Talker processing in Mandarin-speaking congenital amusics

Jing Shao[a,b], Lan Wang[b], and Caicai Zhang[c,d,*]

[a]School of Humanities, Shanghai Jiao Tong University, China

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

[c]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China

[d]Research Centre for Language, Cognition, and Neuroscience, The Hong Kong Polytechnic University, Hong Kong SAR, 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: Talker processing deficit in congenital 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).

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

Abstract

Purpose: The ability to recognize individuals from their vocalizations is an important trait of human beings. In the current study we aimed to examine how congenital amusia, an inborn pitch-processing disorder, affects discrimination and identification of talkers' voices.

Method: Twenty Mandarin-speaking amusics and 20 controls were tested on talker discrimination and identification in four types of contexts that varied in the degree of language familiarity: Mandarin real words, Mandarin pseudowords, Arabic words and reversed Mandarin speech.

Results: The language familiarity effect was more evident in the talker identification task than the discrimination task for both participant groups, and talker identification accuracy decreased as native phonological representations were removed from the stimuli. Importantly, amusics demonstrated degraded performance in both native speech conditions that contained phonological/linguistic information to facilitate talker identification and non-native conditions where talker voice processing primarily relied on phonetics cues including pitch. Moreover, the performance in talker processing can be predicted by the participants' musical ability and phonological memory capacity.

Conclusion: The results provided a first set of behavioral evidence that individuals with amusia are impaired in the ability of human voice identification. Meanwhile, it is found that amusia is not only a pitch disorder, but is likely to affect the phonological processing of speech, in terms of making using of phonological information in native speech to analyze a talker's identity. The above findings expanded the understanding of the nature and scope of congenital amusia.

Keywords: congenital amusia; talker voice processing; language familiarity; phonological memory; Mandarin Chinese

Introduction

Amusia is a lifelong neurogenetic disorder of fine-grained pitch processing in music (Peretz et al., 2002), with an estimated prevalence rate of approximately 1.5-4% (Peretz & Vuvan, 2017). Individuals with amusia have deficits in the processing of pitch (Foxton, Dean, Gee, Peretz, & Griffiths, 2004; Hyde, Peretz, Hyde, & Peretz, 2004; Peretz et al., 2002) and also show impaired short-term memory for pitch (Tillmann, Lévêque, Fornoni, Albouy, & Caclin, 2016; Tillmann, Schulze, & Foxton, 2009). Moreover, a number of studies have revealed that the deficit in amusia is likely to be domain-general, impeding speech pitch processing, including intonation and emotional state perception (Jiang et al., 2012; Jiang, Hamm, Lim, Kirk, & Yang, 2010; Liu, Patel, Fourcin, & Stewart, 2010; Lu, Ho, Liu, Wu, & Thompson, 2015; Thompson, Marin, & Stewart, 2012). Evidence has also shown that tonal language speakers with amusia were impoverished in the identification and discrimination of native tones compared to typical listeners (Liu et al., 2016; Nan et al., 2010; Shao, Lau, Tang, & Zhang, 2019; Shao, Zhang, Peng, Yang, & Wang, 2016). Furthermore, some studies have suggested that the categorical perception of native lexical tones was impaired, with amusics demonstrating null or reduced benefit from between-category tone discriminations, which is a sign of impaired categorical perception (Huang, Liu, Dong, & Nan, 2015; Jiang, Hamm, Lim, Kirk, & Yang, 2012; Zhang, Shao, & Huang, 2017), suggesting that high-level phonological processing of lexical tones might be impaired in tonal language speakers with amusia.

In addition to the aforementioned findings, a recent study has revealed that Chinese-speaking amusics may be impaired in lexical tone normalization in terms of using phonological cues in the speech context for adapting to talker variation (Shao & Zhang, 2018; Zhang, Shao, & Chen, 2018). Zhang et al., (2018) found that typical listeners performed better in tone

normalization in the conditions where phonological cues were available for estimating a talker's tone space (meaningless and meaningful speech contexts) than the conditions where the phonological cues were diminished (reversed speech and nonspeech contexts). In contrast to controls, amusics particularly showed degraded performance in the meaningful and meaningless contexts where phonological cues were available, indicating that amusics have a deficit in making use of phonological cues in surrounding context in the process of tone normalization.

In addition, several studies have reported impoverished phonological awareness in amusics (Jones, Lucker, Zalewski, Brewer, & Drayna, 2009; Sun, Lu, Ho, & Thompson, 2017). Jones et al. (2009) reported lower phonological and phonemic awareness abilities in amusics compared to controls on all measures, including auditory word discrimination, syllable segmentation, and the Comprehensive Test of Phonological Processing (CTOPP). In another study, Sun et al. (2017) investigated whether pitch abilities in amusics were associated with phonological abilities. Four subtests from the CTOPP-2 were adopted. At the group level, there was no significant difference in the phonological performance between amusics and controls. However, eight amusics who showed severe pitch impairment showed significantly lower scores in the subtest of Elision, suggesting impaired phonological awareness.

Taken together, converging evidence has underlined the impaired phonological processing of lexical tones (as indexed by reduced categorical perception and reduced effects of native speech contexts with phonological cues to normalize tones) and reduced phonological awareness in amusia. All these findings are related to speech perception in the linguistic dimension. However, linguistic information and a talker's voice information overlap in the speech signal and interact with each other during perception (Green, Tomiak, & Kuhl, 1997; Mullennix & Pisoni, 1990). On the one hand, extensive studies have suggested that detailed talker-specific

information influences the perception of phonological representations (Creel & Bregman, 2011) and speech perception is less accurate and takes longer time when the talker variability increases (Green et al., 1997; Lee, 2009; Lee, Tao, & Bond, 2010; Mullennix & Pisoni, 1990; Nusbaum & Morin, 1992; Shao et al., 2019; Shao & Zhang, 2019; Strange, Verbrugge, Shankweiler, & Edman, 1976; Wong & Diehl, 2003). On the other hand, talker voice identification is also influenced by the linguistic content in the speech signal (Goggin, Thompson, Strube, & Simental, 1991; C. P. Thompson, 1987). Typical listeners are more accurate at recognizing talkers' voices in their native language than in an unfamiliar language, a native language advantage that has been reported by a variety of studies (Goggin et al., 1991; Goldstein, Knight, Bailis, & Conover, 1981; Hollien, Majewski, & Doherty, 1982; Perrachione & Wong, 2007; C. P. Thompson, 1987; Winters, Levi, & Pisoni, 2008). For instance, Goggin et al., (1991) found that English and Spanish monolinguals were more accurate at recognizing talkers in their native language. Furthermore, Goggin and colleagues investigated the effects of linguistic content on voice identification. In addition to the regular/original English passage, three conditions differing in the amount of semantic, syntactic and lexical cues were generated: mixed words (the words were rearranged to be anomalous and the paragraphs were semantically nonsense, but preserving the syntactic structures), mixed syllables (the words were nonsense by rearranging the syllables from the original passages, in which only one-syllable words were retained) and reversed text (the original passage was time-reversed, in which normal phonological cues were destroyed) conditions. The results showed that as the listening materials progressively deviated from the native language, the identification accuracy of talkers' voices systematically decreased. In the reversed-text condition, the performance was even worse than in the foreign language context. Taken together, these results also supported the facilitating effects of familiar language, as the

talker identification accuracy declined with the syntactic, semantic and phonological properties of the listening materials removed. In addition, short-term training (six sessions) of recognizing talkers' voices cannot overcome the language effect (Perrachione & Wong, 2007).

Another line of research showed that talker identification was also influenced by listeners' ability in phonological skills (Perrachione, Del Tufo, & Gabrieli, 2011; Perrachione, del Tufo, Ghosh, & Gabrieli, 2011). For instance, Perrachione, Del Tufo, et al. (2011) examined the performance of listeners with dyslexia, who are known as having a core phonological deficit, in recognizing different talkers' voices in Mandarin and English. English listeners with dyslexia performed equally poorly as the controls in identifying the voices of Mandarin talkers. However, they were impaired in recognizing talkers' voices in English compared to controls. These findings suggested that when phonological representations are accessible in the native language, listeners with dyslexia may be unable to utilize the mapping between the phonetic variation in talker voices and the underlying phonological representation to identify the talkers (Perrachione, Del Tufo, et al., 2011). In contrast, in the Mandarin condition, neither group had access to the phonological representation in the foreign language, and the group difference was absent.

It was proposed that the difference in performance between familiar and unfamiliar languages can be attributed to two different types of talker information that are present in the speech signal: language-independent and language-specific information (Wester, 2012; Winters et al., 2008). Language-independent talker information is available across different languages and includes properties such as age, gender, and vocal tract length. Language-specific talker information is tied to linguistic content, such as the information of how a particular talker articulates native speech sounds. In the non-native language context, where there is no familiar linguistic information tied to the talker' voice, talker processing becomes less accurate.

Perrachione, Del Tufo, et al., (2011) further explained this effect as causally related to the ability to utilize the correspondence between the phonetic variations in talkers' voices and the abstract phonological representations. If phonological representations are absent, as in an unfamiliar foreign language, or if the ability to access to phonological representations is impoverished, as in the case of dyslexia, talker identification becomes deficient.

While amusia is widely known primarily as a pitch-processing deficit, few studies have examined how amusia affects talker voice processing. The ability to recognize conspecifics from their vocalizations is an important adaptive trait among animals and human (Perrachione, Del Tufo, et al., 2011). Previous studies on amusia have mostly focused on linguistic pitch processing, but in the task of talker identification, attention is not directed to the linguistic content but to a talker's voice. As a result, there is a research gap as to whether and how the deficit in amusia will hinder the explicit processing of the talker dimension in speech signals.

To fill the gap, we examined the performance of Mandarin-speaking amusics and musically intact controls on talker discrimination and identification in four types of contexts with the available linguistic cues gradually decreased: Mandarin real words, Mandarin pseudo-words, Arabic words and reversed Mandarin words. In light of the previous findings (Goggin et al., 1991; Goldstein et al., 1981; Hollien et al., 1982; Perrachione, Del Tufo, et al., 2011; Perrachione & Wong, 2007; C. P. Thompson, 1987; Winters et al., 2008), we expect talker processing to mainly depend on *language-independent information* and general auditory processing in conditions where native linguistic content was destroyed (as in reversed words) or absent (as in Arabic words), and on *language-specific information* and linguistic/phonological processing in native speech conditions (Mandarin real words and Mandarin pseudo-words), at least in typical controls. Accordingly, a native language advantage is expected, with more

accurate talker voice perception obtained in native speech conditions than non-native/reversed speech conditions. As for the performance of amusics, we predict that they will perform significantly worse than controls in non-native/reversed speech conditions, due to their known deficiency in auditory pitch processing. It is an open question as to whether amusics would process native speech conditions in a phonological or phonetic/auditory mode (Strange, 2011). The *phonological mode* of perception is typically used by (unimpaired) adult listeners, who detect phonologically relevant information in native speech highly automatically and efficiently, with over-learned and automatic routines. In contrast, the *phonetic mode*, which is slower and less accurate, is typically used by non-native listeners. If the former, we expect the amusics to exhibit a language familiarity effect in native speech conditions, similar to controls; crucially, amusics would be less capable of utilizing phonological cues to perform the talker identification task due to their degraded phonological processing ability, thus exhibiting worse performance than controls in native speech conditions. On the other hand, if amusics process native speech in a phonetic/auditory mode, we expect the amusics to show no language familiarity effect, together with a global impairment in talker voice processing in native as well as non-native conditions. Examining talker voice perception in listeners with amusia and listeners with intact musical ability in conditions differing in linguistic characteristics will broaden our understanding of the scope of deficits of amusia, and help specify their deficits in different levels of talker voice processing.

In Experiment 1, we designed a talker discrimination task to probe to what extent amusics can discern differences between talkers' voices in the four conditions. We adopted the discrimination task because it does not require the listeners to establish the talker category and the task demand is relatively low. As amusia is primarily a pitch-processing deficit, we aimed to

investigate whether amusics would show inferior performance in the easier discrimination task, which mostly involves the phonetic comparison between two talker's voices. In Experiment 2, we adopted a training task followed by a talker identification task, to determine whether the two groups of listeners were able to identify the talker's voice after training, and to explore the possible interaction between the linguistic conditions and talker processing. In this task, it is required to establish the abstract talker categories through training, and higher-order cognitive processes are involved. We included both the discrimination and identification tasks, as they tap into different levels of processing and are expected to reveal the potential talker-processing deficit in amusics in a more comprehensive manner. Finally, as previous studies have revealed that the performance in talker processing or talker's dialect categorization is typically correlated with their phonological memory capacity (Long, Fox, & Jacewicz, 2016; Perrachione, Del Tufo, et al., 2011), in Experiment 3, we further examined the phonological memory of the same subjects who have participated in Experiment 1 and 2, using a nonword repetition task. The correlation between the subjects' phonological memory capacity and their talker voice processing performance was examined. The three experiments were carried out on the same day for each participant, following the same sequential order (Experiment 1-2-3), with breaks provided between experiments to avoid fatigue. We administrated the discrimination task before the identification task, to avoid possible learning effects in the training phase of the identification task, which could potentially affect the performance in the discrimination task. The three experiments lasted about 2 hours in total (excluding breaks).

Experiment 1. Talker discrimination

Method

Participants

Twenty Mandarin-speaking amusics and 20 musically intact controls participated in this experiment. Amusics and controls were matched one by one in age, gender, and years of education. All participants were native speakers of Mandarin and university students in Shenzhen at the time of the experiment. They were all right-handed, with no reported hearing impairment, history of neurological illness or formal musical training (instrument or vocal). None of the participants reported any knowledge of Arabic, had ever lived in Arabic-speaking countries or had any Arabic-speaking friends or family members prior to participation in the study. Amusics and controls were identified using the Montreal Battery of Evaluation of Amusia (MBEA) (Peretz, Champod, & Hyde, 2003). The MBEA consists of six subtests: three of them are pitch-based tests (scale, contour, and interval), two of them are duration-based tests (rhythm and meter), and the last one is a melody memory test. All amusic participants scored below 71% (Nan et al., 2010) in the global score, which is the mean of all six subtests, whereas all control participants scored higher than 80%. Independent-samples t-tests confirmed that amusics' global scores were significantly lower than those of controls ($t(38) = -18.243$, $p < 0.001$). Amusics also performed significantly worse than controls in each subtest (ps < 0.001). Summarized demographic characteristics of the participants are shown in Table 1. The detailed geographic backgrounds of the participants are provided in supplemental materials. 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 participants in compliance with the experiment protocols.

Stimuli

The stimuli included four types of conditions - Mandarin real words, Mandarin pseudo-words, Arabic real words, and reversed Mandarin speech which was generated from the Mandarin real

words, all of which were disyllabic utterances. The pseudo-words were lexical gaps where the syllables exist in Mandarin, but the syllable-tone combination is illegal and yields no meaningful word (e.g., /nei35 min55/). Each type of stimuli consisted of a total of 140 words. Six female bilingual Mandarin-L1/Arabic-L2 speakers were recorded producing the words in Mandarin (real and pseudo) and Arabic (real). Their Arabic proficiency was reported to be near native. The detailed geographic backgrounds of the six talkers are shown in supplemental materials. For each type of stimuli, the disyllabic words were presented to the six talkers via E-prime 2.0 in a random order. In each trial, one word was visually displayed on the computer screen, and the talkers were asked to read aloud the words naturally. Both Chinese characters and pinyin were shown on the screen in the Mandarin real words condition, but only pinyin was shown in the pseudo-word condition. For the Arabic words, the Arabic script was shown on the screen. The six talkers were instructed to read the words with a brief pause between two neighbouring words, in order to ease the segmentation by Praat (Boersma & Weenink, 2014). Prior to the recording, the talkers were given the whole word lists to get familiar with the materials, especially the pseudo-words, to make sure that their production was natural. The recordings were made at a sampling rate of 22050 Hz with 16 bits per sample. Each stimulus type included three repetitions of each word, totalling 7560 words (3 stimulus type × 140 words × 6 talkers × 3 repetitions).

For each talker and each stimulus type, one clearly produced token was selected and segmented from the recordings using Praat (Boersma & Weenink, 2014). For each type of stimuli, the target words were normalized in length to the mean duration of all the words in that condition using PSOLA in Praat (712 ms for Mandarin real words, 875 ms for Mandarin pseudo-words, and 732 ms for Arabic words). The word duration was normalized as a whole unit, proportionally lengthening/shortening the duration of phonetic elements of the original stimuli.

The overall acoustic features such as formant structures, F0 contour and amplitude contour were preserved. The average acoustic intensity of each target word was manipulated to 75 dB. Lastly, the reversed speech condition was created from the Mandarin real words by time-reversing the words using Praat, sharing the same duration and intensity with the Mandarin real words. Among the 140 words in each stimulus type, 120 were used in the talker discrimination task and the remaining 20 were used in the talker identification task described below. The fundamental frequency (F0) distribution (mean and SD) of each talker in each stimulus condition is summarized in Table 2.

Procedure

E-prime 2.0 was used to present the stimuli and collect the responses. All participants completed four blocks of talker AB discrimination tasks, corresponding to the four types of stimuli. Altogether, there were 15 different talker pairs among the six talkers. Within each stimulus condition, 15 different talker pairs were presented in forward and backward order and six same talker pairs were repeated five times, generating equal numbers of different and same talker pairs, which were intermixed and randomly presented in a block. The two words in each trial (same or different talkers) were always different, in order to increase the task demand and encourage listeners to process the talker information at a relatively abstract level. In each trial, a fixation first occurred on the computer screen for 500 ms, followed by the presentation of two stimuli separated by an inter-stimulus-interval of 500 ms via the headphones. Subjects were instructed to judge whether the two words were spoken by the same talker or different talkers by pressing "left arrow" (same) or "right arrow" (different) on a keyboard within 3 seconds. If no responses were detected within 3 seconds, the procedure will proceed to the next trial automatically. No feedback was given in the discrimination task. The stimuli were presented at a comfortable

listening level and the volume was kept identical across the participants. There was a break after every 20 trials within each stimulus condition and also a break after each condition.

The presentation of stimuli was blocked by the condition, with each block containing only trials of one type of stimuli. The presentation order of the four blocks was counterbalanced among the participants as much as possible, which means that the orders differed among the participants within each group. Importantly, the block order was kept identical between matched amusic and control participants in a one to one manner. Before the experiment, one practice block was presented to the participants to familiarize them with the experimental procedure. The practice block consists of eight trials, containing stimuli of the same condition (e.g., reversed speech) as those that occurred in the first experimental block for each participant.

Results

The discrimination performance and reaction time (RT) were analyzed. Trials with null responses were discarded (the amusic group: 3.2%; the control group: 1.9%). The discrimination performance was analyzed in terms of d' (Macmillan & Creelman, 2005), which was calculated as the z-score value of the hit rate ("different" responses to different talker pairs) minus that of the false alarm rate ("different" responses to same talker pairs). Linear mixed-effects models were fitted on the d' scores with *group* (amusics and controls) and *condition* (Mandarin real words, Mandarin pseudo-words, Arabic words and reversed speech) as two fixed effects, and with by-subject random intercept and slope as random effects; two-way interaction was also included as a fixed effect in the models. Models were compared by likelihood ratio tests and p-values were obtained from those tests. RT was measured from the offset of the second stimulus in a talker pair to the time that a response was made. Linear mixed-effects models were fitted on the log-transformed RT data with *group* (amusics and controls) and *condition* (Mandarin real

words, Mandarin pseudo-words, Arabic words and reversed speech) as two fixed effects, and with by-subject random intercept and slope as random effects; two-way interaction was also included as a fixed effect in the models. The procedures were same as those described in the above d' analysis. Figure 1(a) and 1(b) illustrate the d' scores and RT in the discrimination task respectively. The above analyses were performed with R (R Core Team, 2014), using the *lme4* package (Bates, Maechler, & Bolker, 2012), the *lmtest* package (Zeileis & Hothorn, 2002) and the *lsmeans* package (Lenth, 2016).

For the d' scores, there were significant main effects of *group* ($\chi^2(1) = 22.205$, $p < 0.001$) and *condition* ($\chi^2(3) = 71.817$, $p < 0.001$) Post-hoc analyses showed that amusics obtained significantly lower d' scores than controls. The reversed speech condition elicited significantly lower d' scores than the other three conditions (ps < 0.001), while the differences among Mandarin real word, Mandarin pseudo-word and Arabic word conditions were not significant. No other effects were significant.

For the RT, there were significant main effects of *condition* ($\chi^2(3) = 190.9$, $p < 0.001$), and a significant two-way interaction ($\chi^2(3) = 21.691$, $p < 0.001$). The group difference between amusics and controls was not significant in any context. Within the control group, RTs elicited in the Mandarin real word and Arabic word conditions were significantly shorter than those in the Mandarin pseudo-word and reversed speech conditions (ps < 0.001). Within the amusic group, Mandarin real word condition elicited the shortest RT compared to the other three conditions (ps < 0.001), and RT in Mandarin pseudoword condition was significantly longer than Arabic and reversed speech conditions. No other effects were significant.

Discussion

The results from Experiment 1 revealed an effect of group in talker discrimination, with lower d' scores observed in amusics than in controls. Individuals with amusia were worse than controls at discerning talker differences across all four contexts. In the discrimination task, it is not necessary for listeners to establish the talker category to perform the task. Instead, it requires the auditory comparison of talkers' voices within one trial. Although the task demand was relative low in the discrimination task, amusics demonstrated inferior performance than controls, suggesting that talker voice processing is impaired in amusic listeners.

According to the language familiarity effect, performance should be best in the Mandarin real words, followed by Mandarin pseudowords, Arabic words and reversed speech. This prediction was partly confirmed by the performance pattern. As predicted, the reversed speech condition where the linguistic content was completely destroyed elicited significantly lower d' and longer RT than the other three conditions. Interestingly, the Arabic condition elicited comparably high accuracy and fast RT as the Mandarin real word condition, and the Mandarin pseudoword condition also elicited similarly high accuracy as the Mandarin real word condition, though the RT was considerably longer. Since the language familiarity effect was reported in previous studies that mostly adopted a talker identification task (Goggin et al., 1991; Perrachione, Del Tufo, et al., 2011), it is possible that that the magnitude of the language familiarity effect may not be as strong in the talker discrimination task.

Taken together, results from Experiment 1 confirmed that amusics were degraded in talker discrimination compared with controls in all four types of conditions.

Experiment 2. Talker category learning

Method

Participants

Same as described in Experiment 1.

Stimuli

Twenty out of 140 target words in each stimulus type produced by the six female talkers that were described in Experiment 1 were used in Experiment 2.

Procedure

E-prime 2.0 was used to present the stimuli and collect the responses. There were two phases in this experiment: a learning phase and a following test phase. All the four conditions were included in both the learning and test phase. In the *learning* phase, there were 10 words in each condition. All the participants completed a talker identification training task in which they learned to identify the voices of six female speakers, each of whom was presented as a cartoonlike characters on a computer screen. The task was a six-alternative forced-choice identification task with feedback. In each trial, participants were presented with a word, and six talkers' names with the cartoon images were presented on the computer screen simultaneously. The subjects were then instructed to indicate which of the six talkers produced the word by pressing the number keys 1-6 corresponding to the talkers one to six. After each response, a feedback screen was presented. The feedback information included: (1) accuracy: if the response made by the participant was correct (in blue text) or incorrect (in red text), (2) the correct response: the talker who produced the word was shown on the screen with the corresponding cartoon image and name. If the response was correct, the participant proceeded to the next trial, but if the response was incorrect, the incorrect trial was repeated until a correct response was selected. The ten stimuli were repeated 12 times, which gave rise to 120 trials in each stimulus condition.

Like in Experiment 1, the stimuli of each condition were presented in a separate block. The presentation order of the four conditions was counterbalanced across the participants and kept identical between matched amusic and control participants. Before each task, a practice block, containing 12 trials, was given to the participants to familiarize them with the procedure.

Immediately after the learning phase, there was a *test* phase. Twenty target words were used as stimuli in each condition. Among them, ten were from the learning phase, and another ten were new to the listeners. It should be noted that the test phase assessed the outcomes in the learning phase, by using the same ten words as in the learning phase. The task was the same six-alternative forced-choice identification task as used in the learning phase, but no feedback was given and the participants were required to make the response within 5 seconds. If no responses were made within the time limit, the procedure will proceed to the next trial automatically. The stimuli were presented at a comfortable listening level and the volume was kept constant to all of the participants. The accuracy and RT data were collected. The presentation order of the four conditions was counterbalanced across the participants and kept identical between matched amusics and controls.

Results

Trials with null responses were discarded from the analysis (the amusic group: 0.8%; the control group: 1.1%). For the talker identification accuracy, generalized mixed-effects models were fitted on the responses to each trial (1 or 0) with *group* (amusics and controls), *condition* (Mandarin real words, Mandarin pseudo-words, Arabic words and reversed speech) and *word type* (trained and untrained words) as three fixed effects, and with by-subject random intercept and slope as random effects; two-way and three-way interactions were also included as a fixed effect in the models. Models were compared by likelihood ratio tests and p-values were obtained

from those tests, using the same procedure as described in Experiment 1. Figure 2(a) shows the talker identification accuracy.

There were significant main effects of *group* ($\chi^2(1) = 14.901, p < 0.001$), *condition* ($\chi^2(3) = 353.99, p < 0.001$), and *word type* ($\chi^2(1) = 14.846, p < 0.001$). Identification accuracy in the amusic group was significantly lower than the control group. Accuracy on the trained words was significantly higher than the untrained words. For the effect of *condition*, post hoc analysis showed that the identification accuracy in the Mandarin real word context was significantly higher than the other three types (Mandarin pseudoword vs. Mandarin real word: $z = -5.522$, $p < 0.001$; Arabic vs. Mandarin real word: $z = -10.33$, $p < 0.001$; reversed speech vs. Mandarin real word: $z = -17.946$, $p < 0.001$). Mandarin pseudoword context also elicited better talker identification accuracy than the other two conditions (Arabic vs. Mandarin pseudoword: $z = -4.840$, $p < 0.001$; Reversed speech vs. Mandarin pseudoword: $z = -12.564$, $p < 0.001$). Finally, Arabic word context elicited higher accuracy than the reversed speech condition ($z = -7.777$, $p < 0.001$). No other effects were significant.

Linear mixed-effects models were fitted on the log-transformed RT data that were measured from the offset of the stimuli with *group*, *condition* and *word type* as three fixed effects, and with by-subject random intercept and slope as random effects; two-way and three-way interactions were also included as a fixed effect in the models. Models were compared by likelihood ratio tests with the same procedures as described in Experiment 1. Figure 2(b) shows the identification RT.

There were significant main effects of *condition* ($\chi^2(3) = 123.4, p < 0.001$) and *word type* ($\chi^2(1) = 5.847, p = 0.015$). RT in the trained words was significantly shorter than in the untrained words. RT in the reversed speech condition was the shortest, significantly shorter than the other

three types of contexts (ps < 0.001). The Mandarin pseudoword condition elicited longer RT than Mandarin real word condition (p = 0.001). No other effects were significant.

Discussion

The results in Experiment 2 also showed a clear group effect. Similar to Experiment 1, amusics were less accurate at identifying the talkers' voices than controls in all four types of conditions, suggesting that amusics were also impoverished in the relatively higher-level of talker processing, which requires listeners to establish the talker category. It is possible that listeners with amusia were less efficient at utilizing phonological and phonetic cues in the context to support talker identification due to their impairment in the phonetic and phonological processing of language.

Different from Experiment 1, the results in Experiment 2 demonstrated a robust language familiarity effect. The accuracy was highest in the Mandarin real word condition, followed by the Mandarin pseudo-word, Arabic word and reserved word condition, indicating that the talker identification accuracy significantly deteriorated when the available linguistic content decreased or the phonological characteristics of the speech materials became less familiar. These patterns confirmed the previous findings that talker identification is facilitated when listeners can comprehend or are familiar with the speech materials (Goggin et al., 1991; Perrachione & Wong, 2007; C. P. Thompson, 1987).

To conclude, results from Experiment 2 demonstrated a strong language familiarity effect. They also showed that amusics were impoverished in talker identification under all conditions.

Experiment 3. Phonological memory test - nonword repetition

In a typical nonword repetition task, participants are instructed to listen to a nonword and repeat it exactly as it was heard. This task requires the participants to access and maintain new

phonological codes, and is thought to measure the quality of the phonological representations in working memory (e.g., Pelczarski & Yaruss, 2016). This task has been extensively used to measure phonological memory in both children and adults (Archibald & Gathercole, 2007; Dollaghan & Campbell, 1998; Long et al., 2016; Pelczarski & Yaruss, 2016; Williams, Payne, & Marshall, 2013). In this experiment, we used the nonword repetition task to assess phonological memory in amusics, with an aim to examine whether inefficient or impaired phonological memory may lead to difficulty in the maintenance of the phonological code for subsequent use in speech and language planning, thereby contributing to the deficient talker processing in amusia.

Method

Participants

Same as in Experiment 1 and 2.

Materials and procedure

Previous studies used nonwords with various word lengths, including monosyllabic, disyllabic, trisyllabic, and quadrisyllabic non-words (Dollaghan & Campbell, 1998; Edwards, Beckman, & Munson, 2004; Gathercole, Willis, Emslie, & Baddeley, 1991; Williams et al., 2013). As suggested in Gathercole et al., (1991), in which the number of syllables varied from one to four, word length influenced the performance of nonword repetition. Participants' performance declined when the non-words were disyllabic and dropped below 60% when the syllable number reached four. To make sure that the task is demanding but at the same time avoiding a possible floor effect, we designed 20 disyllabic and 20 trisyllabic nonwords in the current study, which is compatible with the word length in previous studies. The syllables in the nonwords do not exist in Mandarin. The syllable structures were CVCV and CVCVCV for the two types of nonwords. To increase the task demand and phonological complexity, all vowels in the nonwords were

diphthongs in Mandarin. Furthermore, the nonwords carried lexical tone information, which may exert more demand on the memory. A female native speaker of Mandarin recorded the stimuli. The recordings were made at a sampling rate of 22050 Hz with 16 bits per sample. Each item was repeated three times by the speaker. One clearly produced token of each target was selected and segmented from the recordings using Praat (Boersma & Weenink, 2014). The disyllabic nonwords were normalized to 930 ms and the three-syllable nonwords were normalized to 1460 ms, which roughly equalled to the mean duration of selected tokens in each condition. The average acoustic intensity of each target word was scaled to 75 dB.

E-prime 2.0 was used to present the stimuli and record the response from the controls and amusics. Disyllabic and trisyllabic nonwords were presented in two separate blocks, and the items were randomized within each block. Participants were informed that they would hear some nonwords and asked to repeat them as accurately as possible. The stimuli were presented through headphones and the productions from participants were recorded automatically via the Soundin function in E-prime. Each response item was stored as a wav file. The recordings were transcribed and scored by a trained phonetician afterwards. A simple binary scoring procedure was used to score the repetition accuracy, in which a correct repetition attempt was scored as 1 and an attempt in which one or more phonemes (including consonant, vowel and lexical tone) were incorrectly produced was scored as 0. The phonetician was independent and naïve to the experimental design. The first author checked the scoring and transcription of 10% of the production data (two amusics and two controls) and confirmed that the scoring and transcription by the phonetician were accurate.

Results and discussion

As we are interested in the overall difference between amusics and controls in the phonological memory performance, we conducted independent samples t-test, instead of *group* by *word length* ANOVA. Figure 3(a) demonstrates the nonword repetition accuracy. The results revealed that amusics repeated the nonwords significantly less accurately than controls ($t(38) = -4.669$, $p < 0.001$), confirming that amusics have reduced phonological memory. This means that the quality of the phonological representations held in working memory in the amusic brain may be degraded compared with controls.

Furthermore, linear and logistic regression models were fitted to examine to what extent the participants' performance on talker discrimination indexed by d' scores and on talker identification indexed by accuracy could be predicted by their phonological memory capacity. The d' score and identification accuracy (1 or 0) was the input in the linear and logistic regression models respectively, and phonological memory capacity was the predictor. The discrimination index d' and identification accuracy are plotted as a function of phonological memory in Figure 3(b) and (c). The results demonstrated that the scores of the nonword repetition task could account for the participants' performance in the talker discrimination task ($p < 0.001$) and in the talker identification task ($p < 0.001$).

In addition to the phonological memory capacity, we also examined to what extent the participants' musical abilities can predict their performance on talker discrimination and identification by fitting multiple linear regression and logistic regression models. In the linear regression models, the input was d' scores and the predictors were the scores in the six MBEA subtests (scale, contour, interval, rhythm, meter and memory), with the two groups collapsed. The six predictors were added to the models consecutively, and likelihood ratio tests were conducted to compare two models with and without a certain predictor. The d' scores are plotted

as a function of the six subtests of MBEA in Figure 4. The multiple linear regression models showed that the MBEA scores were significantly associated with the d' ($p < 0.001$). Among the six subtests, only the melodic memory subtest contributed significantly to the d' ($p = 0.016$). In the logistic regression model, the input was identification accuracy (1 or 0), and the predictors were the participants' accuracy in the six MBEA subtests. The identification accuracy is plotted as a function of the six subtests in Figure 5. The results showed that the effects of the six subtests of MBEA were all significant (ps < 0.001), with the pitch-related subtests and memory subtest eliciting the stronger effect. These patterns suggested that musical abilities could predict the talker processing performance well, especially in the talker identification task, where higher-order cognitive processes were involved.

We further included both the scores of the six MBEA subtests and nonword repetition accuracy as predictors. The results showed that nonword repetition accuracy can account for the additional variance beyond the effects of musical abilities in the talker identification task ($p < 0.001$), but not in the talker discrimination task. These results may suggest that phonological memory capacity plays a more important role in the talker identification task.

General discussion

Identification of a talker's voice through the speech signal is an important skill for human beings. The current study examined the amusics' ability to discriminate and recognize talkers' voices in four conditions differing in the amount of available phonetic cues and phonological representations. We hypothesized that in the non-native conditions such as foreign language and reversed speech conditions, where only *language-independent information* of a talker's voice is present, amusics would demonstrate significantly worse performance than controls. While in the native speech conditions (Mandarin real words and pseudo-words), *language-specific*

linguistic/phonological information is tied to a talker's voice. In these conditions, if amusics processed native speech materials in the phonological mode (Strange, 2011), they are expected to exhibit a language familiarity effect combined with worse performance than controls, due to their reported deficiency in phonological processing. On the other hand, if amusics processed native speech in the phonetic/auditory mode, they are expected to show no language familiarity effect, and a global impairment in talker voice processing.

Previous studies have widely reported the importance of language familiarity in talker identification (Goggin et al., 1991; Goldstein et al., 1981; Hollien et al., 1982; Perrachione & Wong, 2007; C. P. Thompson, 1987; Winters et al., 2008). In the current study, similar effects were obtained. Listeners demonstrated better talker identification performance in the conditions with richer phonological cues, such that the highest talker identification accuracy was observed in the Mandarin real word condition, followed by the Mandarin pseudoword condition, the Arabic word condition, and finally the reversed Mandarin word condition. As mentioned earlier, both Mandarin real word and pseudoword conditions were native speech contexts with phonological cues (i.e., native phonemes and tones), but in the pseudoword condition, the combinations of the syllables and tones were illegal, which reduced the semantic content. On the other hand, the Arabic word condition contained non-native phonological representations to Mandarin listeners. The reversed speech condition sounded like foreign language, but all legal phonological cues were destroyed. The results obtained in the current study are largely consistent with previous findings that talker identification is facilitated by native language contexts, and this effect is reduced when the linguistic content of the speech is gradually eliminated. Indeed, previous findings also revealed that the reversed speech condition generated the lowest accuracy scores (Goggin et al., 1991), and when replacing all syllables of a spoken message with the

syllable [ma] but maintaining the global prosodic patterns, talker identification performance also deteriorated (Schiller, Koster, & Duckworth, 1997). It should also be noted that in the current study, the language familiarity effect was more robust in the talker identification task than the discrimination task, as the discrimination task did not elicit a clear pattern of language familiarity effects, probably because the discrimination task tapped into a relatively low-level of talker processing.

Compared to controls, amusics demonstrated degraded talker processing performance in terms of the talker discrimination and identification accuracy, and in non-native and native speech conditions. The inferior performance of amusics in talker discrimination and non-native conditions can be largely explained by their low-level pitch-processing deficit, which presumably impedes their ability to utilize *language-independent* phonetic cues for talker processing. As for amusics' worse performance in talker identification, especially in native speech conditions, it is more likely to be due to their degraded ability to use phonological representations to faciliate talker identification. Note that amsuics showed a native language advantage, exhibiting better performance in native speech conditions than in non-native conditions, a pattern largely similar to the performance of controls. This result indicates that amusics did make use of phonological cues in native speech conditions (i.e., the phonological mode) to some extent, but their worse phonological processing ability presumbaly led to worse talker identification in these conditions. That being said, we cannot completely rule out the possibility that amusics processed the native speech conditions in the phonetic/auditory mode, and their general pitch deficit or other factors such as working memory at least partly contributed to their worse performance in talker identification in native conditions.

Altogether, these results suggest that Mandarin-speaking amusics are likely to be impaired in talker discrimination and identification in both levels: talker processing via utilizing phonological cues (*language-specific information*) in native speech contexts and talker processing via analyzing phonetic cues (*language-independent information*) in non-native speech contexts. Therefore, corroborating previous findings of impaired phonological processing in amusics (Huang et al., 2015; Jiang, Hamm, Lim, Kirk, & Yang, 2012; Wang & Peng, 2014; Zhang et al., 2017), results of the current study suggested that utilizing phonological cues to recognize a talker's voice - another dimension of the speech signal, is also impaired. As mentioned earlier, it has been found that both Mandarin-speaking and Cantonese-speaking amusics exhibited no or reduced benefit for between-category discriminations of tone stimuli relative to within-category discriminations, which indicates lack of or reduced categorical perception of tones (Jiang, Hamm, Lim, Kirk, & Yang, 2012; Zhang et al., 2017). The findings of the current study converged with these results in suggesting that amusia is more than a pitch-processing deficit, which also extends to the phonological level and affects the analysis of the phonological cues to process the talker dimension in speech signals.

Interestingly, the finding that amusics were able to make use of phonological cues to some extent in talker identifiation in native speech conditions is divergent from that found in dyslexia (Perrachione, Del Tufo, et al., 2011). It was reported that English-speaking listeners with dyslexia did not show better performance in identifying voices speaking the native language than a foreign language, in contrast to English-speaking listeners with normal reading ability who demonstrated a native language advantage. Furthermore, dyslexic listeners were significantly impaired compared with controls in their ability to recognize voices speaking their native language, but were as accurate as controls when identifying the voices speaking the non-native

language. These different patterns point to potential differences in the underlying impairment of these two disorders. Converging evidence underscores a core deficit in the phonological component in dyslexia (Fuerst, 2008; Goswami, Gerson, & Astruc, 2010), in that the access to the phonological representation of words is compromised. In contrast, amusia is an innate pitch-processing disorder. The deficit in amusia is primarily driven by impairment in low-level auditory pitch processing, which at the same time prevails to the phonological processing of speech sounds. Given the deficits of amusics in auditory pitch processing, it is not surprising that it affects the use of auditory cues, especially pitch cues, in non-native speech contexts, leading to deficient talker processing in those conditions. By comparing the different patterns of dyslexia and amusia, the findings of the current study may provide insight into the relationship between language impairment and talker processing.

We found that Mandarin-speaking amusics showed smaller phonological memory capacity compared with controls. While previous studies have reported reduced phonological awareness in amusia (Jones et al., 2009; Sun et al., 2017), these results extended the current understanding of amusia by demonstrating that phonological memory is also degraded in amusics. Sun et al. (2017) found that a subgroup of Australian English speakers with severe pitch deficits showed significantly lower scores in the subtest of Elision, which primarily measures phonological awareness. However, amusics' performance in the other subtests that measure phonological short-term memory and rapid naming ability was not significantly different from controls. The results in current study demonstrated a significant group difference in phonological memory, and the participants' capacity in phonological memory was further found to account for their performance in talker identification tasks, which is consistent with previous findings on dyslexia (Long, Fox, & Jacewicz, 2016; Perrachione, Tufo, et al., 2011). The nonword repetition task

measures the quality of phonological representations held in working memory and how well the participants can access and maintain novel phonological codes (i.e., nonwords) from the phonological store. The degraded nonword repetition performance in amusics may suggest that amusics were less able to access and maintain novel phonological codes from the phonological store. It has been established that the ability to recognize talkers' voices depends on the ability to compute the differences between the incidental phonetics of a specific articulation and the abstract phonological representations of the words that the articulation contains (Perrachione, Del Tufo, et al., 2011). It is possible that their impoverished ability to access and maintain the phonological code may have impeded the utilization of native phonological representations in the process of talker identification.

Conclusion

Altogether, the results in the present study provided the first evidence that individuals with amusia were impaired in human voice identification – an important trait of human beings, and expanded our understanding of the nature and scope of amusia. As plenty of previous studies have revealed that amusics were less efficient in the dimension of linguistic pitch processing in terms of lexical tone and intonation perception (Hutchins, Gosselin, & Peretz, 2010; Jiang, Hamm, Lim, Kirk, & Yang, 2012; Shao et al., 2019; Zhang et al., 2018, 2017), our results showed for the first time that amusics are also impaired in talker voice processing and filled the gap in the literature. The degraded performance in amusia was found in both native speech conditions that contained available phonological representations to facilitate talker processing, and non-native conditions that only contained auditory cues, implying that amusia is not only a pitch-processing disorder, but may also have a negative impact on the phonological analysis of

speech during talker voice processing to some extent. These results also provided evidence that the abilities in speech perception and voice perception may be closely integrated in speakers with amusia. While we conclude that deficits in phonetic and phonological processing best explained the results of the current study, the possibility cannot be completely ruled out that other factors, such as attention, may have contributed to the observed group difference in talker voice processing to some extent. A previous study has reported an attention deficit in approximately 40% of the amusic individuals (Jones, Zalewski, Brewer, Lucker, & Drayna, 2009). Future studies may further examine whether and how factors other than phonetic and phonological processing affect talker voice processing in amusics.

Acknowledgements

We thank Ms. Yiran Wei for help with data collection.

References

Archibald, L. M. D., & Gathercole, S. E. (2007). Nonword repetition and serial recall: Equivalent measures of verbal short-term memory? *Applied Psycholinguistics*, *28*(4), 587–606. https://doi.org/10.1017/S0142716407070324

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: linear mixed-effects models using S4 classes. R package version 0.999375-42. 2011.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer.

Creel, S. C., & Bregman, M. R. (2011). How Talker Identity Relates to Language Processing. *Linguistics and Language Compass*. https://doi.org/10.1111/j.1749-818X.2011.00276.x

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment.

Journal of Speech, Language, and Hearing Research, *41*(5), 1136–1146.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*.

Foxton, J. M., Dean, J. L., Gee, R., Peretz, I., & Griffiths, T. D. (2004). Characterization of deficits in pitch perception underlying "tone deafness." *Brain*, *127*(4), 801–810. https://doi.org/10.1093/brain/awh105

Fuerst, D. R. (2008). Learning Disabilities: From Identification to Intervention. *Child Neuropsychology*, *14*(3), 286–288. https://doi.org/10.1080/09297040701455171

Gathercole, S. E., Willis, C., Emslie, H., & Baddeley, A. D. (1991). The influences of number of syllables and wordlikeness on children's repetition of nonwords. *Applied Psycholinguistics*, *12*(3), 349–367. https://doi.org/10.1017/S0142716400009267

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, *19*(5), 448–458. https://doi.org/10.3758/BF03199567

Goldstein, A. G., Knight, P., Bailis, K., & Conover, J. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society*, *17*(5), 217–220. https://doi.org/10.3758/BF03333718

Goswami, U., Gerson, D., & Astruc, L. (2010). Amplitude envelope perception, phonology and prosodic sensitivity in children with developmental dyslexia. *Reading and Writing*, *23*(8), 995–1019. https://doi.org/10.1007/s11145-009-9186-6

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception and Psychophysics*, *59*(5), 675–692.

https://doi.org/10.3758/BF03206015

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*, *10*, 139–148. https://doi.org/10.1121/1.1914230

Huang, W.-T., Liu, C., Dong, Q., & Nan, Y. (2015). Categorical perception of lexical tones in mandarin-speaking congenital amusics. *Frontiers in Psychology*, *6*(829). https://doi.org/10.3389/fpsyg.2015.00829

Hutchins, S., Gosselin, N., & Peretz, I. (2010). Identification of changes along a continuum of speech intonation is impaired in congenital amusia. *Frontiers in Psychology*, *1*(DEC), 1–8. https://doi.org/10.3389/fpsyg.2010.00236

Hyde, K. L., Peretz, I., Hyde, K. L., & Peretz, I. (2004). Brains That Are out of Tune but in Time, 1–6. https://doi.org/10.1111/j.0956-7976.2004.00683.x

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). 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

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., & Yang, Y. (2012). Impaired categorical perception of lexical tones in Mandarin-speaking congenital amusics. *Memory and Cognition*, *40*(7), 1109–1121. https://doi.org/10.3758/s13421-012-0208-2

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

Lee, C. (2009). Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *The Journal of the Acoustical Society of America*, *125*(2), 1125–1137. https://doi.org/10.1121/1.3050322

Lee, C., Tao, L., & Bond, Z. S. (2010). Identification of multi-speaker Mandarin tones in noise by native and non-native listeners. *SPEECH COMMUNICATION*, 1–11. https://doi.org/10.1016/j.specom.2010.01.004

Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. https://doi.org/10.18637/jss.v069.i01

Liu, F., Chan, A. H. D., Ciocca, V., Roquet, C., Peretz, I., & Wong, P. C. M. (2016). Pitch perception and production in congenital amusia: Evidence from Cantonese speakers. *The Journal of the Acoustical Society of America*, *140*(1), 563–575. https://doi.org/10.1121/1.4955182

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

Long, G. B., Fox, R. A., & Jacewicz, E. (2016). Dyslexia Limits the Ability to Categorize Talker Dialect. *Journal of Speech Language and Hearing Research*, *59*(5), 900. https://doi.org/10.1044/2016_JSLHR-S-15-0106

Lu, X., Ho, H. T., Liu, F., Wu, D., & Thompson, W. F. (2015). Intonation processing deficits of emotional words among Mandarin Chinese speakers with congenital amusia: An ERP study. *Frontiers in Psychology*, *6*(MAR), 385. https://doi.org/10.3389/fpsyg.2015.00385

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah:

Lawrence Erlbaum Associates.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390. https://doi.org/10.3758/BF03210878

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

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Y. Tohkura Sagasaka, and E. Vatikiotis-Bateson (Ed.), *Speech Perception, Speech Production, and Linguistic Structure* (pp. 113–134). Tokyo: OHM.

Pelczarski, K. M., & Yaruss, J. S. (2016). Phonological memory in young children who stutter. *Journal of Communication Disorders*, *62*, 54–66. https://doi.org/https://doi.org/10.1016/j.jcomdis.2016.05.006

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/http://dx.doi.org/10.1016/S0896-6273(01)00580-3

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

Peretz, I., Gosselin, N., Tillmann, B., Cuddy L., L., Gagnon, B., Trimmer G., C., … Bouchard, B. (2008). On-line identification of congenital amusia. *Music Perception*, *25*(4), 331–343. https://doi.org/10.1525/mp.2008.25.4.331

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

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science*, *333*(6042), 595. https://doi.org/10.1126/science.1207327

Perrachione, T. K., del Tufo, S. N., Ghosh, S. S., & Gabrieli, J. D. E. (2011). Phonetic variability in speech perception and the phonological deficit in dyslexia. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*, (August), 1578–1581.

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899–1910. https://doi.org/10.1016/j.neuropsychologia.2006.11.015

Pfeifer, J., & Hamann, S. (2015). Revising the diagnosis of congenital amusia with the Montreal Battery of Evaluation of Amusia. *Frontiers in Human Neuroscience*, *9*, 161. https://doi.org/10.3389/fnhum.2015.00161

Schiller, N. O., Koster, O., & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *Forensic Linguistics*, *4*(1), 1350–1771. https://doi.org/10.1558/ijsll.v4i1.1

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. (2018). Context integration deficit in tone perception in Cantonese speakers with congenital amusia. *The Journal of the Acoustical Society of America*, *144*(4), EL333–EL339. https://doi.org/10.1121/1.5063899

Shao, J., & Zhang, C. (2019). Talker normalization in typical Cantonese-speaking listeners and congenital amusics: Evidence from event-related potentials. *NeuroImage: Clinical*, *23*, 101814. https://doi.org/10.1016/j.nicl.2019.101814

Shao, J., Zhang, C., Peng, G., Yang, Y., & Wang, W. S.-Y. (2016). Effect of noise on lexical tone perception in Cantonese-speaking amusics. In *Proceedings of the Interspeech*. San Francisco, U.S.A.

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, *39*(4), 456–466. https://doi.org/10.1016/j.wocn.2010.09.001

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *The Journal of the Acoustical Society of America*, *60*(1), 213–224.

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

Team, R. C. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*(2), 121–131. https://doi.org/10.1002/acp.2350010205

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

Tillmann, B., Lévêque, Y., Fornoni, L., Albouy, P., & Caclin, A. (2016). Impaired short-term memory for pitch in congenital amusia. *Brain Research*, *1640*, 251–263. https://doi.org/10.1016/j.brainres.2015.10.035

Tillmann, B., Schulze, K., & Foxton, J. M. (2009). Congenital amusia: A short-term memory deficit for non-verbal, but not verbal sounds. *Brain and Cognition*, *71*(3), 259–264. https://doi.org/http://dx.doi.org/10.1016/j.bandc.2009.08.003

Wang, X., & Peng, G. (2014). Phonological processing in Mandarin speakers with congenital amusia. *Journal of the Acoustical Society of America*, *136*(6), 3360–3370.

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781–790.

Williams, D., Payne, H., & Marshall, C. (2013). Non-word repetition impairment in autism and specific language impairment: Evidence for distinct underlying cognitive causes. *Journal of Autism and Developmental Disorders*, *43*(2), 404–417. https://doi.org/10.1007/s10803-012-1579-8

Winters, S. J., Levi, S. V, & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America*, *123*(6), 4524–4538. https://doi.org/10.1121/1.2913046

Wong, P. C. M., Ciocca, V., Chan, A. H. D., Ha, L. Y. Y., Tan, L. H., & Peretz, I. (2012). Effects of culture on musical pitch perception. *PLoS ONE*, *7*(4), e33424. https://doi.org/10.1371/journal.pone.0033424

Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, *46*(2), 413–421.

Zeileis, A., & Hothorn, O. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10.

Zhang, C., Shao, J., & Chen, S. (2018). Impaired perceptual normalization of lexical tones in Cantonese-speaking congenital amusics. *The Journal of the Acoustical Society of America*, *144*(2), 634–647. https://doi.org/https://doi.org/10.1121/1.5049147

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

Table 1. Demographic characteristics of the amusic and control participants.

	Amusics	Controls
No. of participants	20 (10 M, 10 F)	20 (10 M, 10F)
Age (range)	23.81 ± 3.1 years (19.1-32.0 years)	23.4 ± 3.2 years (19.0-31.2 years)
MBEA (SD)		
Scale	61.1 (12.7)	92.2 (6.2)
Contour	64.3 (11.0)	95.2 (4.5)
Interval	61.2 (7.7)	93.2 (5.0)
Rhythm	69.7 (15.3)	95.6 (5.7)
Meter	56.7 (9.7)	81.8 (12.2)
Memory	76.7 (12.7)	96.8 (3.1)
Global	65 (5.5)	92.5 (3.9)

Note: Amusics and controls were identified using the Montreal Battery of Evaluation of Amusia (MBEA) (Peretz et al., 2003). Amusics scored lower than 71% in the global score, which is the mean of all six subtests, whereas controls scored higher than 80%. M = male; F = female.

Table 2. Mean and SD (in parenthesis) of F0 (Hz) for each talker in each stimulus condition.

	Mandarin & Reversed Mandarin words	Pseudowords	Arabic words
Talker 1	244.5 (40.7)	229.5 (35.2)	233.5 (30.3)
Talker 2	248.2 (29.1)	222.6 (24.1)	219.2 (25.9)
Talker 3	247.7 (36.3)	235.3 (36.9)	235.6 (21.3)
Talker 4	260.0 (41.7)	243.6 (36.4)	230.4 (32.6)
Talker 5	209.1 (33.6)	212.4 (40.1)	219.9 (34.9)
Talker 6	220.4 (34.7)	175.2 (34.2)	208.5 (32.2)

Figure 1. Results of the talker discrimination task for amusics (in blue) and controls (in red) in the four stimulus conditions: Mandarin real word, Mandarin pseudoword, Arabic word and Reversed Mandarin word. (a) d' score; (b) RT. In each condition, the amusics' results are displayed on the left, and controls' results are displayed on the right.

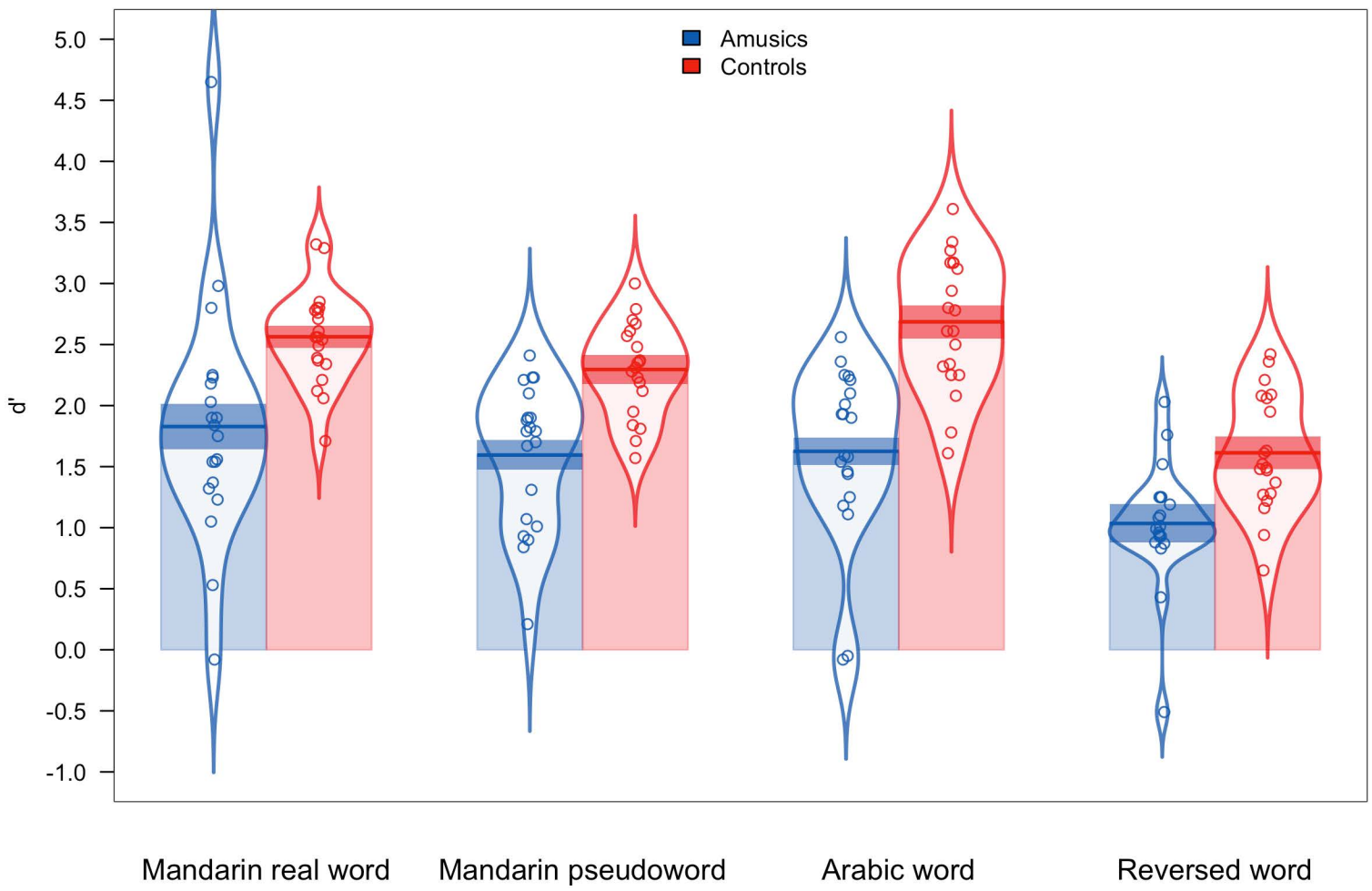
Figure 2. Results of the talker identification task for amusics (in blue) and controls (in red) in the four stimulus conditions: Mandarin real word, Mandarin pseudoword, Arabic word and Reversed Mandarin word. (a) Identification accuracy; (b) RT. In each condition, the amusics' results are displayed on the left, and controls' results are displayed on the right.

Figure 3. Results of the nonword repetition task. (a) Accuracy of the nonword repetition task for amusics (in blue, on the left) and controls (in red, on the right); (b) talker discrimination sensitivity index d' score plotted as a function of the phonological memory capacity for amusics (in blue circle) and controls (in red triangle); (c) talker identification accuracy plotted as a function of the phonological memory capacity for amusics (in blue circle) and controls (in red triangle).

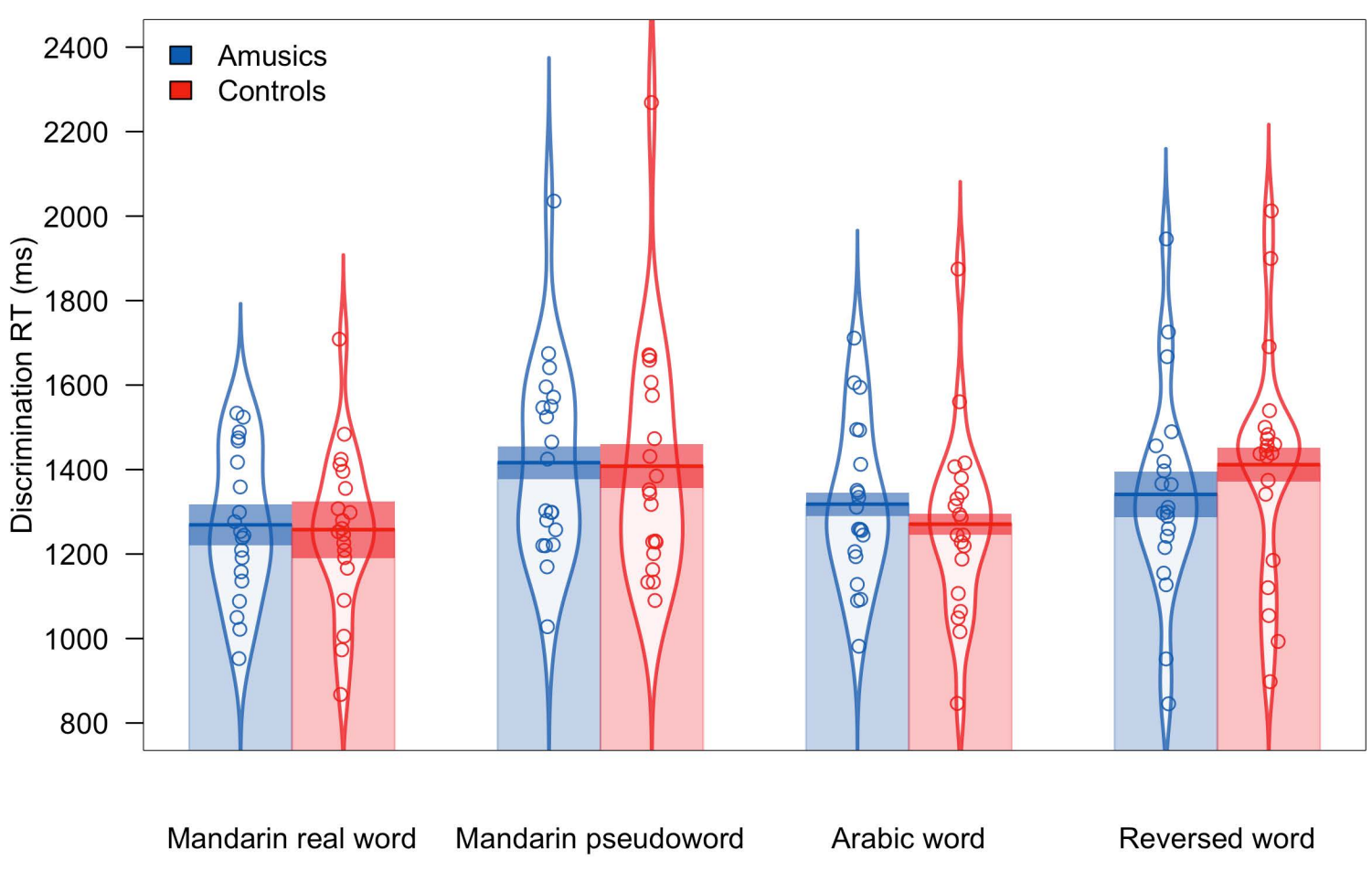
Figure 4. Talker discrimination accuracy plotted as a function of the six subtests of MBEA for amusics (in blue circle) and controls (in red triangle).

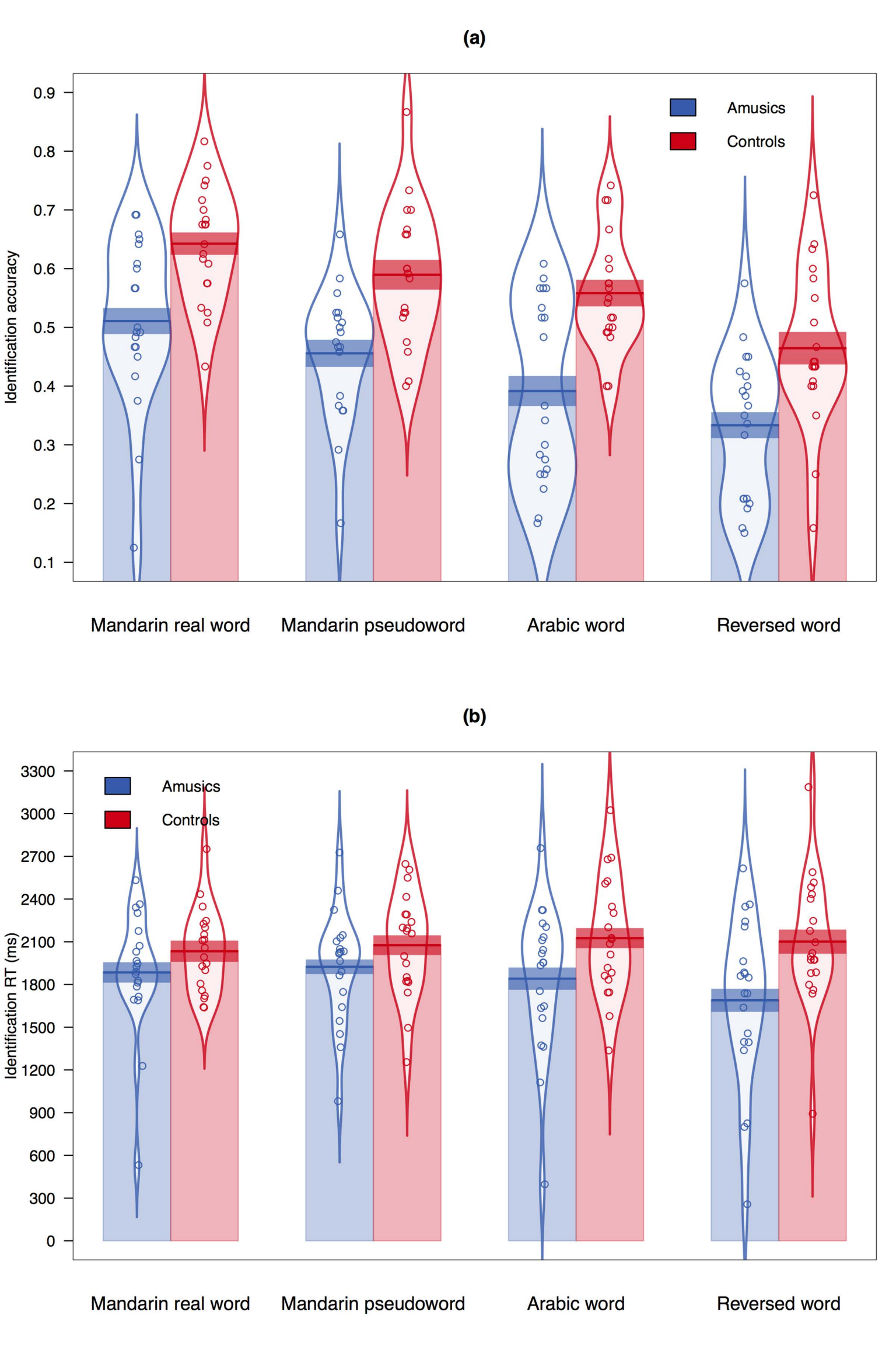
Figure 5. Talker identification accuracy plotted as a function of the six subtests of MBEA for amusics (in blue circle) and controls (in red triangle).

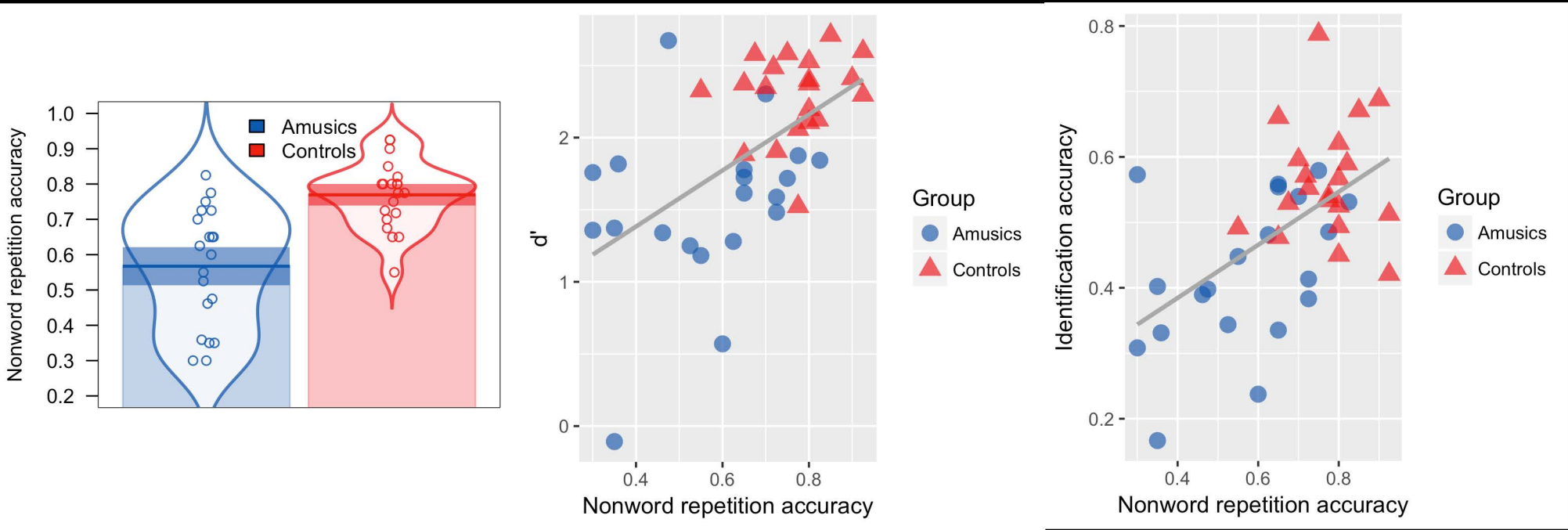
(a)

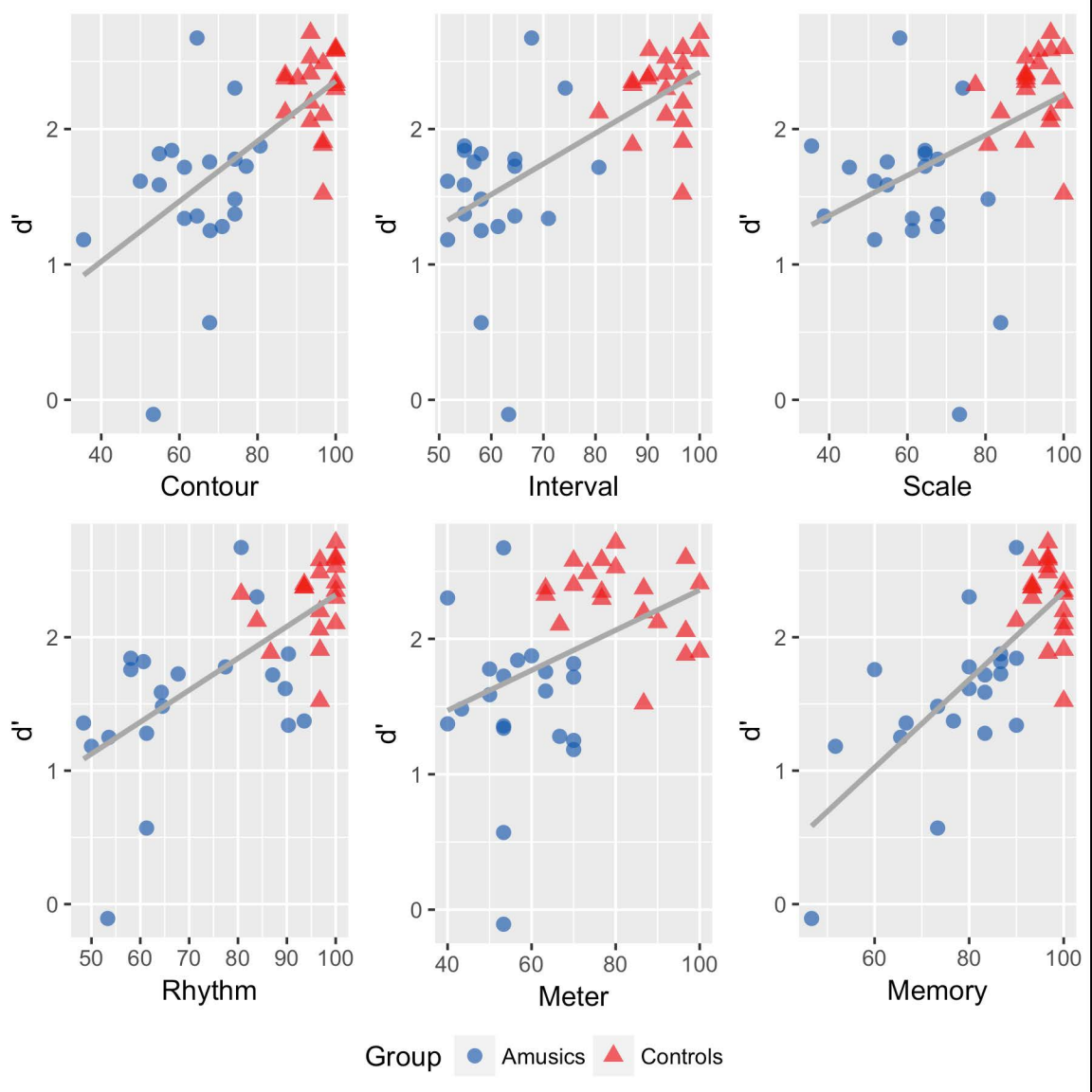

Amusics
Controls
d'
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
Mandarin real word
Mandarin pseudoword
Arabic word
Reversed word

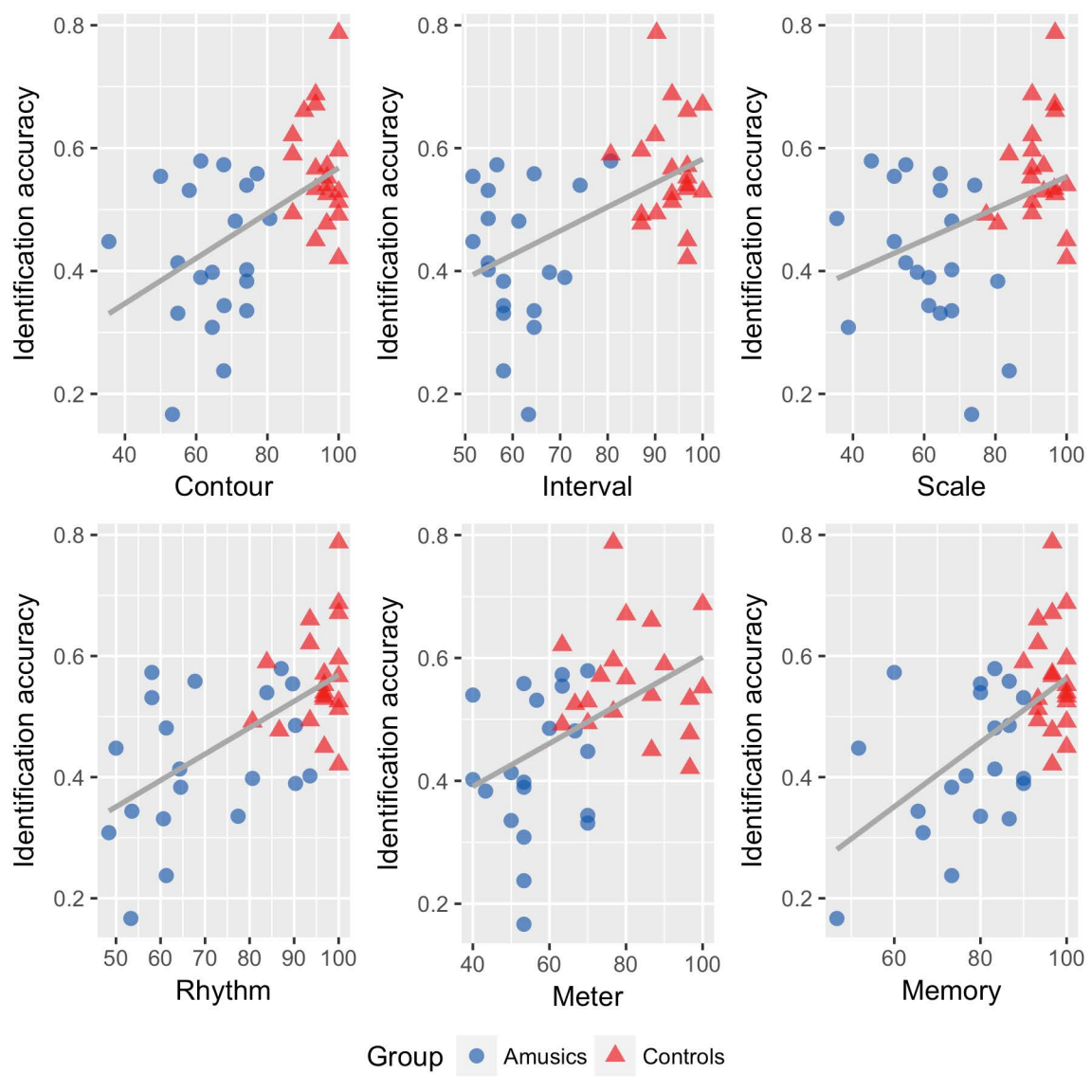
(b)


Amusics
Controls
Discrimination RT (ms)
2400
2200
2000
1800
1600
1400
1200
1000
800
Mandarin real word
Mandarin pseudoword
Arabic word
Reversed word


(a)
Identification accuracy
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Amusics
Controls
Mandarin real word
Mandarin pseudoword
Arabic word
Reversed word
(b)
Identification RT (ms)
3300
3000
2700
2400
2100
1800
1500
1200
900
600
300
0
Amusics
Controls
Mandarin real word
Mandarin pseudoword
Arabic word
Reversed word


Nonword repetition accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Amusics
Controls
d'
2
1
0
0.4
0.6
0.8
Nonword repetition accuracy
Group
Amusics
Controls
Identification accuracy
0.8
0.6
0.4
0.2
0.4
0.6
0.8
Nonword repetition accuracy
Group
Amusics
Controls


d'
Contour
Interval
Scale
Rhythm
Meter
Memory
Group
Amusics
Controls


Identification accuracy
0.8
0.6
0.4
0.2
40
60
80
100
Contour
50
60
70
80
90
100
Interval
Scale
Rhythm
Meter
Memory
Group
Amusics
Controls