor of phonated tone or intonation identification, where F0 was a dominant cue. It may be because melodic organization, temporal organization and melodic memory are highly correlated with each other (see Table 2), and temporal organization or melodic memory may be able to explain more unique variances than melodic organization in these models. Another explanation is that the three pitch-based MBEA subtests do not purely assess pitch processing, but also involve musical knowledge. In either case, the finding that temporal organization was a significant predictor of phonated tone identification reinforces the view above that acoustic duration cues contributed to phonated tone identification. The contribution of melodic memory to intonation identification (phonated or whispered) may be because the intonation tasks used long sentence materials, where memory capacities of the melodic patterns are crucial for the identification performance. However, it is not entirely clear why melodic memory also significantly predicted whispered tone identification. That being said, all these results must be replicated in future studies for more rigorous interpretation.

To conclude, we found that Mandarin-speaking amusics showed degraded performance of lexical tone and intonation identification in both phonated and whispered modes compared to musically intact listeners. The results indicated that although only around 7% of amusics selfreport that they have difficulties in understanding other people's speech in daily life (Liu et al., 2015), their deficits affected phonated and whispered lexical tone and intonation processing in the laboratory. The results of the current study are consistent with the hypothesis that the impairment of amusia is domain general, rather than limited to the musical domain. Moreover, our findings indicate that the impairment is not confined to pitch processing, but extend to other aspects beyond pitch processing (Jones et al., 2009; Jones, Lucker, et al., 2009; Lehmann et al., 2015; Liu et al., 2015; Whiteford & Oxenham, 2017; Zhang et al., 2017). It is likely that amusia is a syndromic disorder frequently accompanied by deficiencies of other kinds (Jones et al., 2009; Jones, Lucker, et al., 2009). This study is the first to examine whispered speech perception in amusics, and revealed that amusics have impairments in other aspects of the linguistic domain, which sheds further light on the nature of the deficits underlying amusia. These findings also have real-world implications for the diagnosis and treatment of amusia. However, there are some remaining issues to be addressed in future studies. First, future studies with a large sample of amusics should separate them into pitch- and time-based forms of amusia (Peretz & Vuvan, 2017) and further examine if there are subgroup differences in phonated and whispered tone and intonation perception. It should be noted that the MBEA temporal organization subtests are complex tasks that assess more than duration processing. It is recommended that future studies use tasks that probe into duration processing (e.g., duration threshold tasks) to examine whether time-based amusics truly have duration (and intensity) processing deficits and how these problems contribute to their perception in phonated and whispered speech. Second, and related to the first point, future large-scale studies may separate Mandarin-speaking amusics into lexical tone agnostics and those without severe tone perception deficits (Nan et al., 2010; Huang et al., 2015a; Huang et al., 2015b; Nan et al., 2016), and examine if there are subgroup differences in phonated and whispered tone and intonation perception. Future studies should also investigate whether the finding of the current study generalizes to amusic individuals in other tonal languages (e.g., Cantonese) or non-tonal language (e.g., English), and with different tasks (e.g., discrimination task). It will also be of interest to use event-related potentials (ERPs) to probe passive and active processing of whispered speech in amusics and examine if there are any processing differences between the two listening conditions (Moreau et al., 2003; Zhang & Shao, 2018).

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (NSFC: 11504400; http://www.nsfc.gov.cn/) and the Research Grants Council of Hong Kong (ECS: 25603916; https://www.ugc.edu.hk/eng/rgc/) to CZ. This work was also supported by the grant of the National Key R & D Program of China (2020YFC2004100) to LW. We thank Mr. Yulin Wen for help with data collection. We thank Dr. Xiaocong Chen for help with statistical analysis.

Footnote

[1]We conducted simultaneous linear regression and reported the results in the supplementary materials (see supplementary Table10). Nonetheless, the results of simultaneous regression generated some puzzling results. Although the correlations between the scores of melodic organization and identification performance are positive (see Table 2), the standardized coefficients for melodic organization turned out to be negative in the regression models (although not significant). It might be because the correlation among the three MBEA composites are very high ($ps < .001$). That is, the multicollinearity of the independent variables is strong. As a result, we entered the predictors into the regression models in a stepwise manner. The significant predictors in the stepwise regression models were almost identical to those in the simultaneous model, except that melodic memory was a significant predictor in whispered intonation

identification in the stepwise model (but not in the simultaneous model). Forward and backward models generated the same results as the stepwise regression models.

References

Albouy, P., Mattout, J., Bouet, R., Maby, E., Sanchez, G., Aguera, P.-E., Daligault, S., Delpuech, C., Bertrand, O., Caclin, A., & Tillmann, B. (2013). Impaired pitch perception and memory in congenital amusia: The deficit starts in the auditory cortex. *Brain*, *136*(5), 1639–1661. https://doi.org/10.1093/brain/awt082

Albouy, P., Mattout, J., Sanchez, G., Tillmann, B., & Caclin, A. (2015). Altered retrieval of melodic information in congenital amusia: Insights from dynamic causal modeling of MEG data. *Frontiers in Human Neuroscience, 9*. https://doi.org/10.3389/fnhum.2015.00020

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain*, *125*(2), 238–251. https://doi.org/10.1093/brain/awf028

Boersma, P., & Weenink, D. (2001). *Praat: Doing Phonetics by Computer*. https://uvafon.hum.uva.nl/praat/

Chang, C., & Yao, Y. (2007, August 6). *Tone Production in whispered Mandarin*. Proceedings of the 16th international congress of phonetics sciences, Saarbrücken, Germany.

Cheung, Y. L., Zhang, C., & Zhang, Y. (2021). Emotion processing in congenital amusia: The deficits do not generalize to written emotion words. *Clinical Linguistics & Phonetics*, *35*(2), 101–116. https://doi.org/10.1080/02699206.2020.1719209

Cousineau, M., Oxenham, A. J., & Peretz, I. (2015). Congenital amusia: A cognitive disorder limited to resolved harmonics and with no peripheral basis. *Neuropsychologia*, *66*, 293–301. https://doi.org/10.1016/j.neuropsychologia.2014.11.031

Couvignou, M., Peretz, I., & Ramus, F. (2019). Comorbidity and cognitive overlap between developmental dyslexia and congenital amusia. Cognitive Neuropsychology, 1–17. https://doi.org/10.1080/02643294.2019.1578205

Couvignou, M., & Kolinsky, R. (2021). Comorbidity and cognitive overlap between developmental dyslexia and congenital amusia in children. Neuropsychologia, 107811. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2021.107811

Douglas, K. M., & Bilkey, D. K. (2007). Amusia is associated with deficits in spatial processing. *Nature Neuroscience*, *10*(7), 915–921. https://doi.org/10.1038/nn1925

Eklund, I., & Traunmüller, H. (1997). Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica*, *54*(1), 1–21. https://doi.org/10.1159/000262207

Foxton, J. M. (2004). Characterization of deficits in pitch perception underlying “tone deafness.” *Brain*, *127*(4), 801–810. https://doi.org/10.1093/brain/awh105

Gao, M. (2002). *Tones in whispered Chinese: Articulatory features and peceptual cues* [Master Thesis]. University of Victoria.

Gussenhoven, C., & Chen, A. (2000). *Universal and language-specific effects in the perception of question intonation*. 91–94.

Heeren, W. F. L. (2015). Vocalic correlates of pitch in whispered versus normal speech. *The Journal of the Acoustical Society of America*, *138*(6), 3800–3810. https://doi.org/10.1121/1.4937762

Heeren, W., & Heuven, V. J. V. (2009). *Perception and production of boundary tones in whispered Dutch*. 2411–2414.

Higashikawa, M., Nakai, K., Sakakura, A., & Takahashi, H. (1996). Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study. *Journal of Voice*, *10*(2), 155–158. https://doi.org/10.1016/S0892-1997(96)80042-7

Hirschberg, J., & Ward, G. (1992). The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics*, *20*(2), 241–251. https://doi.org/10.1016/S0095-4470(19)30625-4

Ho, A. T. (1977). Intonation Variation in a Mandarin Sentence for Three Expressions: Interrogative, Exclamatory and Declarative. *Phonetica*, *34*(6), 446–457. https://doi.org/10.1159/000259916

Huang, W.-T., Liu, C., Dong, Q., & Nan, Y. (2015a). Categorical perception of lexical tones in mandarin-speaking congenital amusics. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00829

Huang, W.-T., Nan, Y., Dong, Q., & Liu, C. (2015b). Just-noticeable difference of tone pitch contour change for Mandarin congenital amusics. *The Journal of the Acoustical Society of America*, *138*(1), EL99–EL104. https://doi.org/10.1121/1.4923268

Hyde, K. L., & Peretz, I. (2003). “Out-of-pitch” but still “in-time.” *Annals of the New York Academy of Sciences*, *999*(1), 173–176. https://doi.org/10.1196/annals.1284.023

Irwin, R. J., Hautus, M. J., & Stillman, J. A. (1992). Use of the receiver operating characteristic in the study of taste perception. *Journal of Sensory Studies*, *7*(4), 291–314. https://doi.org/10.1111/j.1745-459X.1992.tb00196.x

Jensen, M. K. (1958). Recognition of word tones in whispered speech. *WORD*, *14*(2–3), 187–196. https://doi.org/10.1080/00437956.1958.11659663

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., Chen, X., & Yang, Y. (2012). Amusia Results in Abnormal Brain Activity following Inappropriate Intonation during Speech Comprehension. *PLoS ONE*, *7*(7), e41411. https://doi.org/10.1371/journal.pone.0041411

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., & Yang, Y. (2010). Processing melodic contour and speech intonation in congenital amusics with Mandarin Chinese. *Neuropsychologia*, *48*(9), 2630–2639. https://doi.org/10.1016/j.neuropsychologia.2010.05.009

Jiao, L., Ma, Q., Wang, T., & Xu, Y. (2015). *Perceptual Cues of Whispered Tones: Are They Really Special?* 5.

Jiao, L., & Xu, Y. (2019). Whispered Mandarin has no production-enhanced cues for tone and intonation. *Lingua*, *218*, 24–37. https://doi.org/10.1016/j.lingua.2018.01.004

Jones, J. L., Lucker, J., Zalewski, C., Brewer, C., & Drayna, D. (2009). Phonological processing in adults with deficits in musical pitch recognition. *Journal of Communication Disorders*, *42*(3), 226–234. https://doi.org/10.1016/j.jcomdis.2009.01.001

Jones, J. L., Zalewski, C., Brewer, C., Lucker, J., & Drayna, D. (2009). Widespread Auditory Deficits in Tune Deafness. *Ear & Hearing*, *30*(1), 63–72. https://doi.org/10.1097/AUD.0b013e31818ff95e

Kallail, K. J., & Emanuel, F. W. (1984). Formant-Frequency Differences Between Isolated Whispered and Phonated Vowel Samples Produced by Adult Female Subjects. *Journal of Speech, Language, and Hearing Research*, *27*(2), 245–251. https://doi.org/10.1044/jshr.2702.251

Lehmann, A., Skoe, E., Moreau, P., Peretz, I., & Kraus, N. (2015). Impairments in musical abilities reflected in the auditory brainstem: Evidence from congenital amusia. *European Journal of Neuroscience*, *42*(1), 1644–1650. https://doi.org/10.1111/ejn.12931

Li, B., & Guo, Y. (2012, May 26). *Mandarin Tone Contrast in Whisper*. Third International Symposim on Tonal Aspects of Languages, Nanjing, China.

Li, B., & Rong, R. (2012). Tones in whispered Mandarin. *2012 8th International Symposium on Chinese Spoken Language Processing*, 422–425. https://doi.org/10.1109/ISCSLP.2012.6423539

Li, X., & Xu, B. (2005). Formant comparison between whispered and voiced vowels in Mandarin. *Acta Acustica United with Acustica*, *91*(6), 1079–1085.

Lima, C. F., Brancatisano, O., Fancourt, A., Müllensiefen, D., Scott, S. K., Warren, J. D., & Stewart, L. (2016). Impaired socio-emotional processing in a developmental music disorder. *Scientific Reports*, *6*(1), 34911. https://doi.org/10.1038/srep34911

Liu, F., Jiang, C., Thompson, W. F., Xu, Y., Yang, Y., & Stewart, L. (2012). The Mechanism of Speech Processing in Congenital Amusia: Evidence from Mandarin Speakers. *PLoS ONE*, *7*(2), e30374. https://doi.org/10.1371/journal.pone.0030374

Liu, F., Jiang, C., Wang, B., Xu, Y., & Patel, A. D. (2015). A music perception disorder (congenital amusia) influences speech comprehension. *Neuropsychologia*, *66*, 111–118. https://doi.org/10.1016/j.neuropsychologia.2014.11.001

Liu, F., Patel, A. D., Fourcin, A., & Stewart, L. (2010). Intonation processing in congenital amusia: Discrimination, identification and imitation. *Brain*, *133*(6), 1682–1693. https://doi.org/10.1093/brain/awq089

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin Lexical Tones when F0 Information is Neutralized. *Language and Speech*, *47*(2), 109–138. https://doi.org/10.1177/00238309040470020101

Ma, J. K.-Y., Ciocca, V., & Whitehill, T. L. (2006). Effect of intonation on Cantonese lexical tones. *The Journal of the Acoustical Society of America*, *120*(6), 3978–3987. https://doi.org/10.1121/1.2363927

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed). Lawrence Erlbaum Associates.

Matsuda, M., & Kasuya, H. (1999). *Acoustic nature of the whisper. 1*, 147–140.

Mignault Goulet, G., Moreau, P., Robitaille, N., & Peretz, I. (2012). Congenital amusia persists in the developing brain after daily music listening. *PLoS ONE*, *7*(5), e36860. https://doi.org/10.1371/journal.pone.0036860

Moreau, P., Jolicœur, P., & Peretz, I. (2013). Pitch discrimination without awareness in congenital amusia: Evidence from event-related potentials. *Brain and Cognition*, *81*(3), 337–344. https://doi.org/10.1016/j.bandc.2013.01.004

Nan, Y., Sun, Y., & Peretz, I. (2010). Congenital amusia in speakers of a tone language: Association with lexical tone agnosia. *Brain*, *133*(9), 2635–2642. https://doi.org/10.1093/brain/awq178

Nan, Y., Huang, W., Wang, W., Liu, C., & Dong, Q. (2016). Subgroup differences in the lexical tone mismatch negativity (MMN) among Mandarin speakers with congenital amusia. *Biological Psychology*, *113*, 59–67. https://doi.org/10.1016/j.biopsycho.2015.11.010

Ohala, J. J. (1983). Cross-Language Use of Pitch: An Ethological View. *Phonetica*, *40*(1), 1–18. https://doi.org/10.1159/000261678

Patel, A. D., Wong, M., Foxton, J., Lochy, A., & Peretz, I. (2008). Speech intonation perception deficits in musical tone deafness (congenital amusia). *Music Perception*, *25*(4), 357–368. https://doi.org/10.1525/mp.2008.25.4.357

Peretz, I., Ayotte, J., Zatorre, R. J., Mehler, J., Ahad, P., Penhune, V. B., & Jutras, B. (2002). Congenital amusia: A disorder of fine-grained pitch discrimination. *Neuron*, *33*(2), 185–191. https://doi.org/10.1016/S0896-6273(01)00580-3

Peretz, I., Brattico, E., & Tervaniemi, M. (2005). Abnormal electrical brain responses to pitch in congenital amusia. *Annals of Neurology*, *58*(3), 478–482. https://doi.org/10.1002/ana.20606

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders: The montreal battery of evaluation of amusia. *Annals of the New York Academy of Sciences*, *999*(1), 58–75. https://doi.org/10.1196/annals.1284.006

Peretz, I., Cummings, S., & Dubé, M.-P. (2007). The genetics of congenital amusia (tone deafness): A family-aggregation study. *The American Journal of Human Genetics*, *81*(3), 582–588. https://doi.org/10.1086/521337

Peretz, I., & Vuvan, D. T. (2017). Prevalence of congenital amusia. *European Journal of Human Genetics*, *25*(5), 625–630. https://doi.org/10.1038/ejhg.2017.15

Phillips-Silver, J., Toiviainen, P., Gosselin, N., Piché, O., Nozaradan, S., Palmer, C., & Peretz, I. (2011b). Born to dance but beat deaf: A new form of congenital amusia. *Neuropsychologia*, *49*(5), 961–969. https://doi.org/10.1016/j.neuropsychologia.2011.02.002

Shao, J., Lau, R. Y. M., Tang, P. O. C., & Zhang, C. (2019). The Effects of Acoustic Variation on the Perception of Lexical Tone in Cantonese-Speaking Congenital Amusics. *Journal of Speech, Language, and Hearing Research*, *62*(1), 190–205. https://doi.org/10.1044/2018_JSLHR-H-17-0483

Shao, J., & Zhang, C. (2020). Dichotic Perception of Lexical Tones in Cantonese-Speaking Congenital Amusics. *Frontiers in Psychology*, *11*, 1411. https://doi.org/10.3389/fpsyg.2020.01411

Shao, J., Zhang, C., Peng, G., Yang, Y., & Wang, W. S.-Y. (2016). *Effect of Noise on Lexical Tone Perception in Cantonese-Speaking Amusics*. 272–276. https://doi.org/10.21437/Interspeech.2016-891

Sharifzadeh, H. R., McLoughlin, I. V., & Russell, M. J. (2012). A Comprehensive Vowel Space for Whispered Speech. *Journal of Voice*, *26*(2), e49–e56. https://doi.org/10.1016/j.jvoice.2010.12.002

Sun, Y., Lu, X., Ho, H. T., & Thompson, W. F. (2017). Pitch discrimination associated with phonological awareness: Evidence from congenital amusia. *Scientific Reports*, *7*(1), 44285. https://doi.org/10.1038/srep44285

Thompson, W. F. (2007). Exploring variants of amusia: Tone deafness, rhythm impairment, and intonation insensitivity. *Proceedings of the Inaugural International Conference on Music Communication Science*, 159–163.

https://researchers.mq.edu.au/en/publications/exploring-variants-of-amusia-tone-deafness-rhythm-impairment-and-

Thompson, W. F., Marin, M. M., & Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proceedings of the National Academy of Sciences*, *109*(46), 19027–19032. https://doi.org/10.1073/pnas.1210344109

Vuvan, D. T., Nunes-Silva, M., & Peretz, I. (2015). Meta-analytic evidence for the non-modularity of pitch processing in congenital amusia. *Cortex*, *69*, 186–200. https://doi.org/10.1016/j.cortex.2015.05.002

Vuvan, D. T., Paquette, S., Mignault Goulet, G., Royal, I., Felezeu, M., & Peretz, I. (2018). The Montreal Protocol for Identification of Amusia. Behavior Research Methods, 50(2), 662–672. https://doi.org/10.3758/s13428-017-0892-8

Wang, X., & Peng, G. (2014). Phonological processing in Mandarin speakers with congenital amusia. *The Journal of the Acoustical Society of America*, *136*(6), 3360–3370. https://doi.org/10.1121/1.4900559

Whiteford, K. L., & Oxenham, A. J. (2017b). Auditory deficits in amusia extend beyond poor pitch perception. *Neuropsychologia*, *99*, 213–224. https://doi.org/10.1016/j.neuropsychologia.2017.03.018

Xu, Y. (2013). *ProsodyPro—A tool for large-scale systematic prosody analysis*. 7–10.

Yan, J. (2016). *Experimental study on foreign accent of Chinese learners*. Beijing language and culture university.

Yang, L., Li, Y., & Xu, B. (2005). The establishment of a Chinese whisper database and perceptual experiment (in Chinese). *Journal of NanJing University (Natural Sciences)*, *41*(3), 311–317.

Zhang, C., Shao, J., & Huang, X. (2017). Deficits of congenital amusia beyond pitch: Evidence from impaired categorical perception of vowels in Cantonese-speaking congenital amusics. *PLOS ONE*, *12*(8), e0183151. https://doi.org/10.1371/journal.pone.0183151

Zhang, C., & Shao, J. (2018). Normal pre-attentive and impaired attentive processing of lexical tones in Cantonese-speaking congenital amusics. *Scientific Reports*, *8*(1), 8420. https://doi.org/10.1038/s41598-018-26368-7

Zhang, G., Shao, J., Huang, X., Wang, L., & Zhang, C. (2018). Unequal Impairment of Native and Non-native Tone Perception in Cantonese-speaking Congenital Amusics. *9th International Conference on Speech Prosody 2018*, 562–566. https://doi.org/10.21437/SpeechProsody.2018-114

Zhang, W., & Dong, W. (2004). *Advanced tutorial of SPSS statistical analysis (in Chinese)*. Higer Education Press.

Table 1. Demographic characteristics of participants. The results of independent-samples t-tests comparing the amusics and controls in age and the scores of MBEA test are also reported here. n.s. = not significant. The *p*-value was corrected for multiple comparisons on the MBEA tests (.05/8 = .006).

	Amusics	Controls	*t*-value	*p*-value	Cohen's *d*
Male/Female (total)	9/10 (19)	9/10 (19)	/	/	/
Mean Age (range)	24.37 (20-30)	24.42 (20-31)	-0.07	n.s.	/
MBEA					
Scale (SD)	55.16% (14.24)	85.90% (11.16)	-7.41	$p < 0.001$	2.40
Contour (SD)	58.56% (15.50)	94.23% (4.76)	-9.59	$p < 0.001$	3.11
Interval (SD)	58.57% (7.85)	93.02% (3.78)	-17.24	$p < 0.001$	5.59
Rhythm (SD)	61.13% (13.75)	93.54% (6.88)	-9.20	$p < 0.001$	2.98
Meter (SD)	50.53% (10.44)	84.39% (12.07)	-9.25	$p < 0.001$	3.00
Memory (SD)	71.06% (16.12)	96.49% (3.92)	-6.68	$p < 0.001$	2.17
Pitch composite score (SD)	51.69 (8.16)	81.94 (4.62)	-14.01	$p < 0.001$	4.56
Global (SD)	58.84% (7.32)	91.26% (4.65)	-16.30	$p < 0.001$	5.29

Table 2. Results of bivariate correlations between the dependent MBEA scores and identification performance, and between the three MBEA scores. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Phonated tone	Whispered tone	Phonated intonation	Whispered intonation	Melodic organization	Temporal organization
Melodic organization	0.50**	0.39*	0.57***	0.44**		
Temporal organization	0.61***	0.47**	0.57***	0.45**	0.91***	
Melodic memory	0.47**	0.54***	0.65***	0.50**	0.82***	0.78***

Table 3. Results of stepwise linear regression models with the MBEA scores as predictors on tone and intonation identification. Note: The values represent standardized regression coefficients for the predictors retained in the model. Empty cells indicate that the predictor was not retained in the model. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Phonation mode	MBEA predictors			Adjusted R^2 of the model
	Melodic organization	Temporal organization	Melodic memory	
Tone				
Phonated mode		0.61***		0.35***
Whispered mode			0.54***	0.27***
Intonation				
Phonated mode			0.65***	0.41***
Whispered mode			0.50**	0.23**

Table 4. Results of multinomial logistic regression models with the acoustic cues as predictors on phonated tone identification. Note: The tone following versus was used as the baseline for the contrast (e.g., T1 was the baseline in the T2 vs. T1 pair). Controls were used as the baseline for comparison with amusics. The values represent regression coefficients (*B* (odds ratio)) for the predictors. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For significant interactions between group and acoustic cues, cases where a certain acoustic cue led to worse performance in amusics than in controls were marked in bold. In contrast, cases where a certain acoustic cue led to better performance in amusics than in controls were marked with #.

Tones	Main effects of predictors						
	F0 mean	F0 SD	Duration	Intensity	F1	F2	Group
T2 vs. T1	-0.03*** (0.97)	0.35*** (1.41)	-0.01*** (0.99)	-0.28 *** (0.75)	0.002*** (1.002)	-0.001** (0.99)	-15.57** (1.74e-7)
T3 vs. T1	-0.11*** (0.90)	0.43*** (1.54)	0.02*** (1.02)	-0.57*** (0.57)	0.005*** (1.005)	0.0003 (1.0003)	-15.84* (1.32e-7)
T4 vs. T1	-0.001 (0.99)	0.39*** (1.48)	-0.04*** (0.96)	0.23*** (1.26)	-0.005*** (0.99)	-0.002*** (0.99)	-12.10* (5.78e-6)
T3 vs. T2	-0.08*** (0.93)	0.09** (1.09)	0.04*** (1.04)	-0.28*** (0.75)	0.003*** (1.003)	0.001*** (1.001)	-0.28 (0.76)
T4 vs. T2	0.03*** (1.03)	0.04*** (1.04)	-0.03*** (0.97)	0.52*** (1.68)	-0.007*** (0.99)	-0.002*** (0.99)	3.50 (33.22)
T4 vs. T3	0.11*** (1.12)	-0.05 (0.96)	-0.07*** (0.94)	0.80*** (2.22)	-0.01*** (0.99)	-0.002*** (0.99)	3.78 (43.89)
Interactions between group and acoustic cues							

T2 vs. T1	0.01** (1.01)	-0.14*** (0.87)	0.008*** (1.008)	0.16** (1.17)	-0.001 (0.99)	0.0004*# (1.0004)	/
T3 vs. T1	0.06*** (1.06)	-0.19*** (0.83)	-0.01*** (0.99)	0.26** (1.30)	-0.002**# (0.99)	-0.0002 (0.99)	/
T4 vs. T1	-0.005 (0.99)	-0.12*** (0.88)	0.02*** (1.02)	0.06 (1.07)	0.001 (1.001)	0.001***# (1.001)	/
T3 vs. T2	0.05*** (1.05)	-0.05 (0.95)	-0.02*** (0.98)	0.10 (1.11)	-0.001*# (0.99)	-0.001**# (0.99)	/
T4 vs. T2	-0.02** (0.99)	0.02 (1.02)	0.009*** (1.01)	-0.10 (0.91)	0.002**# (1.002)	0.0005**# (1.0005)	/
T4 vs. T3	-0.06*** (0.94)	0.07*# (1.07)	0.03*** (1.03)	-0.20* (0.82)	0.003***# (1.003)	0.001***# (1.001)	/

Table 5. Results of binominal regression models with the acoustic cues as predictors on phonated intonation identification. Note: The statement was used as the baseline for contrast with the question. Controls were used as the baseline for comparison with amusics. The values represent regression coefficients (*B* (odds ratio)) for the predictors. 'F0 dir' means F0 direction. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For significant interactions between group and acoustic cues, cases where a certain acoustic cue led to worse performance in amusics than in controls were marked in bold. In contrast, cases where a certain acoustic cue led to better performance in amusics than in controls were marked with #.

Main effects of predictors							
Sentence					Final syllable		Group
F0 mean	F0 SD	F0 dir	Duration	Intensity	F0 SD	Duration	
0.02 *** (1.02)	0.11 *** (1.11)	2.17 *** (8.75)	-0.0002 (0.99)	1.19 *** (3.29)	0.02 * (1.02)	0.02 *** (1.02)	43.99 *** (1.27e+19)
Interactions between group and acoustic cues							
-0.007 ******* **(0.99)**	**-0.05** ******* **(0.95)**	**-1.07** ******* **(0.34)**	-0.0002 (0.99)	**-0.55** ******* **(0.58)**	-0.02 (0.98)	**-0.01** ******* **(0.99)**	/

Table 6. Results of multinomial logistic regression models with the acoustic cues as predictors on whispered tone identification. Note: The tone following versus was used as the baseline for the contrast (e.g., T1 was the baseline in the T2 vs. T1 pair). Controls were used as the baseline for comparison with amusics. The values represent regression coefficients (*B* (odds ratio)) for the predictors. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For significant interactions between group and acoustic cues, cases where a certain acoustic cue led to worse performance in amusics than in controls were marked in bold. In contrast, cases where a certain acoustic cue led to better performance in amusics than in controls were marked with #.

Tones	Main effects of predictors				
	Duration	Intensity	F1	F2	Group
T2 vs. T1	-0.002*** (0.99)	0.004 (1.004)	0.001*** (1.001)	0.001*** (1.001)	0.78 (2.18)
T3 vs. T1	0.008*** (1.008)	0.04** (1.04)	-0.002*** (0.99)	-0.0001 (0.99)	1.17 (3.22)
T4 vs. T1	-0.01*** (0.99)	-0.08*** (0.92)	0.002*** (1.002)	0.001*** (1.001)	-0.99 (0.42)
T3 vs. T2	0.01*** (1.01)	0.04* (1.04)	-0.003*** (0.99)	-0.001*** (0.99)	0.39 (1.48)
T4 vs. T2	-0.01*** (0.99)	-0.09*** (0.92)	0.001** (1.001)	0.001*** (1.001)	-1.77 (0.17)
T4 vs. T3	-0.02*** (0.98)	-0.12*** (0.89)	0.004*** (1.004)	0.001*** (1.001)	-2.16 (0.12)
Interaction between the group and acoustic cues					
T2 vs. T1	**0.002* (1.002)**	-0.009 (0.99)	-0.0003 (0.99)	-0.0005**# (0.99)	/

T3 vs. T1	-0.001 (0.99)	-0.009 (0.99)	0.001 (1.001)	-0.0003*# (0.99)	/
T4 vs. T1	**0.003***** **(1.003)**	0.005 (1.008)	-0.0003 (0.99)	**-0.0005**** **(0.99)**	/
T3 vs. T2	**-0.003***** **(0.99)**	0.0002 (1.0002)	**0.001*** **(1.001)**	0.0002 (1.0002)	/
T4 vs. T2	0.002 (1.002)	0.01 (1.01)	0.00006 (1.00006)	0.000008 (1.000008)	/
T4 vs. T3	**0.004***** **(1.004)**	0.01 (1.01)	-0.001 (0.99)	-0.0002 (0.99)	/

Table 7. Results of binominal logistic regression models with the acoustic cues as predictors on whispered intonation identification. Note: The statement was used as the baseline for contrast with the question. Controls were used as the baseline for comparison with amusics. The values represent regression coefficients (*B* (odds ratio)) for the predictors. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For significant interactions between group and acoustic cues, cases where a certain acoustic cue led to worse performance in amusics than in controls were marked in bold. In contrast, cases where a certain acoustic cue led to better performance in amusics than in controls were marked with #.

Main effects of predictors				
Sentence		Final syllable		Group
Duration	Intensity	Duration	Intensity	
-0.001*** (0.99)	0.03 (1.03)	0.01*** (1.01)	0.02 (1.02)	-0.62 (0.54)
Interaction between the group and acoustic cues				
0.0003* (1.0003)	0.03 (1.03)	-0.002 (0.99)	-0.02 (0.98)	/

Figure 1. The F0 contours of the four Mandarin tones produced in the phonated mode by two native speakers.

Figure 2. Real-time F0 contours of a statement-question pair produced by the male speaker. This sentence is ‘高兵喝鸡汤./?’ /kau55 piəŋ55 xə55 tɕi55 tʰɑŋ55/ ‘Gao Bing drinks chicken soup./?’, in which all the syllables carried T1.

Figure 3. The tone identification accuracy in the phonated and whispered mode in the amusics and controls.

Figure 4. The d’ of intonation identification in the phonated and whispered mode in the amusics and controls.

The supplementary information file includes a total of ten tables and five figures. The tables showed the word and sentence stimuli, acoustic cues of the four tones and two intonation patterns in the phonated and whispered mode, confusion matrices and other supportive statistical analysis results. The figures displayed the relationships between the participants' tone and intonation identification performance (in both phonation modes) and their MBEA scores and acoustic cues.

Figure 1

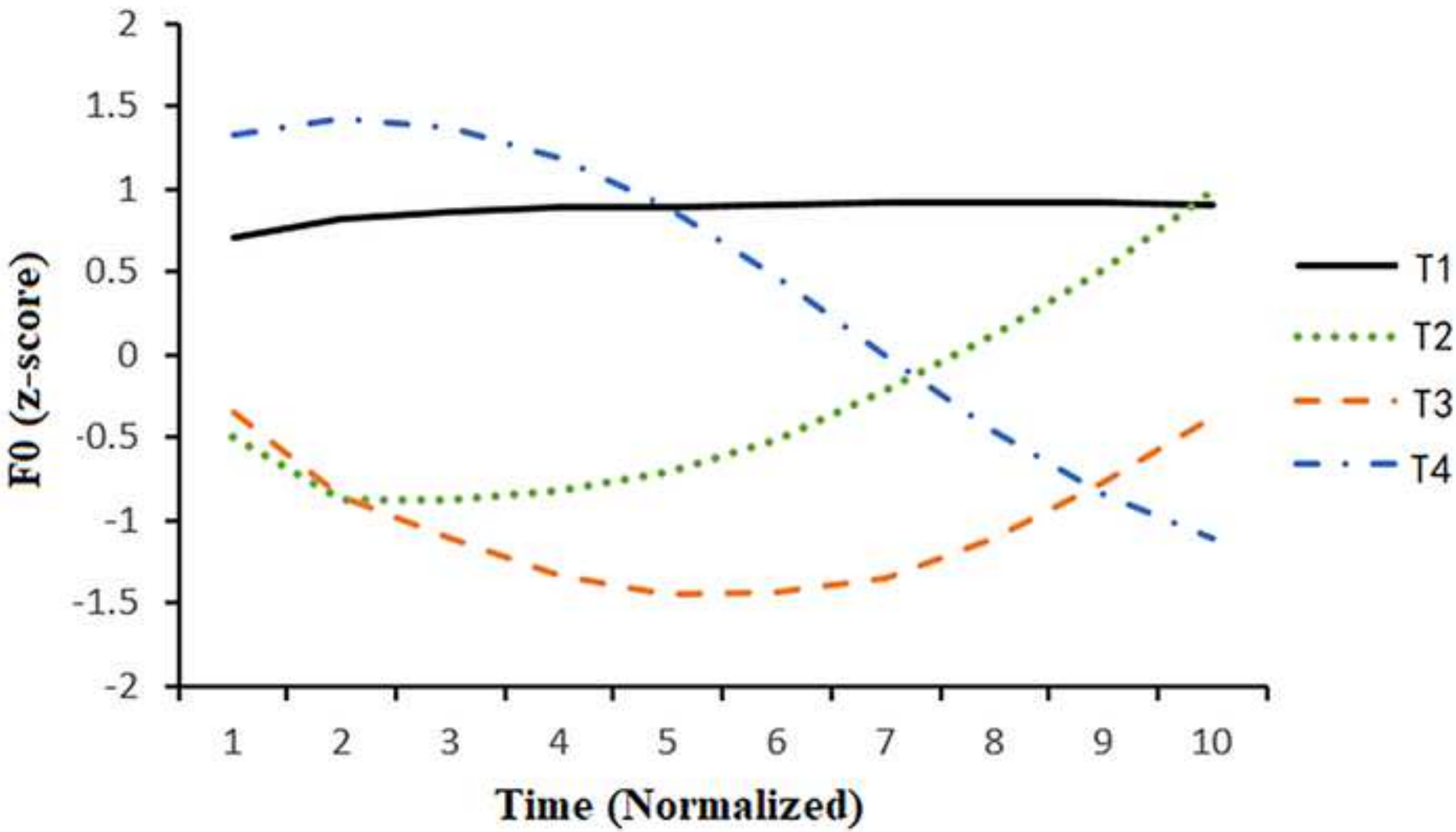
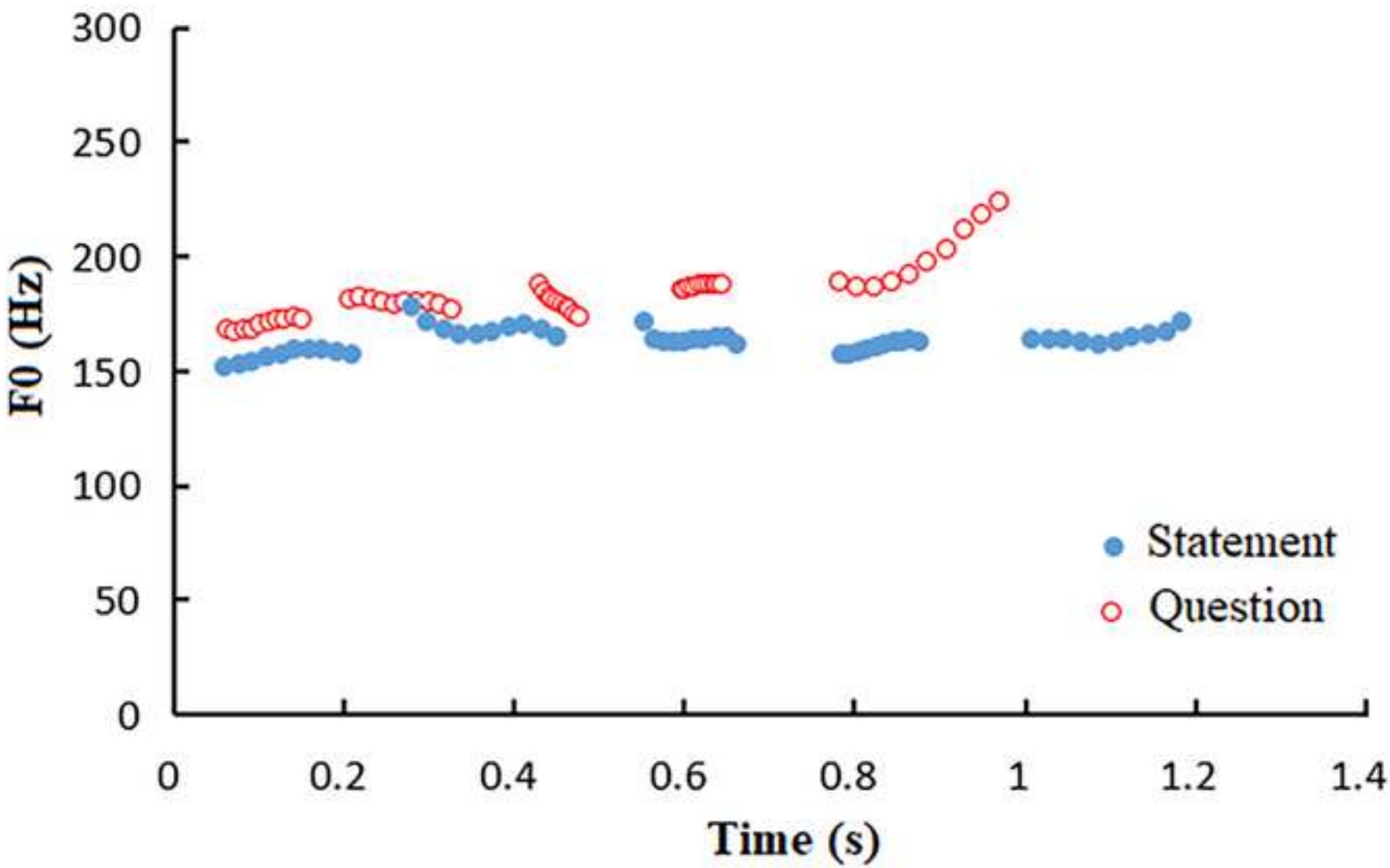

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
F0 (z-score)
1
2
3
4
5
6
7
8
9
10
Time (Normalized)
T1
T2
T3
T4

Figure 2


300
250
200
150
100
50
0
F0 (Hz)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Time (s)
Statement
Question

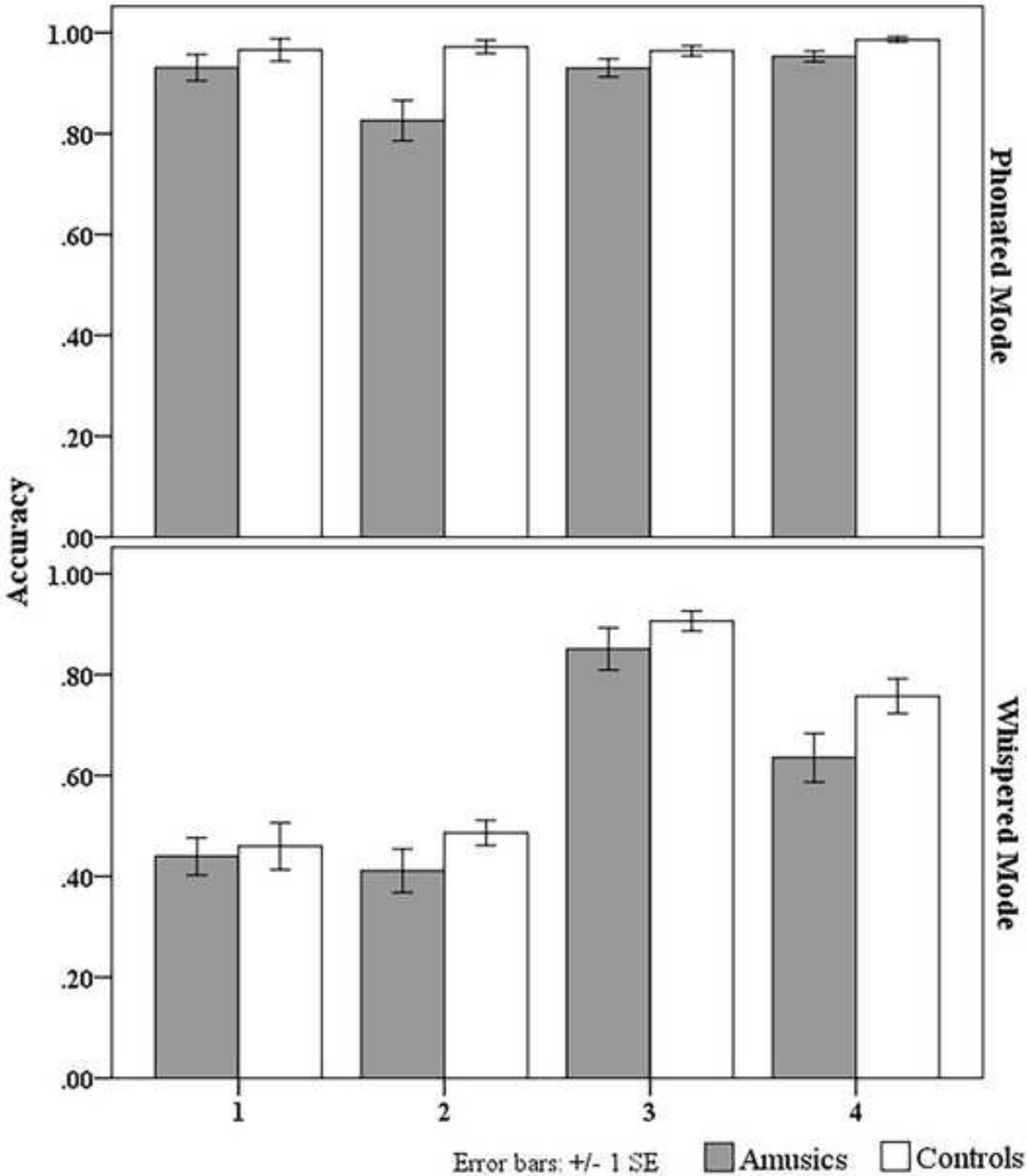
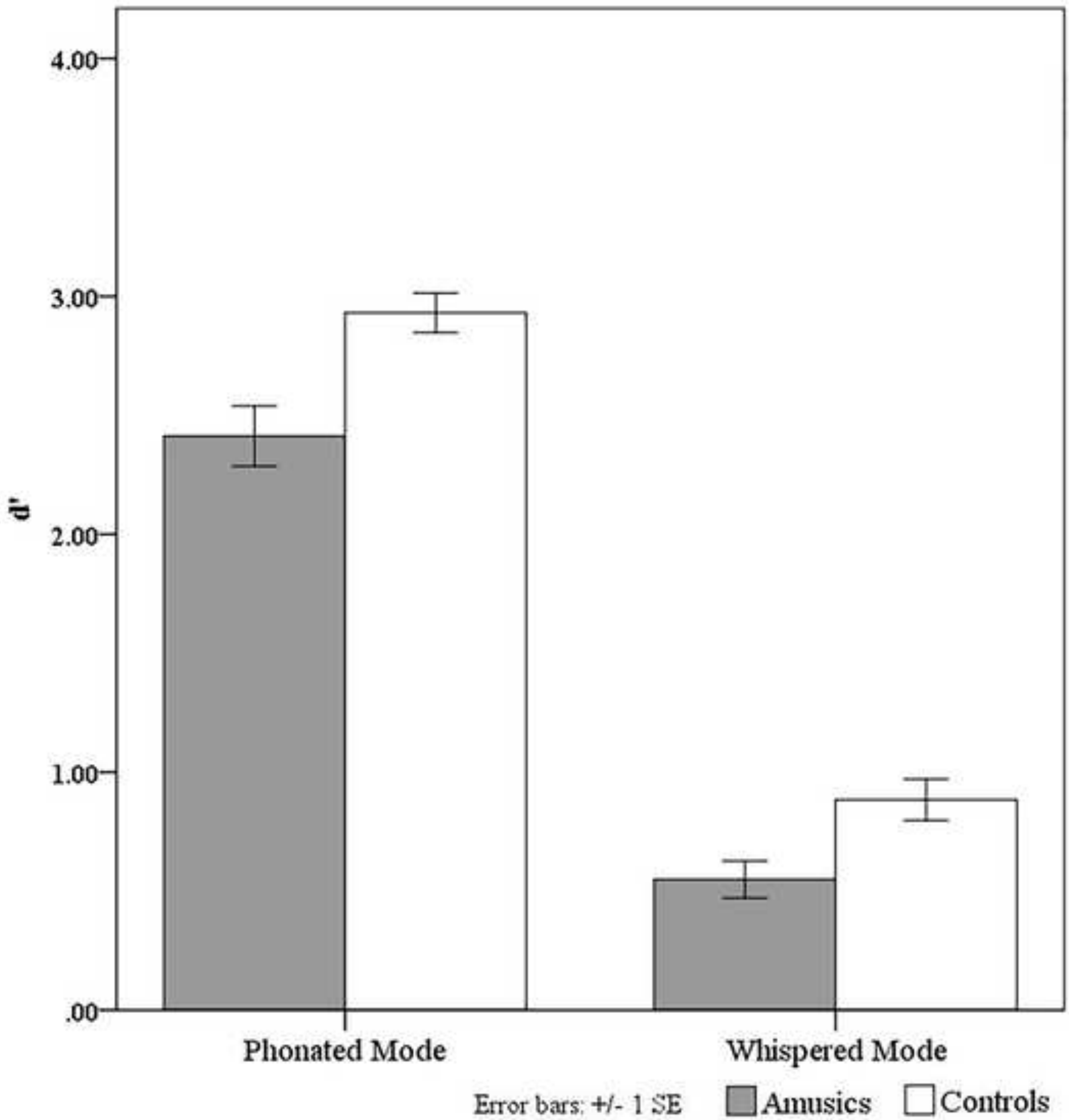

1.00
.80
.60
.40
.20
.00
Accuracy
Phonated Mode
Whispered Mode
1
2
3
4
Error bars: +/- 1 SE
Amusics
Controls

Figure 4


4.00
3.00
2.00
1.00
.00
d'
Phonated Mode
Whispered Mode
Error bars: +/- 1 SE
Amusics
Controls

Supplementary Information

Table 1. A list of syllables and sentences for production and perception experiments.

Syllables									
Tone	/ta/	/ti/	/tu/	/pa/	/pi/	/tʃu/	/a/	/i/	/u/
1	搭 (build)	低 (low)	督 (supervise)	八 (eight)	逼 (force)	猪 (pig)	啊 (oh)	一 (one)	屋 (house)
2	答 (answer)	敌 (enemy)	毒 (poison)	拔 (pull)	鼻 (nose)	竹 (bamboo)	啊 (eh)	遗 (pity)	无 (nothing)
3	打 (fight)	底 (bottom)	赌 (bet)	把 (handle)	笔 (pen)	主 (lord)	啊 (what)	已 (already)	五 (five)
4	大 (big)	地 (land)	肚 (belly)	爸 (father)	币 (coin)	祝 (wish)	啊 (ah)	易 (easy)	误 (error)

Sentences			
4 syllables	张薇开车。/? /tʂɑŋ55 uei55 kʰai55 tʂʰə55/ (Zhang Wei drives the car./?)	5 syllables	高兵喝鸡汤。/? /kau55 piəŋ55 xə55 tɕi55 tʰɑŋ55/ (Gao Bing drinks chicken soup./?)
	王梅划船。/? /uɑŋ35 mei35 xua35 tʂʰuan35/ (Wang Mei boats./?)		罗婷学轮滑。/? /luo35 tʰiəŋ35 ɕyɛ35 luən35 xua35/ (Luo Ting learns skating./?)
	李敏点火。/? /li214 miən214 diɛn214 huo214/ (Li Min makes a fire./?)		李伟买雨伞。/? /li214 uei214 mai214 y214 san214/ (Li Wei buys an umbrella./?)
	叶亮睡觉。/? /iɛ51 liɑŋ51 ʂuei51 tɕiɑu51/ (Ye Liang sleeps./?)		赵志看电视。/? /tʂau51 tʂʅ51 kʰan51 tiɛn51 ʂʅ51/ (Zhao Zhi watches TV./?)
	李刚讲课。/? /li214 kɑŋ55 tɕiɑŋ214 kʰə51/ (Li Gang gives a lesson./?)		李刚交水费。/? /li214 kɑŋ55 tɕiau55 ʂuei214 fei51/ (Li Gang pays water fee./?)

6 syllables	张薇担心肖英。 /tʂɑŋ55 uei55 tan55 ɕiən55 ɕiɑu iəŋ55/ (Zhang Wei worries about Xiao Ying./?)	7 syllables	高兵今天喝鸡汤。/? /kau55 piəŋ55 tɕiən55 tʰiɛn55 xə55 tɕi55 tʰɑŋ55/ (Gao Bing drinks chicken soup today./?)
	王梅怀疑刘宁。/? /uɑŋ35 mei35 xuai35 i35 liou35 niəŋ35/ (Wang Mei suspects Liu Ning.)		罗婷明年学轮滑。/? /luo35 tʰiəŋ35 miəŋ35 niɛn35 ɕyɛ35 luən35 xua35/ (Luo Ting will learn skating next year./?)
	李敏反感刘雨。/? /li214 miən214 fan214 kan214 liou35 y214/ (Li Min is disgusted with Liu Yu)		李伟五点买雨伞。/? /li214 uei214 u214 tiɛn214 mai214 y214 san214/ (Li Wei buys an umbrella at 5 o'clock./?)
	叶亮害怕赵丽。/? /iɛ51 liɑŋ51 xai51 pʰa51 tʂɑu51 li51/ (Ye Liang is afraid of Zhang Li./?)		赵志半夜看电视。/? /tʂɑu51 tʂɿ51 pan51 iɛ51 kʰan51 tiɛn51 ʂɿ51/ (Zhao Zhi watches TV at midnight./?)
	李刚讨厌吕梦。/? /li214 kɑŋ55 tʰau214 iɛn51 ly214 məŋ51/ (Li Gang hates Lü Meng./?)		李刚九号交水费。/? /li214 kɑŋ55 tɕiou214 xau51 tɕiau55 ʂuei214 fei51/ (Li Gang pays water fee on ninth./?)
10 syllables	张薇担心肖英开车发晕。/? /tʂɑŋ55 uei55 tan55 ɕiən55 ɕiɑu55 iəŋ55 kʰai55 tʂʰə55 fa55 yən55/ (Zhang Wei worries about Xiao Ying having a carsickness./?		
	王梅怀疑刘宁划船着迷 。/? /uɑŋ35 mei35 xuaɪ35 i35 liou35 nɪəŋ35 xua35 tʂʰuan35 tʂɑu35 mi35/ (Wang Mei suspects Liu Ning of indulging in boating./?)		
	李敏反感刘雨点火取暖。/? /li214 miən214 fan214 kan214 liou35 y214 tiɛn214 xuo214 tɕʰy214 nuan214/ (Li Min is disgusted with Liu Yu making a fire for warmth./?)		
	叶亮害怕赵丽睡觉做梦。/? /iɛ51 liɑŋ51 xai51 pʰa51 tʂɑu51 li51 ʂuei51 tɕiau51 tsuo51 məŋ51/ (Ye Liang is afraid of Zhang Li dreaming when sleeping./?)		

	李刚讨厌吕梦讲课紧张。/? /li214 kɑŋ55 tʰau214 iɛn51 ly214 məŋ51 tɕiɑŋ214 kʰə51 tɕiən214 tʂɑŋ55/ (Li Gang hates Lü Meng to be nervous when teaching./?)

Table 2. Acoustic characteristics of the four Mandarin tones produced in the phonated mode and results of one-way ANOVAs conducted to compare the four tones on each acoustic cue (the *p*-value was corrected for multiple comparisons: .05/6 = .008).

Tone	$F0_{mean}$ (Hz)	$F0_{SD}$	Duration (ms)	Intensity (dB)	F1 (Hz)	F2 (Hz)
T1 (SD)	194.11 (8.75)	3.53 (2.39)	516.95 (64.93)	72.06 (3.46)	586.05 (360.61)	1655.66 (842.74)
T2 (SD)	154.83 (11.06)	23.03 (4.01)	470.17 (54.68)	71.44 (2.70)	577.33 (369.32)	1664.7 (828.89)
T3 (SD)	128.04 (6.02)	15.59 (3.21)	662.97 (60.79)	68.01 (1.85)	574.02 (375.87)	1621.73 (875.03)
T4 (SD)	173.68 (8.2)	34.58 (5.93)	355.29 (68.84)	72.81 (2.98)	613 (354.73)	1677.88 (834.75)
p value	$p < .001$	$p < .001$	$p < .001$	$p = .005$	n.s.	n.s.
$\eta^2_{partial}$	0.90	0.89	0.78	0.33	0.02	0.01

Table 3. Acoustic characteristics of the four Mandarin tones produced in the whispered mode and results of one-way ANOVAs conducted to compare the four tones on each acoustic cue (the *p*-value was corrected for multiple comparisons: .05/4 = .0125).

Tone	Duration (ms)	Intensity (dB)	F1 (Hz)	F2 (Hz)
T1 (SD)	515.94 (42.05)	54.02 (5.35)	874.46 (259.63)	1867.19 (621.35)
T2 (SD)	472.93 (44.75)	54.08 (5.23)	833.65 (290.29)	1826.28 (632.03)
T3 (SD)	629.10 (63.24)	51.74 (5.48)	809.94 (290.36)	1789.99(666.22)
T4 (SD)	365. 81 (58.34)	55.81 (4.17)	838.73 (276.04)	1872.67 (602.65)
p value	$p < .001$	n.s.	n.s.	n.s.
$\eta^2_{partial}$	0.78	0.08	0.08	0.03

Table 4. Acoustic characteristics of statements and questions in the phonated mode and results of t-tests conducted to compare statement and question on each acoustic cue (the *p*-value was corrected for multiple comparisons: .05/9 = .005).

Sentence					
Intonation	F0 mean (Hz)	F0 SD	F0 direction	Duration (ms)	Intensity (dB)
Statement (SD)	163.28 (28.16)	22.34 (10.81)	-0.42 (0.40)	1745.91 (541.99)	69.46 (1.41)
Question (SD)	213.36 (26.77)	29.27 (14.96)	0.22 (0.60)	1463.45 (446.62)	72.41 (1.12)
p value	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$
Cohen's *d*	1.82	0.53	1.26	0.57	2.32
Final syllable					
Intonation	F0 mean (Hz)	F0 SD	Duration (ms)	Intensity (dB)	
Statement (SD)	152.32 (36.02)	19.33 (12.31)	292.69 (47.06)	67.05 (2.16)	
Question (SD)	222.48 (46.96)	24.12 (11.48)	313.13 (38)	72.51 (2.19)	
p value	$p < .001$	n.s.	n.s. ($p = .012$)	$p < .001$	
Cohen's *d*	1.68	0.4	0.48	2.51	

Table 5. Acoustic characteristics of statements and questions in the whispered mode and results of t-tests conducted to compare statement and question on each acoustic cue (the *p*-value was corrected for multiple comparisons: .05/4 = .0125).

Intonation	Sentence		Final Syllable	
	Duration (ms)	Mean Intensity (dB)	Duration (ms)	Mean Intensity (dB)
Statement (SD)	1645.67 (482.22)	53.46 (2.49)	305.11 (61.13)	52.29 (5.17)
Question (SD)	1453.95 (470.21)	53.25 (2.09)	330.95 (42.82)	53.54 (4.00)
p value	$p < .001$	n.s.	$p = .001$	n.s.
Cohen's *d*	0.40	0.10	0.49	0.27

Table 6. Results of correlations among the nine acoustic cues of the phonated sentences. Note: 'S' means sentence (e.g., 'S F0 mean' means the mean F0 of the whole sentences). 'Fs' means final syllable (e.g., 'FS F0 mean' means the mean F0 of the final syllable).

Pearson Correlation/ items	S F0 mean	S F0 SD	S F0 dir	S duration	S intensity	Fs F0 mean	Fs F0 SD	Fs duration	Fs intensity
S F0 mean	1	0.52	0.21	0.06	-0.2	0.84	0.25	0.18	-0.002
S F0 SD		1	-0.03	0.008	-0.23	0.32	0.61	0.18	-0.03
S F0 dir			1	-0.09	0.44	0.49	-0.24	0.08	0.44
S duration				1	-0.29	0.04	-0.04	-0.05	0.34
S intensity					1	-0.002	-0.19	-0.19	0.83
Fs F0 mean						1	0.02	0.007	0.17
Fs F0 SD							1	0.15	0.14
Fs duration								1	0.12
Fs intensity									1

Table 7. Confusion matrix of tone identification in the phonated mode. For each tonal category of the stimuli, the target tone response was in bold, and the tone response receiving the highest confusion rate was italicized.

Group	Heard / Original	T1 (%)	T2 (%)	T3 (%)	T4 (%)
Controls	T1	**96.59**	*2.92*	0.19	0.29
	T2	*2.14*	**97.17**	0.29	0.39
	T3	0.39	*2.34*	**96.39**	0.68
	T4	0.29	*0.78*	0.29	**98.64**
Amusics	T1	**93.08**	*4.87*	1.36	0.58
	T2	6.63	**82.55**	*9.16*	1.56
	T3	1.75	*4.39*	**92.98**	0.88
	T4	1.56	*1.85*	1.27	**95.32**

Table 8. Confusion matrix of tone identification in the whispered mode. For each tonal category of the stimuli, the target tone response was in bold, and the tone response receiving the highest confusion rate was italicized.

Group	Heard / Original	T1 (%)	T2 (%)	T3 (%)	T4 (%)
Controls	T1	**46.00**	12.38	16.96	*24.66*
	T2	12.86	**48.64**	*29.53*	8.97
	T3	1.27	*7.50*	**90.64**	0.58
	T4	*14.23*	4.58	5.46	**75.73**
Amusics	T1	**43.96**	*21.05*	18.42	16.57
	T2	17.64	**41.13**	*32.94*	8.28
	T3	4.58	*8.28*	**85.09**	2.05
	T4	*18.23*	9.36	8.87	**63.55**

Table 9. Confusion matrix of intonation identification.

Phonated mode	Heard / Original	Question (%)	Statement (%)	Whispered mode	Question (%)	Statement (%)
Controls	Question	**96.89**	3.11	Controls	**51.21**	48.79
	Statement	1.95	**98.05**		12.42	**87.58**
Amusics	Question	**89.63**	10.37	Amusics	**41.26**	58.74
	Statement	3.47	**96.53**		16.58	**83.42**

Table 10. Results of simultaneous linear regression models with the MBEA scores as predictors on tone and intonation identification. Note: The values represent standardized regression coefficients for the predictors retained in the model. The values in parentheses present VIF. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Phonation mode	MBEA predictors			R^2 of the model
	Melodic organization	Temporal organization	Melodic memory	
Tone				
Phonated mode	-0.38 (7.22)	0.89* (6.2)	0.09 (3.06)	0.62**
Whispered mode	-0.61	0.55	0.60*	0.59**
Intonation				
Phonated mode	-0.07	0.21	0.55*	0.66***
Whispered mode	-0.08	0.21	0.40	0.51*

Figure 1. The relationship with the MBEA scores and tone and intonation identification in both phonation modes.

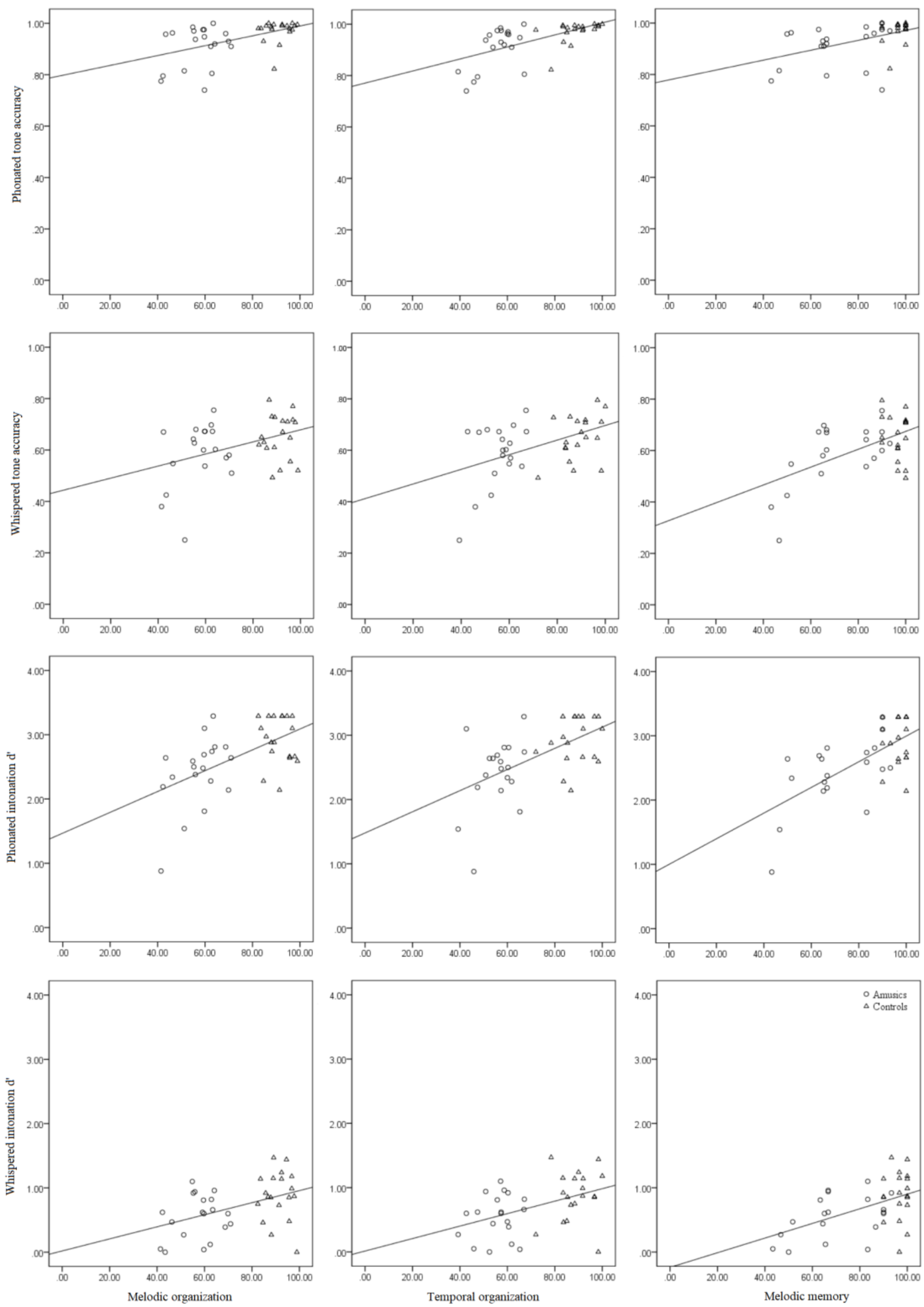
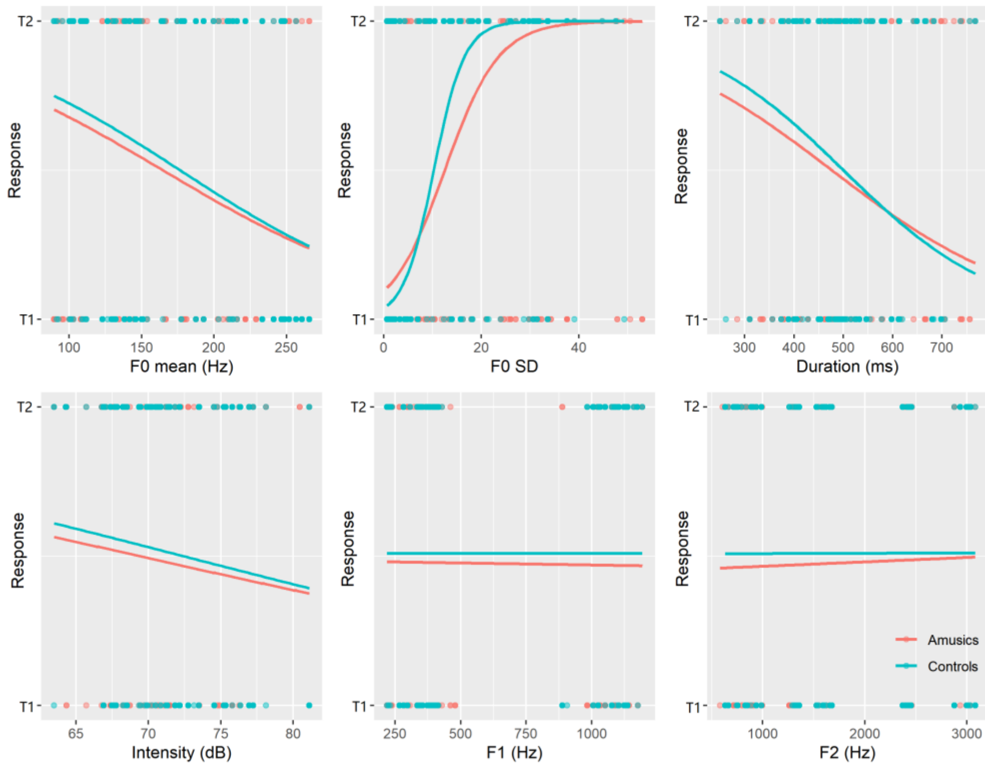
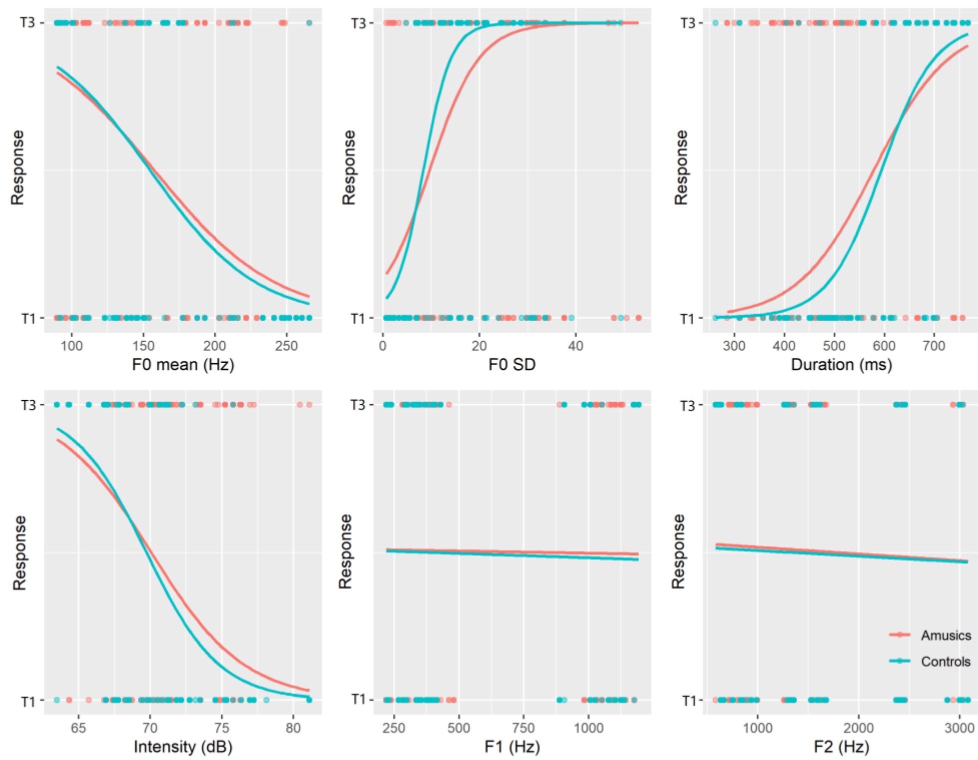


Figure 2. The relationship between the acoustic cues and phonated tone identification.

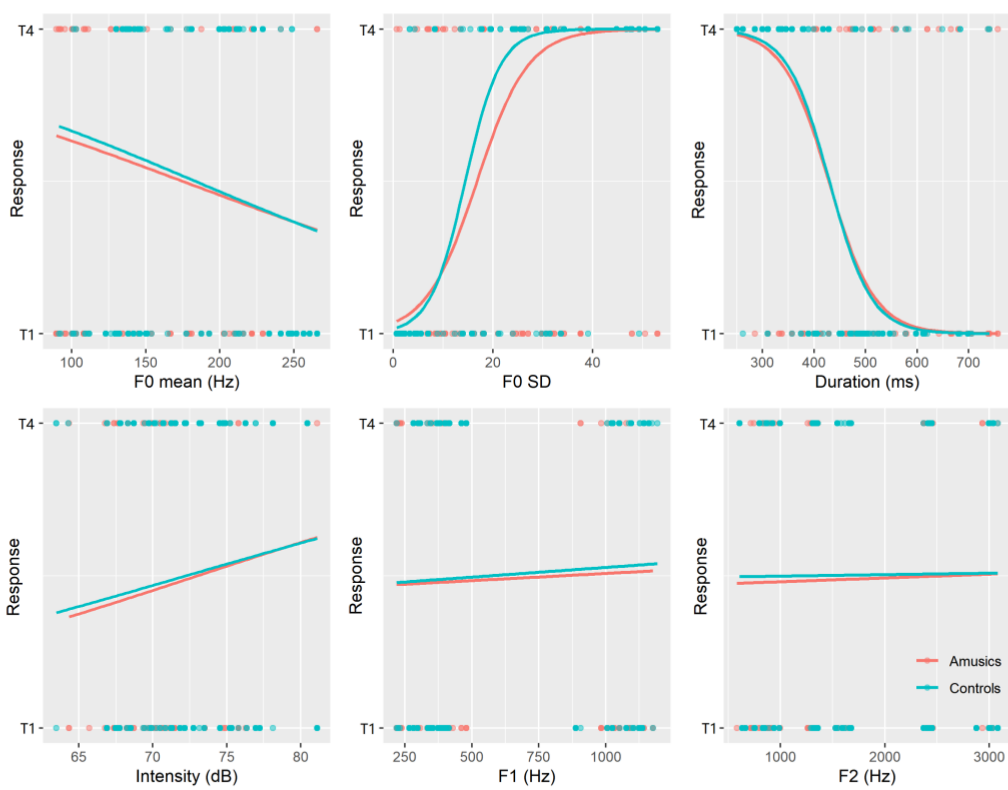
(a) The relationship between the acoustic cues and phonated T2 and T1 identification.



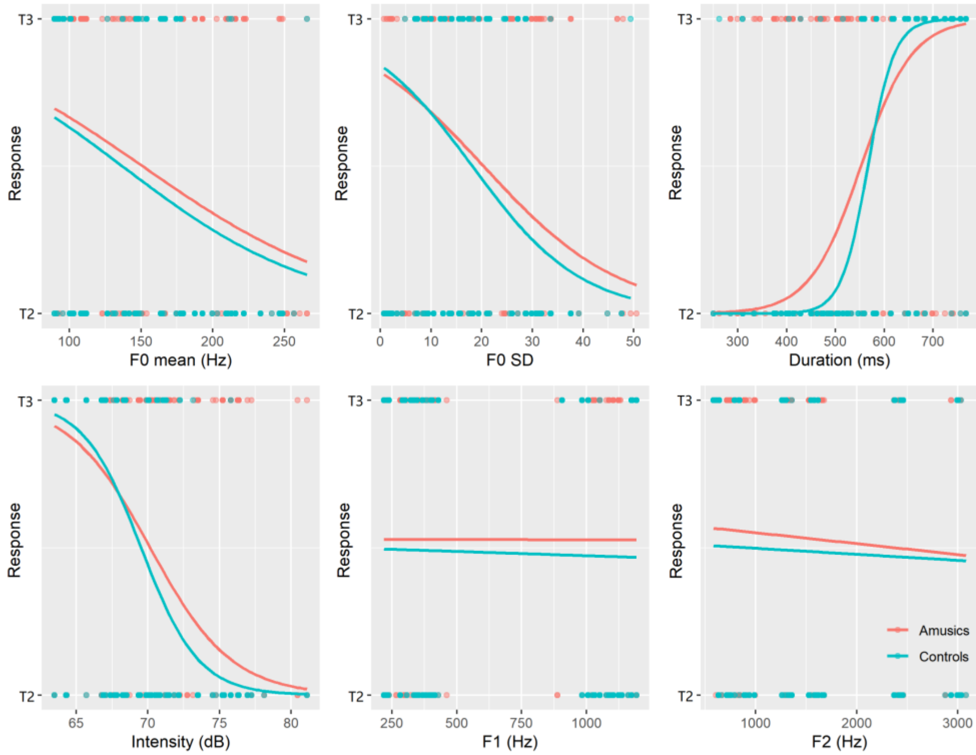
(b) The relationship between the acoustic cues and phonated T3 and T1 identification.



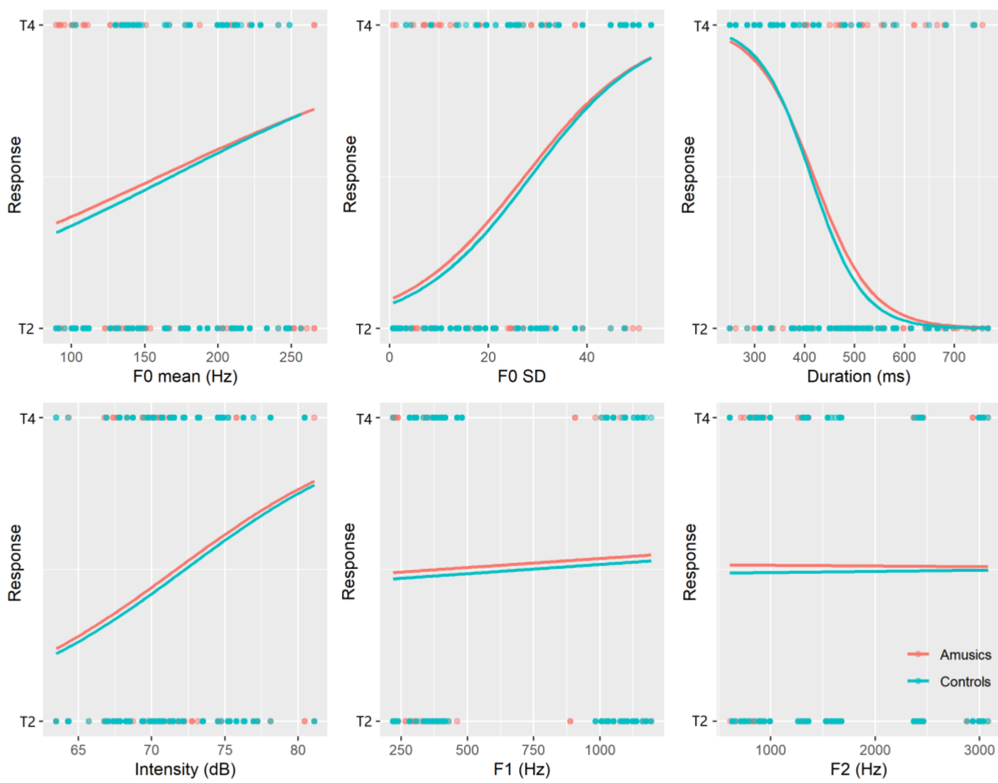
(c) The relationship between the acoustic cues and phonated T4 and T1 identification.



(d) The relationship between the acoustic cues and phonated T3 and T2 identification.



(e) The relationship between the acoustic cues and phonated T4 and T2 identification.



(f) The relationship between the acoustic cues and phonated T4 and T3 identification.

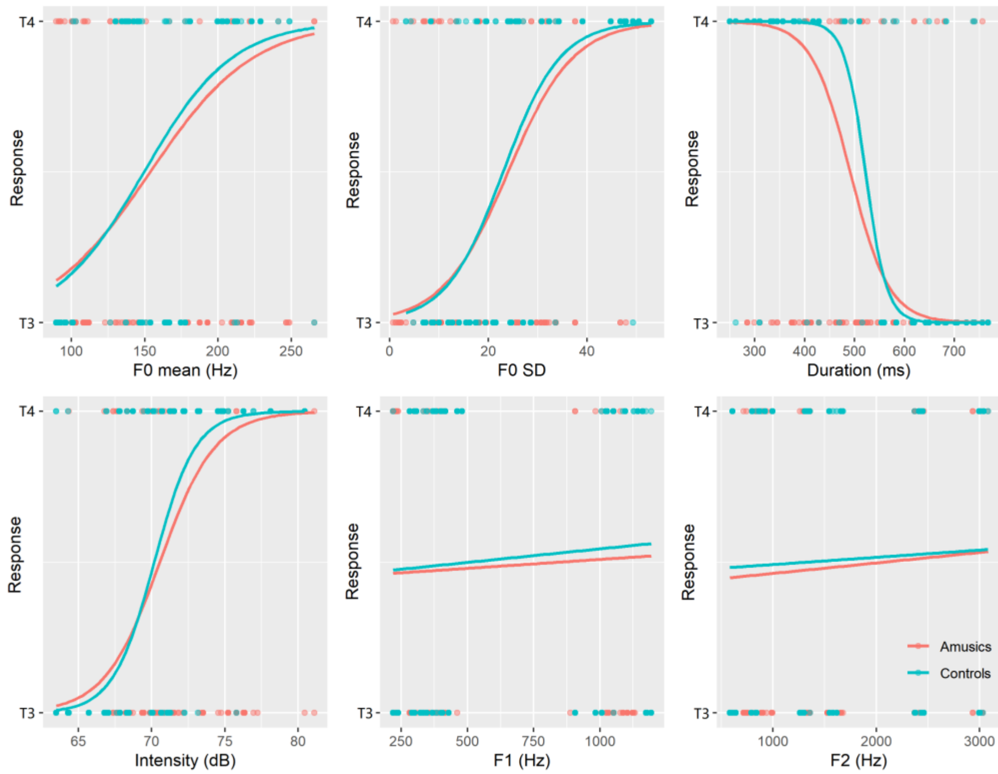


Figure 3. The relationship between the acoustic cues and phonated intonation identification.

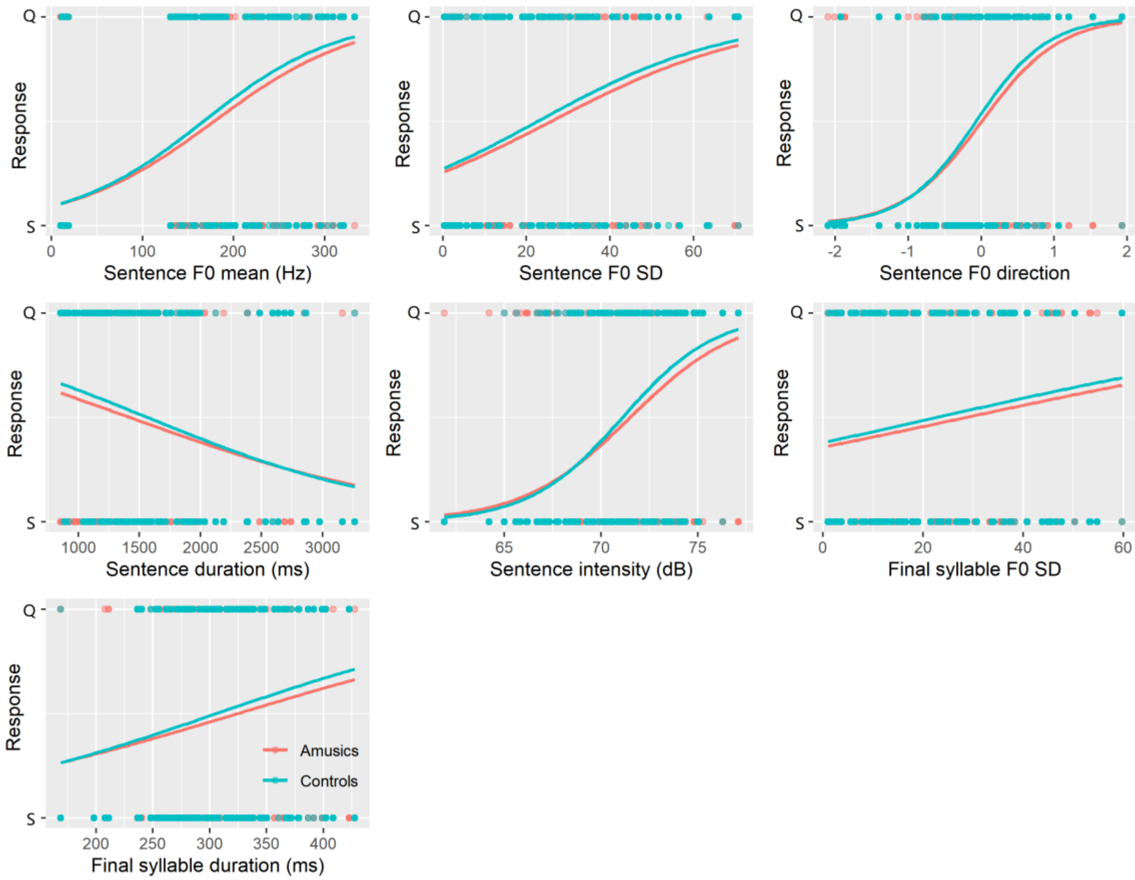
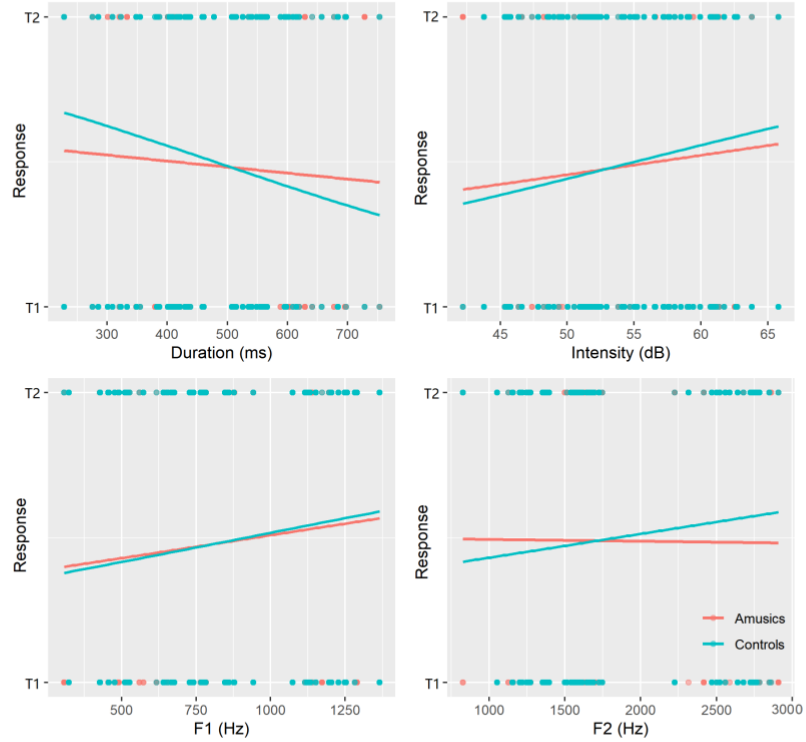
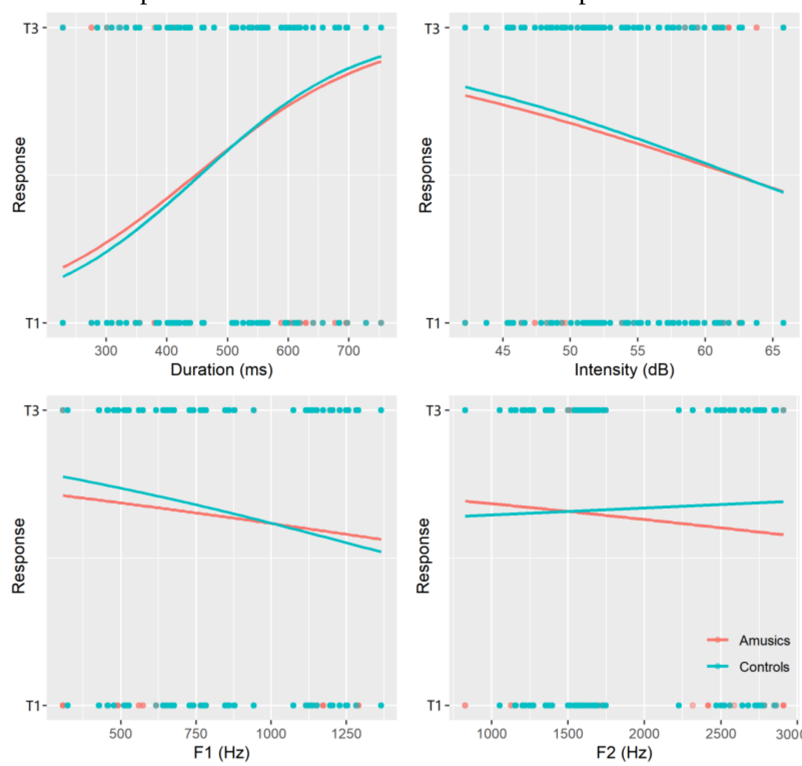


Figure 4. The relationship between the acoustic cues and whispered tone identification.

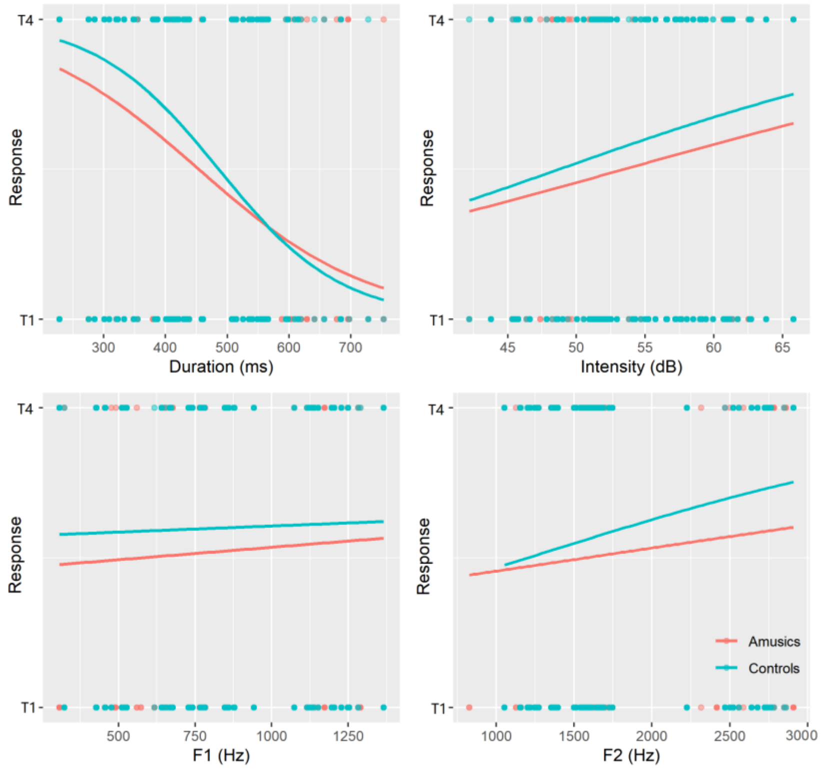
(a) The relationship between the acoustic cues and whispered T2 and T1 identification.



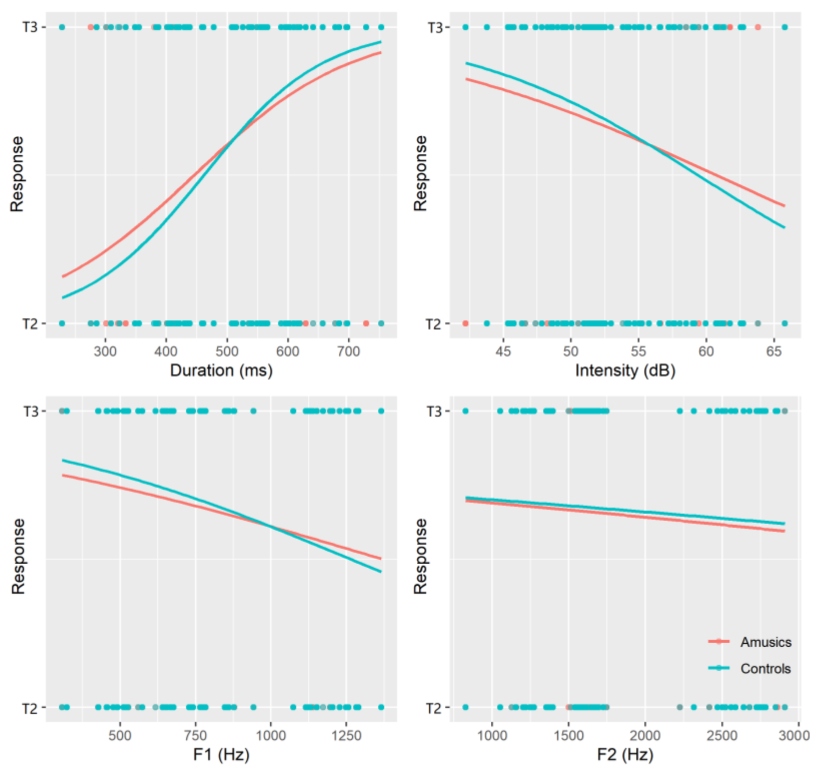
(b) The relationship between the acoustic cues and whispered T3 and T1 identification.



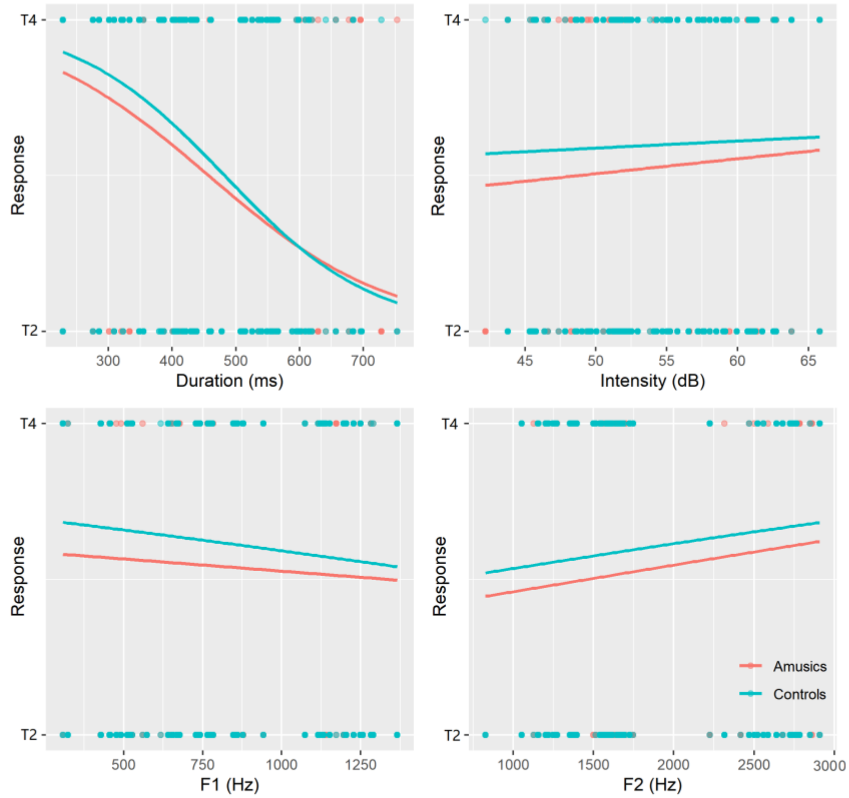
(c) The relationship between the acoustic cues and whispered T4 and T1 identification.



(d) The relationship between the acoustic cues and whispered T3 and T2 identification.



(e) The relationship between the acoustic cues and whispered T4 and T2 identification.



(f) The relationship between the acoustic cues and whispered T4 and T3 identification.

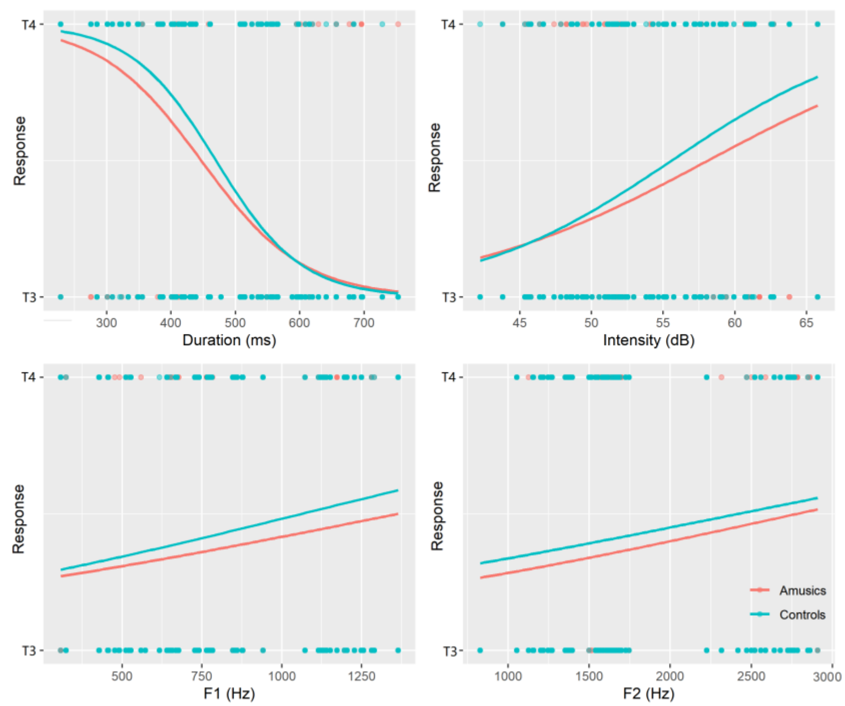


Figure 5. The relationship between the acoustic cues and whispered intonation identification.