IMPAIRED TALKER ROCOGNITION IN MANDARIN-SPEAKING CONGENITAL AMUSICS

Jing Shao[1,2], Lan Wang[2], and Caicai Zhang[1,2]
[1]Department of Chinese and Bilingual Studies, the Hong Kong Polytechnic University
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
jing.shao@polyu.edu.hk, lan.wang@siat.ac.cn,caicai.zhang@polyu.edu.hk

ABSTRACT

The speech signal contains at least two types of information: the linguistic information and a talker's voice. In this study we examined how congenital amusia, a pitch-processing disorder, affects the recognition of talkers' voices. Twenty Mandarin-speaking amusics and 20 controls were tested on talker recognition in four types of contexts that varied in language familiarity: Mandarin real words, Mandarin pseudowords, Arabic words and reversed Mandarin speech. We found that the deficit in amusia affects talker recognition in that amusics demonstrated degraded performance in both native language conditions that contain phonological cues to facilitate talker recognition and non-native conditions where talker recognition primarily relies on phonetics cues including pitch. Altogether, the results suggested that the scope of amusia is beyond the pitch-related processing in linguistic dimension, but also extends to the talker dimension in speech signal.

Keywords: congenital amusia; talker processing; pitch; language familiarity; Mandarin Chinese

1. INTRODUCTION

Amusia is a lifelong neurogenetic disorder of fine-grained pitch processing in music [1], with an estimated prevalence rate of approximately 1.5-4% (Peretz & Vuvan, 2017). The primary deficit in congenital amusia lies in pitch processing [1], [7], [8] and impaired short-term memory for pitch [9], [10]. Moreover, a number of studies have revealed that the deficit in amusia is not domain-specific, but transfers to the language domain. In support of this idea, tonal language speakers with amusia are found to be impoverished in lexical tone perception. For instance, both Mandarin and Cantonese-speaking amusics were less accurate at identifying and discriminating native tones than typical listeners [4], [11], [12]. Furthermore, some studies have suggested that the categorical perception of native lexical tones was also impaired [13]–[15]. The reduced/absent categorical perception of lexical tones suggested that high-level phonological processing of lexical tones may also be impaired in tonal speakers with amusia.

In line with the findings that the high-level phonological processing of lexical tones was impaired, several studies have reported impoverished phonological awareness in amusics [16], [17]. Moreover, a recent study has revealed that amusics are impaired in lexical tone normalization in terms of using the phonological cues in the speech context for adapting to talker variation [18]. Amusics showed reduced context effects in anomalous and meaningful contexts compared with controls, but their performance was largely comparable to controls in nonspeech and reversed speech contexts, indicating that amusics have deficit in making use of phonological cues in the preceding context in the process of tone normalization.

Taken together, converging evidence has shown the impaired phonological processing of lexical tones and reduced phonological awareness in amusia. All these findings are related to speech perception in the linguistic dimension. However, the speech signal contains at least two types of information: the linguistic information and a talker's voice. These two types of information inevitably overlap in the speech signal and interact with each other during speech perception. For instance, speech perception has been found to be less accurate and takes longer response time when the talker variability increases [19]–[21]. On the other hand, recognition of a talker's voice is also influenced by the linguistic content in the speech signal [22], [23]. Typical listeners are more accurate at recognizing talkers' voices in their native language than in an unfamiliar language. For instance, [22] found that native English speakers were most accurate at identifying talkers when speaking English, followed by the Spanish-accented English, and they were worst at recognizing voices speaking Spanish, suggesting the facilitating effect of familiar language in talkers' voice identification.

Another line of research showed that talker identification was also influenced by the listeners' ability in phonological processing [24], [25]. [24] examined the ability of listeners with dyslexia, who are known as having a core phonological deficit, in

recognizing different voices in Mandarin and English. The results showed that English listeners with dyslexia performed equally poorly as the control participants in identifying the Mandarin talkers. However, they were impaired in recognizing different talkers' voices in English compared to controls. These studies suggested that impairment in phonological processing also impeded accurate talker identification.

Talker voice differences are indexed by pitch differences, among other acoustic cues. The literature above has consistently demonstrated the effect of language familiarity and phonological ability on talker recognition. Nonetheless, previous studies on amusia have mostly focused on emotion prosody and linguistic pitch processing, and as such there is a gap as to whether and how the deficit in amusia influences talker processing. As reviewed above, amusics have shown deficits in both general pitch/auditory processing and higher-level phonological processing in the linguistic dimension. It is still not clear how the deficit in amusia influences the perception of talker's voice in different language contents differing in the amount of linguistic cues. The current study examined the performance of Mandarin-speaking amusics on talker recognition in four types of contexts with the available linguistic cues gradually decreased: Mandarin real words, Mandarin pseudo-words, Arabic words and reversed Mandarin speech. Examining the four types of contexts allows us to assess the possible group and linguistic context effects in talker recognition in a comprehensive manner.

2. METHOD

2.1. Participants

Twenty Mandarin-speaking amusics and 20 musically intact controls participated in this experiment. Amusic and control participants were matched one by one in age, gender, and years of education. All participants were native speakers of Mandarin and university students 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 reported any knowledge of Arabic prior to their participation in the current study. None of the listeners had ever lived in Arabic-speaking countries or had any Arabic-speaking friends or family members. Amusics and controls were identified using the Montreal Battery of Evaluation of Amusia (MBEA) [26], which is commonly used to diagnose amusics. 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%. The Demographic characteristics of the participants are summarized in Table 1. The experimental procedures were approved by the Human Subjects Ethics committee of Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. Informed written consent was obtained from participants in compliance with the experiment protocols.

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)

2.2. Stimuli

The stimuli included four types of conditions - Mandarin real words, Mandarin pseudo-words, Arabic real words and reversed Mandarin speech generated from Mandarin real words, all were disyllabic utterances. Each type of stimuli consisted of 20 words. Six female bilingual Mandarin-L1/Arabic-L2 speakers were recorded producing the words in Mandarin (real and pseudo) and Arabic (real). The recordings were made at a sampling rate of 22,050 Hz with 16 bits per sample. Each stimulus type included three repetitions of each word.

For each talker and each stimulus type, one clearly produced token was selected and segmented from the recordings using Praat [27]. The average acoustic intensity of each 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 [27].

2.3. Procedure

E-prime 2.0 was used to present the stimuli and collect the responses. There were two phases in this experiment: learning phase and a following test phase. Among the 20 words in each condition, ten were used in the learning phase. In the *learning* phase, all the participants completed a talker recognition training task in which they learned to identify the voices of six female talkers, each of

whom was presented as a cartoon-like character on a computer screen. The task was a 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 basic trials in each stimulus condition. 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 was given to the participants to familiarize them with the procedure.

Immediately after the learning session, there was a *test* session. Stimuli used in the test session included the ten words in each condition that was trained in the learning session and another ten words that were not trained. The task was the same forced-choice identification task as used in the learning session, but no feedback was given and the participants were required to make the response within 5 seconds. The presentation order of the four conditions was counterbalanced across the participants, and kept identical between matched amusics and controls.

3. RESULTS

The accuracy in the identification task was calculated and analysed as follows. Response to each trial was coded as 1 or 0 (correct or incorrect) for each participant. To compare the accuracy of amusics and controls, 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 Mandarin words) 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 fixed effects in the models. Models were compared by likelihood ratio tests and p-values were obtained from those tests. The above analyses were performed with R (R Core Team, 2014), using the *lme4* package [29], the *lmtest* package [30] and the *lsmeans* package [31]. Figure 1 shows the accuracy.

Figure 1: Talker recognition accuracy in the two groups.

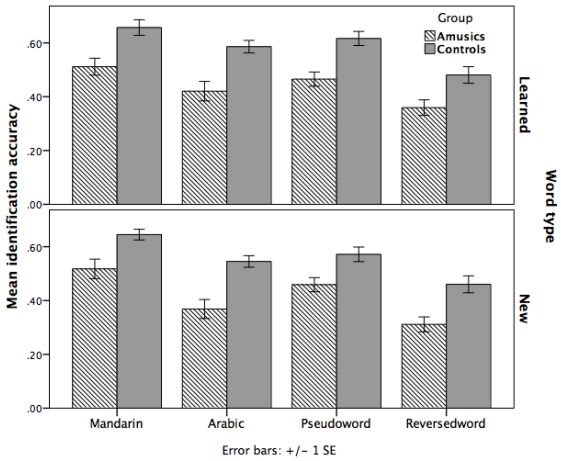
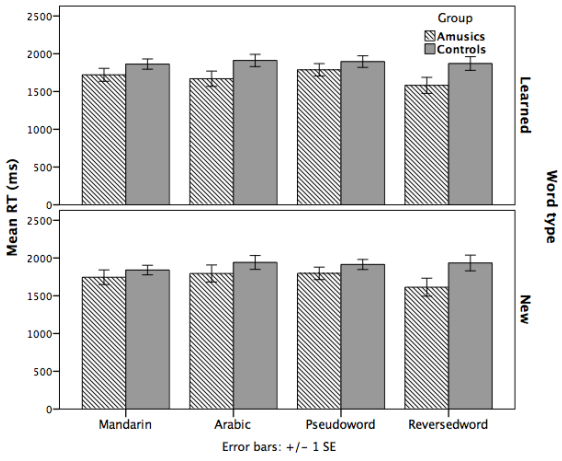


Figure 2: RT in the talker recognition task.



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 (*p*s < 0.001). Mandarin pseudoword context also elicited better talker identification accuracy than the other two conditions, Arabic word ($z = -4.840$, $p < 0.001$) and reversed speech ($z = -12.564$, $p < 0.001$). Finally, the Arabic word context elicited significantly higher accuracy than the

reversed condition (z = -7.777, $p < 0.001$). No other effects were significant.

For the RT analysis, we included both the correct and incorrect trials. It is hypothesized that RT may also reflect the cognitive efforts involved in talker identification, and can be interpreted along with the accuracy data. Linear mixed-effects models were also fitted on the log-transformed RT data 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 fixed effects in the models. Models were compared by likelihood ratio tests. Figure 2 shows the 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 reversed 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.

4. DISCUSSION

The current study examined the amusics' ability to recognize a talker's voice in four conditions differing in the amount of available phonological cues. Firstly, the results confirmed the language effect in talker identification in that control 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., containing native phonemes and tones), but in the pseudoword condition, the combinations of the syllables and tones were illegal, which reduced the semantic content of the words. On the other hand, the Arabic word condition contains no native phonemes to Mandarin speakers. The reversed speech context sounded like foreign language, where the legal phonological cues were removed. The results obtained in the current study are largely consistent with previous findings that talker identification is facilitated by the native language context [22], and this effect disappears when the linguistic content of the speech is eliminated or destroyed [23].

Compared to controls, amusics demonstrated degraded talker recognition performance in terms of the accuracy, and there was no group and condition interaction. These results suggest that Mandarin-speaking amusics were impaired in talker recognition in both levels: talker recognition via utilizing phonological cues in native speech contexts (Mandarin real word and pseudoword) and talker recognition via analysing phonetic cues in the non-native speech contexts (Arabic word and reversed Mandarin speech). Therefore, corroborating with previous findings [18], in which amusics exhibited impairment in lexical tone normalization in utilizing phonological cues in the meaningful and anomalous contexts, results of the current study suggested that when processing another dimension of speech signal, i.e., talker' voice, amusics were also deficient in using the phonological representations in the language to recognize the talker' identity.

It should be noted that amusics, despite displaying overall degraded performance in talker processing, still exhibited better performance in the native speech contexts (Mandarin real word and pseudoword) than in the non-native speech contexts (Arabic word and reversed Mandarin speech), a pattern largely similar to the performance of controls. This result indicates that amusics were able to make use of phonological cues to some extent, displaying certain facilitation of phonological cues in talker recognition. Taken together, the findings of the current study suggested that disorders in amusia does not only affect the linguistic pitch processing, but also influences the talker's voice recognition, another important aspect of speech signal.

5. CONCLUSION

The findings of the current study demonstrated that amusia affects talker identification. The degraded performance in amusia was found in both native language conditions and non-native conditions. Despite being impaired in talker recognition compared to controls, amusics displayed similar patterns concerning the language familiarity effect as controls, suggesting that they seemed to preserve some abilities in utilizing the phonological cues in the native language conditions. The results in the present study have further expanded our understanding of the scope of amusia in that the deficit in amusia also negatively influences talker' voice recognition.

6. ACKNOWLEDGEMENT

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), and the PolyU Start-up Fund for New Recruits.

7. REFERENCES

[1] 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, 185–191.

[2] Peretz, I., Gosselin, N., Tillmann, B., Cuddy, L., Gagnon, B., Trimmer, C.; Paquette, S., Bouchard, B. 2008. On-line identification of congenital amusia. *Music Percept.* 25, 331–343.

[3] 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, e33424.

[4] Nan, Y., Sun,Y., Peretz, I. 2010. Congenital amusia in speakers of a tone language: Association with lexical tone agnosia. *Brain.*133, 2635–2642.

[5] Peretz, I., Vuvan, D. T. 2017. Prevalence of congenital amusia. *Eur. J. Hum. Genet.* 25, 625.

[6] Pfeifer, J., Hamann, S. 2015. Revising the diagnosis of congenital amusia with the Montreal Battery of Evaluation of Amusia. *Front. Hum. Neurosci*. 9, 161.

[7] Foxton, J. M., Jennifer L., Rosemary, G., Peretz, I., and Griffths, D. 2004. Characterization of deficits in pitch perception underlying 'tone deafness. *Brain*. 127, 801–810.

[8] Hyde, K. L., Peretz, I. 2004. Brains that are out of tune but in time. *Psychol. Sci.* 15, 356–360.

[9] Tillmann, B., Schulze, K., Foxton, J. M. 2009. Congenital amusia: A short-term memory deficit for non-verbal, but not verbal sounds. *Brain Cogn.* 71, 259–264.

[10] Tillmann, B., Lévêque, Y., Fornoni, L., Albouy, P., Caclin, A. 2016. Impaired short-term memory for pitch in congenital amusia. *Brain Res*.1640, 251–263.

[11] 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. *J. Acoust. Soc. Am.* 140, 563–575.

[12] 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 *Interspeech*. 272-276.

[13] 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. *Mem. Cognit.* 40, 1109–21.

[14] Huang, W. T., Liu, C., Dong, Q., Nan, Y. 2015. Categorical perception of lexical tones in Mandarin-speaking congenital amusics. *Front. Psychol*. 6, 829.

[15] 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, e0183151.

[16] Jones, J. L., Lucker, J., Zalewski, C., Brewer, C,. Drayna, D. 2009. Phonological processing in adults with deficits in musical pitch recognition. *J. Commun. Disord.*42, 226–234.

[17] Sun, Y., Lu, X., Ho, H. T., Thompson, W. F. 2017. Pitch discrimination associated with phonological awareness: Evidence from congenital amusia. *Sci. Rep.* 7, 44285.

[18] Zhang, C., Shao, J., Chen, S. 2018. Impaired perceptual normalization of lexical tones in Cantonese-speaking congenital amusics. J. Acoust. Soc. Am. 144, 634-647.

[19] Mullennix, J. W., Pisoni, D. B. 1990. Stimulus variability and processing dependencies in speech perception. *Percept. Psychophys*. 47, 379–390.

[20] Nusbaum, H. C., Morin, T. M. 1992. Paying attention to differences among talkers. in *Speech Perception, Speech Production, and Linguistic Structure*. Tokyo. 113–134.

[21] Strange, W., Verbrugge, R. R., Shankweiler, D. P., Edman, T. R. 1976. Consonant environment specifies vowel identity. *J. Acoust. Soc. Am*. 60, 213–224.

[22] Thompson, C. P. 1987. A language effect in voice identification. *Appl. Cogn. Psychol*. 2, 121–131.

[23] Goggin, J. P., Strube, G., Simental, L. R. 1991. The role of language familiarity in voice identification. *Mem. Cognit*. 19, 448–458.

[24] Perrachione, T. K., Del Tufo, S. N., Gabrieli, J. D. E. 2011. Human voice recognition depends on language ability. *Science*, 333, 595.

[25] 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. *Proc. 17th Int. Congr. Phonetic Sci. (ICPhS 2011),* 1578–1581.

[26] Peretz, I., Champod, A. S., Hyde, K. 2003 Varieties of musical disorders. *Ann. N. Y. Acad. Sci.* 999, 58–75, 2003.

[27] Boersma, P., Weenink, D. Praat: Doing phonetics by computer. 2014.