Impaired perceptual normalization of lexical tones in Cantonese-speaking congenital amusics

Caicai Zhang, Jing Shao, and Si Chen

Citation: The Journal of the Acoustical Society of America **144**, 634 (2018); doi: 10.1121/1.5049147
View online: https://doi.org/10.1121/1.5049147
View Table of Contents: https://asa.scitation.org/toc/jas/144/2
Published by the Acoustical Society of America

Impaired perceptual normalization of lexical tones in Cantonese-speaking congenital amusics

Caicai Zhang,[a),b)] Jing Shao,[a)] and Si Chen
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Hong Kong SAR, China

(Received 29 January 2018; revised 17 July 2018; accepted 18 July 2018; published online 8 August 2018)

Human listeners perceive speech sounds relative to acoustic cues in context. In this study the authors examined how congenital amusia, a pitch-processing disorder, affects perceptual normalization of lexical tones according to the distribution of *F*0 cues in context. Sixteen Cantonese-speaking amusics and 16 controls were tested on the effects of shifting *F*0 level in four types of contexts on tone perception: nonspeech, reversed speech, semantically anomalous speech, and meaningful speech contexts. Performance of controls replicated previous studies, showing contrastive changes of tone perception according to the shifted *F*0 level of anomalous and meaningful contexts, which were native speech contexts with phonological cues to estimate a talker's tone space. Effects of nonspeech and reversed contexts were small and inconsistent, and tone perception performance varied depending on the typicality of a talker's *F*0 range. In contrast to controls, amusics showed reduced context effects in anomalous and meaningful contexts, but largely comparable context effects in nonspeech and reversed contexts, indicating a deficit of amusics in tone normalization through phonological cues in native speech contexts. These findings suggest that the ability to perceive speech sounds relative to acoustic cues in context is not a universal endowment, and that this ability is impaired substantially in amusics. © *2018 Acoustical Society of America*.
https://doi.org/10.1121/1.5049147

[TCB] Pages: 634–647

I. INTRODUCTION

There is a tremendous amount of talker variation in speech signals. Talker variation results in acoustic dispersion within the same phonological category as well as overlap between neighboring categories, leading to variance in the mapping of acoustic signals to perceptual representations of speech sounds (Johnson, 2005; Liberman *et al.*, 1967). This phenomenon is known as the "lack of invariance" problem in speech perception research (Johnson, 2005). The process by which listeners map speech signals with talker variation to representations of phonological categories is referred to as talker normalization (Johnson, 2005). Note that the term talker normalization is used in this paper in a neutral sense without theoretical presumptions of abstracting or filtering-out talker details in the mapping process (cf. Gerstman, 1968; Syrdal and Gopal, 1986).

Several mechanisms have been proposed to account for the solution of the lack of invariance problem. The intrinsic normalization mechanism claims that talker variation can be reduced by rescaling/transforming the values of acoustic cues of a target speech sound (e.g., formant frequencies) against other intrinsic acoustic cues that reflect a talker's voice characteristics (e.g., *F*0 and voice quality) carried within the same target sound (Syrdal and Gopal, 1986). On the other hand, the extrinsic normalization mechanism emphasizes the importance of extrinsic cues, e.g., a speech context (Gerstman, 1968; Ladefoged and Broadbent, 1957). According to this mechanism, listeners perceive a target speech sound according to the distribution of acoustic cues in the surrounding context.

A. Perceptual normalization of lexical tones

For lexical tone, talker variation has been demonstrated to pose a challenge for accurate mapping of acoustic *F*0 signals to phonological representations of tones (Peng *et al.*, 2012; Wong and Diehl, 2003). For tone normalization, it has been found that it is important for listeners to adapt to a specific talker's *F*0 range[1] (Francis *et al.*, 2006; Peng *et al.*, 2012; Wong and Diehl, 2003; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013). Although the absolute *F*0 values of a tone exhibit great variation in the productions of different talkers, a high tone tends to be located near the upper end of a talker's *F*0 range, while a low tone tends to be located near the lower end. Perceiving the *F*0 in acoustic signals according to the *F*0 range of a talker, instead of the absolute *F*0 values, therefore provides a means to achieve accurate signal-to-representation mapping.

Previous studies have revealed that although the intrinsic mechanism may play some role in relative pitch perception (Honorof and Whalen, 2005), a talker's intrinsic voice cues (e.g., *F*0 and voice quality) generally have a limited effect on the estimation of an unfamiliar talker's *F*0 range (Bishop and Keating, 2012). When asked to rate the pitch level of a brief isolated *F*0 sample within an unknown

[a)]Also at: Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.
[b)]Electronic mail: caicai.zhang@polyu.edu.hk

voice's *F*0 range, English listeners most strongly relied on the absolute *F*0 in their ratings (Bishop and Keating, 2012). Further attesting to the limited effect of the intrinsic mechanism, Cantonese listeners were found to show poor tone identification accuracy when asked to identify isolated spoken words carrying tones produced by multiple talkers (Peng *et al.*, 2012; Zhang and Chen, 2016; Zhang *et al.*, 2012). Instead of adapting to a specific talker's *F*0 range, Cantonese listeners demonstrated a strong bias of talker typicality, which was defined as the distance of a talker's *F*0 range relative to the population mean *F*0 range. Specifically, tones produced by a talker whose *F*0 range was close to the population mean *F*0 range were more frequently correctly identified, whereas the same tones produced by talkers whose *F*0 range was either *higher* or *lower* than the population mean *F*0 range were perceptually biased toward a neighboring tone with *higher* or *lower* *F*0 height, respectively (Zhang and Chen, 2016). Taken together, these results suggest that listeners may not fully adapt to an unfamiliar talker's *F*0 range by solely relying on intrinsic cues.

Instead, a speech context with cues to a talker's *F*0 range is found to be crucial for talker adaptation (Chen and Peng, 2016; Francis *et al.*, 2006; Huang and Holt, 2009, 2011; Moore and Jongman, 1997; Wong and Diehl, 2003; Zhang, 2018; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013). The effect of acoustic cues in surrounding contexts has been widely demonstrated on the perception of segmentals and suprasegmentals (Holt and Kluender, 2000; Ladefoged and Broadbent, 1957; Sjerps *et al.*, 2011). With regard to tone perception, Zhang and Chen (2016) found that Cantonese listeners were able to correctly identify tones produced by multiple talkers when those words were presented within a speech context from the same talker, and the influence of talker typicality disappeared. Further supporting the effect of context, a plethora of studies have found a contrastive effect of shifting the *F*0 level of a speech context on tone perception (Chen and Peng, 2016; Francis *et al.*, 2006; Huang and Holt, 2009, 2011; Moore and Jongman, 1997; Wong and Diehl, 2003; Zhang, 2018; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013). This contrastive context effect suggests that listeners probably adapt to a talker's tone space through *F*0 cues contained in the speech context, and dynamically evaluate the relative *F*0 location of a target tone within that tone space (Sjerps *et al.*, 2011; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013). When a talker's tone space is shifted upwards to a region with higher *F*0, implying that this talker is speaking with an overall higher pitch, the *F*0 carried by the target word is judged to be a lower tone, and vice versa.

While the effect of shifting *F*0 level in a speech context on tone normalization is widely reported, it remains unclear what cues contained in the context are most crucial for tone normalization. A related question that has long been under debate is whether the context effect is mediated by general auditory processes, such that a nonspeech context with identical acoustic cues as the speech context would elicit similar effects on the perception of a target speech sound (Fowler, 2006; Holt and Kluender, 2000; Laing *et al.*, 2012).

In an attempt to sort out the contribution of different cues to tone normalization and to shed some light on the debate regarding nonspeech contexts, Zhang and Chen (2016) compared the effects of four types of contexts with incrementally more cues on the normalization of Cantonese level tones. The four types of contexts were: nonspeech context (auditory cues), reversed speech context (auditory + phonetic cues), semantically anomalous speech context (auditory + phonetic + phonological cues), and meaningful speech context (auditory + phonetic + phonological + semantic + syntactic cues). While carrying largely identical *F*0 cues, these four types of contexts varied in their proximity to being native, meaningful speech to Cantonese listeners, which could influence how Cantonese listeners used those *F*0 cues contained in the contexts for tone normalization. The nonspeech context only contained auditory pitch cues with no linguistic information, whereas the reversed speech context was a time-reversed version of a meaningful speech context (呢個字係 /li55[2] ko33 tsi22 hɐi22/ "This word is") and sounded like foreign speech (e.g., /ia his (n)o kin/). On the other hand, the anomalous speech context (呢錯視幣 /li55 ts^{h}o33 si22 pɐi22/ "This mistake sees money") and the meaningful speech context were native speech contexts with phonological cues to estimate the *F*0 location of a talker's tone space. Phonological cues refer to cues that can distinguish speech sounds in a listener's native language. In particular, anomalous and meaningful contexts contained syllables with the high level tone (/li55/) and the low level tone (/si22/ and /pɐi22/), which may allow listeners to map out a talker's tone space. The results showed that shifting the *F*0 trajectory upwards and downwards in the nonspeech context hardly had any effect on the normalization of Cantonese level tones, a finding consistent with several previous studies (Chen and Peng, 2016; Francis *et al.*, 2006; Zhang *et al.*, 2012, 2013 but see Huang and Holt, 2009, 2011 for different results on Mandarin tone normalization). The reversed speech context elicited a stronger effect than the nonspeech context, but also failed to affect tone perception contrastively in several conditions. On the other hand, both anomalous and meaningful speech contexts, which contained native phonological cues to estimate a talker's tone space (e.g., mapping out the *F*0 values of the high and low tone in a talker's voice), showed consistent effects on shifting the perception of Cantonese tones contrastively in all conditions. These findings suggest that phonological cues in native speech contexts may be most critical for tone normalization.

B. Congenital amusia and deficits of lexical tone perception

Although many studies have consistently demonstrated the effect of manipulating acoustic cues in context on speech perception (i.e., phonetic context effect), it remains unclear whether the process of perceiving a speech sound relative to the distribution of acoustic cues in context is a universal human endowment, and whether this process may be disrupted in individuals with speech or hearing disorders. As a first attempt to understand the potential impact of disorders

on the phonetic context effect, the current study aims to examine the performance of a group of adult listeners with congenital amusia (amusia hereafter) on tone normalization.

Amusia is a lifelong neurogenetic disorder of fine-grained pitch processing in music (Peretz *et al.*, 2002). Individuals with amusia often have difficulty detecting mistuned melodies or memorizing familiar tunes, with an estimated prevalence rate of approximately 1.5%–4% (Nan *et al.*, 2010; Peretz *et al.*, 2008; Peretz and Vuvan, 2017; Pfeifer and Hamann, 2015; Wong *et al.*, 2012). Although several earlier studies suggested that amusia is a domain-specific pitch deficit (Ayotte *et al.*, 2002; Peretz *et al.*, 2002), recent studies revealed that pitch processing in speech is also impaired in amusics, especially when the pitch differences are small (Jiang *et al.*, 2010; Liu *et al.*, 2012a; Patel *et al.*, 2008; Vuvan *et al.*, 2015). For instance, Liu *et al.* (2010) used naturally produced speech stimuli with small pitch glide differences on the final word that indicated whether the utterance was a statement or question, and found that English-speaking amusics were impaired in statement-question discrimination.

Further corroborating the influence of amusia on linguistic pitch processing, studies on lexical tones have consistently reported that tonal language speakers with amusia exhibited impaired lexical tone perception (Huang *et al.*, 2015; Jiang *et al.*, Yang, 2012; Liu *et al.*, 2015b; Nan *et al.*, 2010; Shao *et al.*, 2016; Wang and Peng, 2014; Zhang and Shao, 2018; Zhang *et al.*, 2017a). Liu *et al.* (2015b) found that Cantonese-speaking amusics were less accurate than musically intact controls in identifying Cantonese spoken words that contrasted in tones. Shao *et al.* (2016) replicated the previous finding of reduced accuracy of tone identification in Cantonese-speaking amusics, and furthermore, found less accurate and slower discrimination of Cantonese tones, although amusics' impairment in tone discrimination was less severe compared to that in tone identification. Nan *et al.* (2010) found that Mandarin-speaking amusics performed generally worse than controls in the identification and discrimination of Mandarin tones, but there was individual variation in the severity of impairment. Taken together, there is converging evidence for the negative impact of amusia on tone perception, which thus provides an ideal case for examining the influence of amusia on the phonetic context effect on tone normalization, a question that has not been examined before.

C. The current study

The motivations for conducting the current study are two-fold.

First, although many previous studies have reported poor performance in tone perception in amusics at the level of a single word, few studies have examined how amusia actually affects tone normalization through the distribution of *F*0 cues in the surrounding context. As mentioned above, plenty of studies have shown that tone perception not only depends on *F*0 cues carried by a word itself, but also involves the normalization process via the distribution of *F*0 cues in the speech context to accommodate talker variation in *F*0 range (Chen and Peng, 2016; Francis *et al.*, 2006; Huang and Holt, 2009, 2011; Moore and Jongman, 1997; Peng *et al.*, 2012; Wong and Diehl, 2003; Zhang *et al.*, 2013, 2016, 2012; Zhang and Chen, 2016). However, little is known about how talker variation is accommodated in tone perception in amusics. In order to better understand the nature and scope of the deficits of amusia in linguistic pitch processing, it is thus important to examine how the phonetic context effect on tone normalization is affected in the amusic population. Since previous studies have reported impairment of amusics in pitch processing over broad temporal units, including sentence intonation processing and musical melodic processing (Jiang *et al.*, 2010; Liu *et al.*, 2010; Patel *et al.*, 2008; Peretz *et al.*, 2002; Peretz and Vuvan, 2017; Vuvan *et al.*, 2015), there is reason to speculate that tonal language speakers with amusia may be impaired in lexical tone perception beyond the scale of a single word. We hypothesize that tonal language speakers with amusia are impaired in tone normalization according to the distribution of *F*0 cues in the surrounding context, showing null or reduced phonetic context effects compared to controls.

Second, studying the impact of disorders such as amusia on the phonetic context effect may shed additional light on the underlying mechanisms of talker normalization. As mentioned above, several studies reported that native speech contexts with phonological cues to adapt to a talker's tone space elicited the strongest effect on tone normalization (Chen and Peng, 2016; Francis *et al.*, 2006; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013). Nonetheless, it is unclear whether those subjects recruited in previous studies were all musically intact listeners. If some listeners were amusical, it could influence their performance on tone normalization. It is thus important to carefully assess the pitch abilities of the listeners, and compare the performance of amusics vs controls on the phonetic context effect, which is the aim of the current study. This approach offers an opportunity to re-assess the underlying mechanisms of tone normalization, and more importantly, allows us to examine whether amusic listeners use contextual cues in a similar manner as control listeners.

In the current study, we compared the performance of Cantonese-speaking amusics and musically intact controls on the normalization of Cantonese level tones in four types of contexts with incrementally more cues: nonspeech context, reversed speech context, anomalous speech context, and meaningful speech context. Level tones are ideal for examining the effect of shifting the *F*0 level on tone normalization, because the three level tones in Cantonese, namely the high level tone (e.g., 醫 /ji55/ "doctor"), the mid-level tone (e.g., 意 /ji33/ "meaning"), and the low level tone (e.g., 二 /ji22/ "two"), contrast a largely similar and flat pitch trajectory at different pitch heights (Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013). Examining tone normalization in four types of contexts provides a comprehensive assessment of the performance of amusics in a full spectrum, by revealing how amusics may use *F*0 cues differently from controls in each type of context. For the controls, we hypothesize that their performance would largely replicate the previous results, showing the strongest and most consistent effect on tone normalization in anomalous and meaningful speech

contexts. As for amusics, there is increasing evidence that their impairment in tone perception is not purely due to a domain-general pitch-processing deficit, but that higher-level phonological processing related to the access of phonological representations of tones is probably impaired (Huang *et al.*, 2015; Jiang *et al.*, 2012; Zhang *et al.*, 2017b). Given the importance of phonological cues for tone normalization, we hypothesize that amusics would be most strongly impaired in tone normalization in native speech contexts with phonological cues, exhibiting the most pronounced reduction of context effects in anomalous and meaningful speech contexts.

II. METHOD

A. Participants

Sixteen Cantonese-speaking amusics and 16 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 Hong Kong Cantonese, and university students in Hong Kong 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). Amusic and control participants were identified using the Online Identification Test of Congenital Amusia (Peretz *et al.*, 2008), which has been used as a screening tool for amusia in previous studies (Liu *et al.*, 2010; Wang *et al.*, 2017; Wang and Peng, 2014; Wong *et al.*, 2012; Zhang *et al.*, 2017a). All participants took this online test in the lab under the instruction of an experimenter. Amusics scored 71% or lower (Nan *et al.*, 2010), whereas controls scored 80% or higher in the global score of this test, which is the mean accuracy of all three subtests (out-of-key, offbeat, and mistuned). Furthermore, all amusics scored 71% or lower in at least one of the two pitch-based subtests (out-of-key and mistuned), or in both. Independent-samples *t*-tests confirmed that amusics' global scores were significantly lower than the controls [$t(30) = -10.768$, $p < 0.001$]. Amusics also performed significantly worse than controls in each subtest (p's < 0.001), though the group difference was noticeably smaller in the offbeat subtest (see Table I), compatible with the notion that amusia is primarily a pitch processing disorder (Peretz *et al.*, 2002; Vuvan *et al.*, 2015). Demographic characteristics of the participants are reported in Table I. The experimental procedures were approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University. Informed written consent was obtained from each participant in compliance with the experiment protocols.

TABLE I. Demographic characteristics of 16 amusic participants and 16 control participants. Amusic and control participants were determined according to the global score from the Online Identification Test of Congenital Amusia (http://www.brams.umontreal.ca/onlinetest) (Peretz *et al.*, 2008). M = male, F = female.

	Amusics	Controls
No. of participants	16 (7 M, 9 F)	16 (7 M, 9 F)
Age (range)	22.76 ± 4.4 years (18.7–32.7 years)	22.72 ± 3.9 years (18.5–32.9 years)
Test of Congenital Amusia		
Out-of-key (SD)	61.1 (12.5)	92.0 (6.6)
Offbeat (SD)	68.6 (11.6)	84.9 (8.5)
Mistuned (SD)	57.7 (10.8)	87.0 (10.8)
Global score (SD)	62.3 (7.9)	88.2 (5.4)

B. Stimuli

The stimuli were identical to those used in Zhang and Chen (2016). The stimuli included four types of contexts—nonspeech, reversed speech, anomalous speech, and meaningful speech contexts. Four native Cantonese speakers with different *F*0 ranges (two female speakers *F*01 and *F*02 and two male speakers *M*01 and *M*02) were recorded reading aloud a meaningful sentence, 呢個字係意 /li55 ko33 tsi22 hɐi22 ji33/ ("This word means"), and an anomalous sentence, 呢錯視幣意 /li55 ts^{h}o33 si22 pɐi22 ji33/ ("This mistake sees money meaning"). These two sentences were produced six times by each speaker. In both sentences, the final word /ji33/ with the mid-level tone was the target word, and the four-syllable utterance before the target served as the context. The meaningful speech context was semantically neutral, carrying no semantic prediction of the target word. The anomalous speech context was a meaningless combination of real syllables whose rhymes and tones were matched with those in the meaningful context. The recordings were made at a sampling rate of 22 050 Hz with 16 bits per sample. The four talkers were instructed to read the two sentences with a brief pause between the context and the target word as naturally as possible, in order to ease the segmentation of the context and target word in the subsequent manipulation.

For each talker, one clearly produced meaningful and anomalous speech context matched in mean, minimum, and maximum *F*0 was selected and segmented from the recordings using Praat (Boersma and Weenink, 2014). The maximum *F*0 generally occurred during the first syllable (/li55/), and the minimum *F*0 generally occurred during the last syllable (/hɐi22/ or /pɐi22/) of the context. The average acoustic intensity of all selected speech contexts was normalized to 55 dB using Praat. The overall *F*0 trajectory of each context was then lowered by 3 semitones, kept unshifted, or raised by 3 semitones using the overlap-add re-synthesis in Praat. When the *F*0 trajectory of a context was lowered or raised, it gave the impression that a talker produced the same context with an overall lower or higher pitch. The magnitude of the *F*0 shift in the context largely reflected the *F*0 distance among the three level tones (Francis *et al.*, 2006; Wong and Diehl, 2003), and has been confirmed to elicit contrastive changes of tone perception among them in previous studies (Zhang and Chen, 2016; Zhang *et al.*, 2013). The segmental information (e.g., consonants and vowels) remained unchanged after the *F*0 manipulation. The mean, minimum, and maximum *F*0 of the meaningful and anomalous speech contexts from the four talkers are reported in Table II.

Reversed speech and nonspeech contexts were created from the meaningful speech contexts. Reversed speech

TABLE II. *F*0 information of the meaningful and anomalous speech contexts and the target words produced by the four talkers. The meaningful and anomalous contexts were matched in mean, minimum, and maximum *F*0. The target words were identical in all four types of contexts.

		F0 shift			Target
Talker	Context	*Lowered*: Mean *F*0 (min, max)	*Unshifted*: Mean *F*0 (min, max)	*Raised*: Mean *F*0 (min, max)	*F*0 (SD)
F01	Meaningful	198.2 (151.9, 275.9)	236.8 (190.2, 314.2)	280.5 (233.7, 357.7)	234.0 (4.9)
	Anomalous	197.9 (144.6, 281.6)	236.4 (183.0, 320.2)	280.1 (227.1, 363.8)	
F02	Meaningful	173.9 (116.0, 265.6)	208.1 (148.7, 300.1)	246.3 (193.4, 337.8)	206.8 (3.0)
	Anomalous	174.9 (124.9, 258.6)	210.1 (159.2, 292.7)	247.3 (197.7, 331.1)	
M01	Meaningful	124.5 (84.8, 184.9)	148.4 (108.1, 208.7)	174.6 (134.7, 234.6)	143.6 (5.6)
	Anomalous	123.7 (83.9, 187.3)	147.5 (107.5, 210.2)	173.5 (133.9, 236.5)	
M02	Meaningful	96.8 (75.0, 137.7)	113.8 (88.8, 157.1)	134.5 (110.0, 177.5)	114.8 (1.7)
	Anomalous	105.2 (76.8, 139.9)	117.8 (90.7, 159.1)	137.0 (111.4, 179.6)	

contexts were generated by time-reversing the meaningful speech contexts with lowered, unshifted, and raised *F*0 for each talker using Praat. The *F*0 contour and intensity profile extracted from the meaningful speech contexts were used to generate the nonspeech contexts. Nonspeech contexts were generated using a triangle wave, which contained only odd-numbered harmonics; as a result the harmonic structure of nonspeech contexts was different from that of natural speech sounds. Following previous studies (Zhang and Chen, 2016; Zhang *et al.*, 2013), the average acoustic intensity of the nonspeech contexts was manipulated to be 75 dB, 20 dB higher than that of the speech contexts, for the reason that the nonspeech contexts sounded softer.

One clearly produced token of the target word was selected for each talker. Each target word was normalized in duration to 450 ms using Praat. Its average acoustic intensity was normalized to 55 dB, equal to the acoustic intensity of speech contexts. The target word was then appended to the end of each context type after a jittered interval of 300–500 ms for each talker. The mean *F*0 of the target word produced by each talker is shown in Table II.

C. Procedure

E-prime 1.1 was used to present the stimuli and collect the responses. The presentation of stimuli was blocked by the context type, with each block containing only trials of one type of context. During a trial, an auditory stimulus containing a preceding context and a target word was presented to participants via headphones. Participants were instructed to pay attention to the whole sound sequence, and to identify the terminal target word as any of the following three Cantonese words—醫 (/ji55/ doctor), 意 (/ji33/ meaning), and 二 (/ji22/ two). These three words were minimally contrastive in the three level tones and are not distinguished by any other cues (e.g., formant frequencies). The subjects were given 2 s to respond after the end of the whole sound sequence, and were instructed to identify the target word by pressing labelled buttons on a computer keyboard as fast and as accurately as possible. Responses made beyond the 2 s time limit were disregarded from analysis. The next trial started 1 s after the end of the previous one. In total there were four blocks, each corresponding to one context type. Within a block, 12 trials (3 *F*0 shifts × 4 talkers) were presented randomly in a sub-block; each sub-block was repeated 6 times, meaning that each stimulus was presented 6 times. There was a break between two blocks.

The presentation order of the four context blocks was counterbalanced among the participants within each group as much as possible and kept identical between each pair of matched amusic and control participants. Before the experiment, one practice block with eight trials containing stimuli from two extra talkers not used in the current experiment was presented to the participants to familiarize them with the experimental procedure. The practice block contained stimuli of the same context type (e.g., nonspeech context) as those that occurred in the first experimental block for each participant. All stimuli were presented to the participants at a comfortable listening level, which was determined during the practice block and kept constant throughout the experiment.

III. RESULTS

A. Results of the expected response rates

The data were first analyzed in terms of the expected response rates, which were defined according to the contrastive context effect mentioned above. Specifically, the target word was expected to be identified as having the high level tone (/ji55/ doctor) in the lowered *F*0 condition, as having the mid-level tone (/ji33/ meaning) in the unshifted *F*0 condition, and as having the low level tone (/ji22/ two) in the raised *F*0 condition. The same identification patterns were expected for stimuli from all four talkers as an indication of perceptual accommodation of talker variation in *F*0 range. The expected response rates were calculated for each context type per talker and *F*0 shift condition, as an index of the magnitude of effect of each type of context. For each trial, the response was coded "1" if it was the expected response, and "0" if not. Trials with no responses received within the 2 s limit were excluded from analysis. The average rate of null responses was 3.68% for the amusic group and 3.92% for the control group. All the analyses were performed with *R*, using the *lme4*, *lmtest*, *lsmeans*, and *glm* packages (R Core Team, 2014). Figure 1 shows the expected response rates for each talker, *F*0 shift, and context condition for the two groups.

Three sets of analyses were conducted on the expected response rates. First, in order to compare the performance of

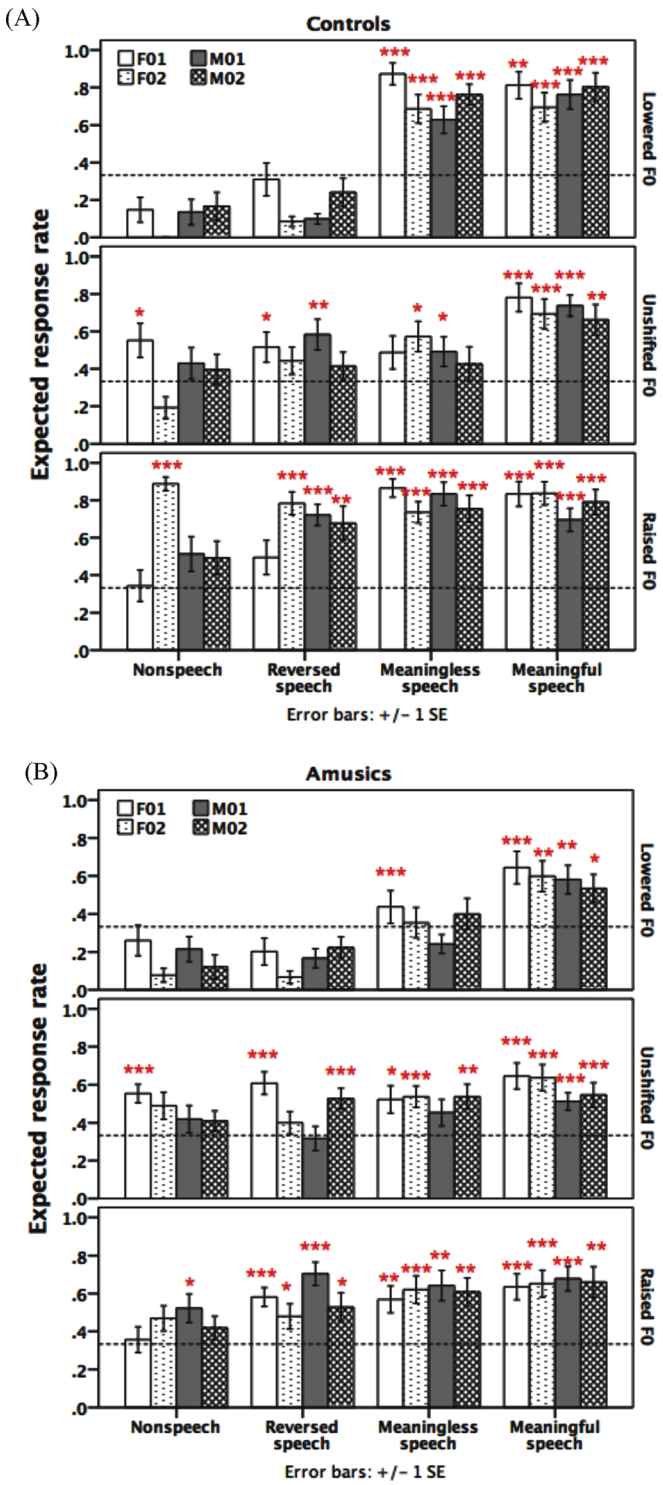

FIG. 1. (Color online) Results of the expected response rates for each talker, *F*0 shift and context condition for each group. (A) The control group. (B) The amusic group. Dotted lines indicate chance performance (0.33). Asterisks refer to the significance level of the generalized mixed-effects models comparing the expected response rates with chance performance. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

amusics and controls in the four types of contexts, generalized mixed-effects models were fitted on the responses to each trial (1 or 0) with *group* (amusics and controls) and *context type* (nonspeech, reversed speech, anomalous speech, and meaningful speech) as two fixed effects, and subjects as a random effect. Two-way interaction was also included as a fixed effect in the models. To test the significance of fixed effects, we started from a simple model (m_0) with only the intercept as a factor, and added the factors *group* and *context type* consecutively. A model with a fixed effect was compared with a baseline model without it by likelihood ratio tests, and *p*-values were obtained from those tests. Details of the models are reported in Table I of the supplemental material.[3]

There were significant main effects of *group* [$\chi^2(1) = 8.5583$, $p = 0.003$] and *context type* [$\chi^2(1) = 634.9$, $p < 0.001$], and a significant two-way interaction [$\chi^2(1) = 58.167$, $p < 0.001$]. Pairwise comparisons were then conducted to examine the group difference within each context type and the context difference within each participant group. The amusic group exhibited significantly lower accuracy than the control group in the anomalous ($z = -4.997$, $p < 0.001$) and meaningful speech contexts ($z = -4.621$, $p < 0.001$), whereas their accuracy was comparable to the controls in the nonspeech ($z = -0.083$, $p = 0.934$) and reversed speech contexts ($z = -41.161$, $p = 0.246$). Within the control group, the differences in expected response rates between the four types of contexts were all significant (p's < 0.001). In particular, the expected response rate was highest in the meaningful speech context, followed by the anomalous speech context, reversed speech context, and nonspeech context (meaningful > anomalous > reversed > nonspeech). Within the amusic group, the results were largely similar. The differences between the four types of contexts were all significant (p's < 0.001), except for that between the nonspeech and reversed speech context ($p = 0.145$) (meaningful > anomalous > reversed = nonspeech).

Second, in order to examine whether the distribution of *F*0 cues in a context had any effect on tone normalization for each participant group, we tested whether the expected response rates were significantly higher than chance performance (0.33 given three allowed options) in each condition. Generalized mixed-effects models were constructed to compare the expected response rates against chance performance for each talker, *F*0 shift, and context condition within each participant group. The responses to each trial (1 or 0) within each condition were the input, and subjects were included as a random effect. We first fitted a generalized linear mixed-effects model with only the intercept as the fixed effect and subjects as the random effect, and compared the intercept to the log odds value corresponding to 33% accuracy [i.e., log (1/2)]. The Wald test statistic was then calculated, which was the proportion of the coefficient estimate and its standard error, and *p*-values were obtained from the test. If a certain type of context facilitated talker normalization, the expected response rates were expected to be significantly higher than chance consistently in all talker and *F*0 shift conditions, irrespective of talker variation and *F*0 manipulations. Detailed results are reported in Table II of the supplemental material.[3]

For the control group, in the nonspeech context condition, the expected response rates were significantly higher than chance only in two conditions: female talker *F*01 in the unshifted *F*0 condition [$M = 0.55$, Standard deviation (SD) $= 0.364$, $z = 1.954$, $p = 0.049$, $b = 032$, Standard error (SE) $= 0.52$], and female talker *F*02 in the raised *F*0 condition

($M = 0.89$, SD = 0.145, $z = 6.030$, $p < 0.001$, $b = 2.19$, SE = 0.48). In the reversed speech context condition, the expected response rates in five out of 12 conditions were significantly higher than chance, including female talker *F*01 ($M = 0.52$, SD = 0.322, $z = 1.977$, $p = 0.04$, $b = 0.12$, SE = 0.41) and male talker *M*01 in the unshifted *F*0 condition ($M = 0.58$, SD = 0.328, $z = 2.634$, $p = 0.08$, $b = 0.41$, SE = 0.42), and three talkers *F*02 ($M = 0.78$, SD = 0.243, $z = 6.030$, $p < 0.001$, $b = 1.61$, SE = 0.45), *M*01 ($M = 0.72$, SD = 0.226, $z = 5.527$, $p < 0.001$, $b = 1.08$, SE = 0.32) and *M*02 ($M = 0.68$, SD = 0.364, $z = 2.607$, $p = 0.009$, $b = 0.42$, SE = 0.81) in the raised *F*0 condition. In the anomalous speech context condition, most conditions reached significance (p's < 0.05), except for two talkers *F*01 and *M*02 in the unshifted condition (p's > 0.05). Finally, in the meaningful speech context condition, the expected response rates were significantly higher than chance in all talker and *F*0 shift conditions (p's < 0.01).

For the amusic group, largely similar results were found despite some small differences mostly in the anomalous context condition. In the nonspeech context condition, the expected response rates were significantly higher than chance only in two conditions: female talker *F*01 in the unshifted *F*0 condition ($M = 0.55$, SD = 0.195, $z = 4.481$, $p < 0.001$, $b = 0.23$, SE = 0.21), and male talker *M*01 in the raised *F*0 condition ($M = 0.52$, SD = 0.302, $z = 2.138$, $p = 0.03$, $b = 0.09$, SE = 0.37). In the reversed speech context condition, the expected response rates reached significance in half the conditions: talkers *F*01 and *M*02 in the unshifted *F*0 condition ($M = 0.61$, SD = 0.235, $z = 4.491$, $p < 0.001$, $b = 0.49$, SE = 0.23; $M = 0.53$, SD = 0.219, $z = 3.691$, $p < 0.001$, $b = 0.11$, SE = 0.21), and all four talkers in the raised *F*0 condition (*F*01: $M = 0.58$, SD = 0.199, $z = 4.491$, $p < 0.001$, $b = 0.32$, SE = 0.32; *F*02: $M = 0.48$, SD = 0.266, $z = 2.097$, $p = 0.03$, $b = -0.09$, SE = 0.29; *M*01: $M = 0.70$, SD = 0.245, $z = 4.827$, $p < 0.001$, $b = 1.04$, SE = 0.36; *M*02: $M = 0.53$, SD = 0.304, $z = 2.368$, $p = 0.01$, $b = 0.33$, SE = 0.26). In the anomalous speech context condition, all conditions reached significance (p's < 0.05) except for the following four conditions: talkers *F*02, *M*01 and *M*02 in the lowered *F*0 condition (p's > 0.05), and male talker *M*01 in the unshifted *F*0 condition ($p > 0.05$). Interestingly, in the anomalous context condition, most of the conditions failing to reach significance were found in the lowered *F*0 condition for the amusic group, different from the results of the control group where the two conditions failing to reach significance were in the unshifted *F*0 condition. This result seems to suggest reduced sensitivity toward lowered contextual *F*0 in amusics. In the meaningful speech context condition, the expected response rates were significantly higher than chance in all conditions (p's < 0.05).

Last, logistic regression models were fitted to examine to what extent the participants' performance on tone normalization within each context type could be predicted by their musical abilities. The input was the expected response rates for each context condition, and the predictors were the participants' accuracy in the three musical subtests (out of key, offbeat, and mistuned subtests), with the two groups collapsed. The three predictors were added to the models consecutively, and likelihood ratio tests were conducted to compare two models with and without a certain predictor, similar to the procedure described above.

For the nonspeech context condition, no effects of any subtest reached significance (p's ≥ 0.05). For the reversed speech context condition, the participants' scores in the out-of-key subtest significantly accounted for the expected response rates [$\chi^2(1) = 4.37$, $p = 0.04$], but the offbeat and mistuned subtest did not reach significance. For the anomalous speech context condition, again, only the effect of the out-of-key subtest was significant [$\chi^2(1) = 91.35$, $p < 0.001$]. For the meaningful speech context condition, the participants' scores in the two pitch-based subtests were found to contribute significantly to the prediction of expected response rates [out-of-key: $\chi^2(1) = 75.75$, $p < 0.001$; mistuned: $\chi^2(1) = 5.06$, $p = 0.02$], but the offbeat subtest was not significant [$\chi^2(1) = 1.55$, $p = 0.21$].

In summary, the analysis of expected response rates revealed important group differences in tone normalization. Amusics exhibited reduced context effects (i.e., lower rates of expected responses) compared to controls in meaningful and anomalous speech contexts, but comparable context effects to controls in reversed speech and nonspeech contexts. These results suggest that amusics exhibited reduced context effects compared with controls primarily in native speech contexts with phonological cues. On top of these group differences, it should be pointed out that largely similar patterns of tone normalization were observed for both controls and amusics, in that shifting *F*0 level in anomalous and meaningful speech contexts significantly changed the perception of the target tone to expected tones in almost all conditions within both groups. Finally, higher scores in the pitch-based musical subtests, especially the out-of-key subtest, were associated with better performance in tone normalization. Participants' scores in the out-of-key subtest significantly predicted the magnitude of the phonetic context effect in the reversed, anomalous, and meaningful speech contexts; their scores in the other pitch-based subtest—mistuned subtest—further predicted the magnitude of the phonetic context effect in meaningful speech contexts.

B. Results of the perceptual height scores

In addition to the expected response rates, perceptual height analysis was conducted to examine the perceptual bias toward high and low level tones in relation to the typicality of the four talkers' *F*0 range within each participant group. As can be seen in Fig. 1, the rate of expected responses was sporadically significantly higher than chance for two talkers in the nonspeech context condition for both groups, implying that the distribution of *F*0 cues in the nonspeech context might have an effect on tone normalization in some conditions. However, the previous study suggested that the sporadically significant effects could be an artefact of the influence of talker typicality[4] and the method for calculating expected tone responses (Zhang and Chen, 2016). For example, the target tone produced by a talker with an *F*0 range lower than the population mean *F*0 range may be misidentified as the low level tone; since the low level tone was the expected tone response in the raised *F*0 condition, this

calculation method may lead to a superficially significant effect for that talker in the raised $F0$ condition. In other words, sporadically significant effects in a few conditions cannot be taken as strong evidence for the effect of the distribution of $F0$ cues in a context; instead, the results may be largely driven by the influence of talker typicality regardless of the distribution of $F0$ cues in the context (raised, unshifted, or lowered).

In order to examine whether listeners were influenced by talker typicality in the current study and, more importantly, whether the control and amusic groups were influenced differently by talker typicality, multiple regression analyses were conducted for the control and amusic groups, respectively. Talker typicality was defined as the distance in semitone between each talker's upper and lower $F0$ range and the gender-matched population mean $F0$ range estimated from a Cantonese speech corpus (Lee *et al.*, 2002), the estimated mean $F0$ range being approximately 200–290 Hz for female talkers, and 110–160 Hz for male talkers. Each response was coded according to the perceptual height of the selected tone. A high level tone response was coded "6," a mid-level tone response was coded "3," and a low level tone response was coded "1," reflecting the pitch height relationship that the high level tone is 3 semitones higher than the mid-level tone ($6 - 3 = 3$), which is 2 semitones higher than the low level tone ($3 - 2 = 1$) (Francis *et al.*, 2006; Wong and Diehl, 2003). Linear regression models were constructed using the *lm* function in *R* (R Core Team, 2014) for each context condition in each group, collapsing the talker and $F0$ shift conditions. The perceptual height scores were the input and talker typicality was the factor in the models. An F-test to test the significance of the regression models and t-tests to test the significance of each factor (upper or lower $F0$ range) were conducted accordingly. Figures 2 and 3 show the perceptual height scores plotted as a function of talker typicality in each context condition for the control and amusic group, respectively.

For the control group, the model on the nonspeech context condition reached significance with an adjusted R^2 of 17.88% [$F(2, 189) = 21.8$, $p < 0.001$], and the typicality of the talkers' lower and upper $F0$ range both contributed significantly to the perceptual height scores (p's < 0.001). A total of 17.88% of the variance in the nonspeech context condition can be explained by a linear relationship of the typicality of the four talkers' lower and upper $F0$ range from the gender-matched population mean $F0$ range. A significant effect was also found in the reversed speech context with an adjusted R^2 of 7.9% [$F(2, 189) = 9.23$, $p < 0.001$]. Again, the typicality of the talkers' lower and upper $F0$ range both contributed significantly to the perceptual height scores (p's < 0.001). Finally, the model on the anomalous speech context also reached significance with a small adjusted R^2 of 2.3% [$F(2, 189) = 3.274$, $p = 0.04$]. However, neither the upper nor lower $F0$ range contributed significantly to the perceptual height scores (p's > 0.05), suggesting that this small effect may not warrant any meaningful interpretation. The model on the meaningful speech context condition was not significant. Lack of significant effect of talker typicality together with the significant effects of contextual $F0$ shift on expected response rates reported above suggest that controls were no longer affected by talker typicality, and largely perceived the target tone according to the distribution of $F0$ cues (raised, unshifted, or lowered) in the meaningful speech context.

For the amusic group, only the model on the reversed speech context reached significance with a small adjusted R^2 of 3.2% [$F(2, 189) = 4.193$, $p = 0.017$]. The typicality of the talkers' lower and upper $F0$ range both contributed significantly to the perceptual height scores (p's < 0.05). Although the model reached significance, the amount of variance in perceptual height scores that could be accounted for by talker typicality (3.2%) was smaller compared to that in the control group (7.9%). Models on the other three types of contexts were not significant.

In summary, the perceptual height analysis confirmed the influence of talker typicality on tone perception in controls, a finding largely consistent with previous studies (Zhang and Chen, 2016). More importantly, this analysis revealed differences between controls and amusics. Although the results of the expected response rates above showed that controls and amusics performed comparably in the nonspeech context, the perceptual height analysis revealed that amusics showed no influence of talker typicality in the nonspeech context, unlike the control group; in addition, amusics exhibited a weaker influence of talker typicality in the reversed speech context compared with controls.

IV. DISCUSSION

In this study, we examined the performance of Cantonese-speaking amusics and musically intact controls on the normalization of Cantonese level tones in four types of contexts that varied in the proximity of being native, meaningful speech for Cantonese listeners. We hypothesized that control listeners would show the strongest and most consistent effects on tone normalization in native speech contexts with phonological cues. We further hypothesized that amusics would be impaired in tone normalization, showing null or reduced context effects, especially in native speech contexts. The results showed that amusics demonstrated reduced context effects in terms of the expected response rates in anomalous and meaningful contexts, and performed comparably to controls in nonspeech and reversed speech contexts. Despite their similar performance in terms of the expected response rates, amusics exhibited no influence of talker typicality in the nonspeech context and reduced influence of talker typicality in the reversed speech context. In the discussion below, we first focus on the reduced context effects in amusics, and then on the reduced influence of talker typicality, in relation to the underlying mechanisms of the linguistic pitch processing deficit of amusics and the mechanisms of talker normalization.

A. Perceptual normalization of lexical tones according to the distribution of *F0* cues in context

Previous studies have widely reported the importance of the distribution of acoustic cues in context for accommodating

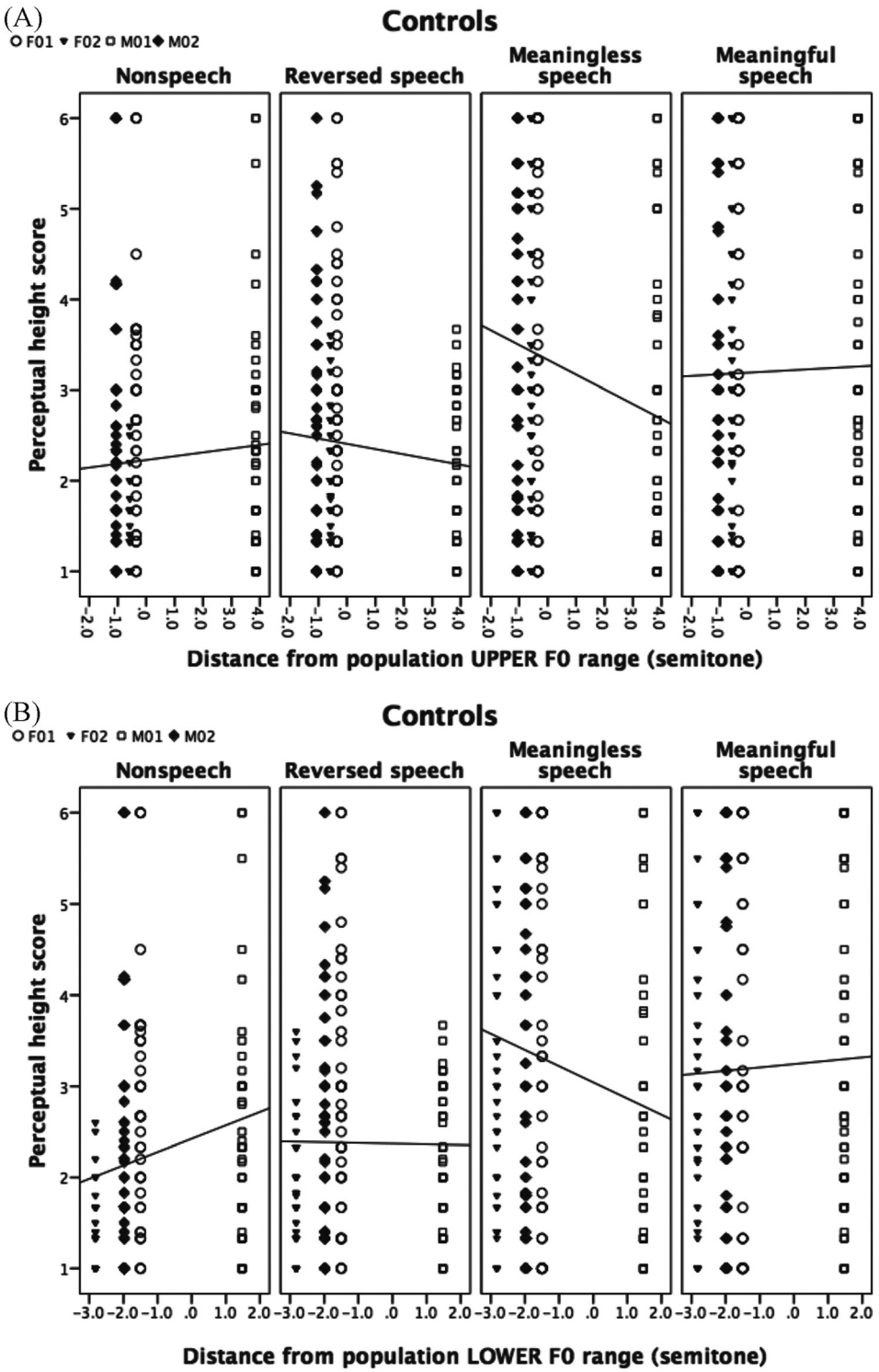

FIG. 2. Perceptual height scores plotted as a function of the distance of the four talkers' lower/upper *F*0 range from the gender-matched population mean *F*0 range for the control group. (A) Distance of the four talkers' upper *F*0 range from the gender-matched population mean *F*0 range. (B) Distance of the four talkers' lower *F*0 range from the gender-matched population mean *F*0 range.

talker variation in the signal-to-representation mapping (Chen and Peng, 2016; Francis *et al.*, 2006; Huang and Holt, 2009, 2011; Moore and Jongman, 1997; Peng *et al.*, 2012; Wong and Diehl, 2003; Zhang *et al.*, 2012, 2013; Zhang, 2018; Zhang *et al.*, 2016; Zhang and Chen, 2016). In particular, native speech contexts with phonological cues are found to elicit the strongest and most consistent effect on tone normalization (Chen and Peng, 2016; Francis *et al.*, 2006; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013).

In this study, the performance of controls largely replicated previous results (Chen and Peng, 2016; Francis *et al.*, 2006; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013), demonstrating stronger effects on tone normalization in contexts with richer cues, such that the strongest effects were observed in the meaningful speech context, followed by the anomalous speech context, the reversed speech context, and finally the nonspeech context. Whereas shifting the *F*0 level in meaningful and anomalous speech contexts elicited significant, contrastive changes in tone perception in almost all conditions, the same *F*0 shift in nonspeech and reversed speech contexts elicited much less consistent effects. Further analysis on the influence of talker typicality suggested that

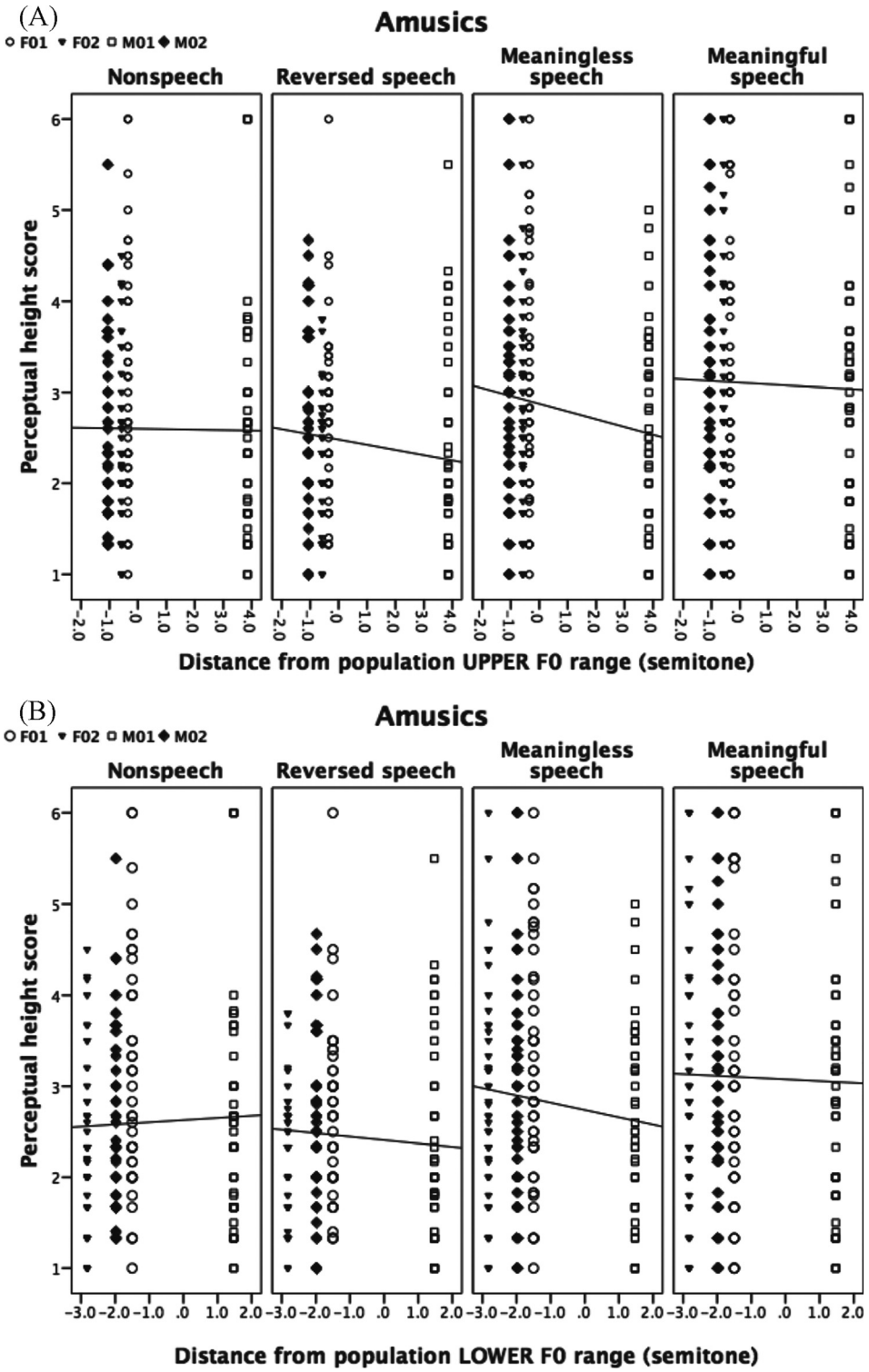

FIG. 3. Perceptual height scores plotted as a function of the distance of the four talkers' lower/upper *F*0 range from the gender-matched population mean *F*0 range for the amusic group. (A) Distance of the four talkers' upper *F*0 range from the gender-matched population mean *F*0 range. (B) Distance of the four talkers' lower *F*0 range from the gender-matched population mean *F*0 range.

the few sporadically significant effects in nonspeech and reversed speech context conditions were likely to be superficial, largely driven by the influence of talker typicality and the method for calculating the expected response rate. We argued that these sporadic effects could not be taken as strong evidence for the phonetic context effect in nonspeech and reversed speech contexts. As mentioned earlier, the nonspeech context only contained auditory pitch cues for perceiving the auditory pitch contrast between the context and target word. The reversed speech context [e.g., /ia his (n)o kin/] sounded foreign to Cantonese listeners and could be deemed as non-native speech. On the other hand, both anomalous and meaningful speech contexts were native speech contexts with phonological cues (i.e., native phonemes and tones) to estimate a talker's tone space. These results are largely consistent with previous findings demonstrating that tone normalization in control listeners is primarily driven by phonological cues contained in native speech contexts (Chen and Peng, 2016; Francis *et al.*, 2006; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013).

Compared to controls, amusics demonstrated reduced context effects in terms of the expected response rates in anomalous and meaningful speech contexts, but performed comparably to controls in nonspeech and reversed speech contexts. These results suggest that Cantonese-speaking amusics were particularly impaired in tone normalization via phonological cues in native speech contexts. The core deficiency of amusics appears to lie in phonological processing—in the dynamic evaluation and adjustment of the tone category according to its relative *F*0 location within a talker's shifted tone space embodied in phonological cues in native speech contexts.

It should be noted that the lack of group differences in nonspeech and reversed speech contexts could not be attributed simply to floor effects. Although both controls and amusics performed largely at chance in the nonspeech context, which might indeed diminish possible group differences, their performance was above chance in the reversed speech context. This result was supported by mixed-effects models conducted to compare the expected response rates against chance in nonspeech and reversed speech contexts (collapsing talker and *F*0 shift conditions). The results showed that the expected response rates were significantly higher than chance in the reversed speech context for both the control group ($z = 5.34$, $p < 0.001$) and amusic group ($z = 4.66$, $p < 0.001$), while the expected response rates in the nonspeech context were not significantly different from chance for either group.

Another question is whether the inferior performance of amusics in tone normalization was because of their poor ability to perceive the pitch distance between the target word and the context. When the contextual *F*0 was shifted upwards or downwards, changing the pitch distance between the target word and the context, the perception of the target tone would be adjusted accordingly. Many studies have reported that amusics have larger (less sensitive) pitch threshold than controls, requiring larger pitch changes to detect differences between two consecutive pitch stimuli in pitch height (e.g., high or low) or pitch direction (e.g., rising or falling) (Liu *et al.*, 2012b). In the current study, the *F*0 of the context was shifted upwards or downwards by three semitones, an *F*0 manipulation determined from the *F*0 distance between the three level tones and confirmed by previous studies (Francis *et al.*, 2006; Wong and Diehl, 2003; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013). This magnitude of *F*0 shift was capable of changing the perception of the target tone contrastively in the control group. It is possible that for some amusics, especially those with very large (less sensitive) pitch threshold, a larger contextual *F*0 shift might be required for them to perceive the *F*0 shift to a similar extent as controls. Nonetheless, it should be noted that this factor alone could not explain the main results. If the deficit of amusics in tone normalization is purely due to their poor ability to perceive pitch contrast, it is reasonable to expect amusics to exhibit similar impairment in tone normalization across all four types of contexts, which carried the same *F*0 shift manipulations. But amusics demonstrated reduced context effects primarily in anomalous and meaningful speech contexts. These results appear to suggest that the deficit of amusics in tone normalization primarily originates from their impoverished ability to make use of phonological cues in native speech contexts. That being said, any impairment of amusics in perceiving pitch distance probably has some contribution to their inferior performance, and presumably both factors (making use of phonological cues in native speech contexts and perceiving the pitch distance) played a role in tone normalization. Future studies may investigate whether enlarging the magnitude of *F*0 shift above all amusics' pitch threshold in anomalous and meaningful speech contexts could render the group differences smaller or even null in tone normalization.

The findings of the current study extended previous studies on amusia in tone perception at the single word level (Huang *et al.*, 2015; Jiang *et al.*, 2012; Liu *et al.*, 2015b; Nan *et al.*, 2010; Shao *et al.*, 2016; Wang and Peng, 2014), demonstrating that amusia affects tone normalization according to the distribution of phonological cues in native speech contexts. These findings suggest that amusia not only affects the perception of tones based on *F*0 cues within a word, but also negatively impacts the perception and adjustment of the tone category according to its location in a talker's shifted tone space. These results are to some extent compatible with the findings that amusics are impaired in intonation processing and musical melodic contour processing, which involves more global pitch processing beyond the scale of a single word or a single musical note (Jiang *et al.*, 2010; Liu *et al.*, 2010).

The current results are consistent with previous findings that phonological processing of tones is impaired in amusics (Huang *et al.*, 2015; Jiang *et al.*, 2012; Wang and Peng, 2014; Zhang *et al.*, 2017b). It has been found that Mandarin-speaking amusics exhibited no benefit for between-category discriminations of tone stimuli relative to within-category discriminations, which indicates lack of categorical perception of tones (Jiang *et al.*, 2012). Reduced benefit for between-category discriminations of lexical tones was also reported in Cantonese-speaking amusics (Zhang *et al.*, 2017b). The findings of the current study converged with these results in suggesting that amusia affects phonological processing of tones in tonal language speakers.

It should be noted that amusics, despite displaying reduced context effects on tone normalization via native phonological cues, still exhibited better than chance performance in almost all conditions in anomalous and meaningful speech contexts, a pattern largely similar to the performance of controls. This result indicates that the perceptual normalization system in the brain of amusics is probably partially intact. Amusics were able to make use of phonological cues to some extent, displaying certain tone normalization abilities. This result further implies that the deficit of amusics in tone normalization might be remediable through training and intervention, a question that awaits future investigation.

Finally, the findings of the current study can shed some light on the mechanisms of talker normalization. As mentioned above, the identification of Cantonese tones produced by multiple talkers is heavily influenced by talker typicality (Bishop and Keating, 2012; Peng *et al.*, 2012; Zhang and Chen, 2016; Zhang *et al.*, 2012), and native speech contexts

with phonological cues to map out a talker's tone space are most critical for tone normalization (Chen and Peng, 2016; Francis *et al.*, 2006; Zhang and Chen, 2016; Zhang *et al.*, 2012, 2013). The current results corroborated these previous findings, demonstrating that anomalous and meaningful speech contexts with native phonological cues induced the strongest and most consistent effects on tone normalization in control listeners. Interestingly, these results are also consistent with a different line of studies on talker voice processing. It has been reported that talker identification is more accurate in listeners' native language than non-native language, suggesting that listeners benefit from native phonological cues in the processing of a talker's voice characteristics (e.g., Perrachione *et al.*, 2009). As talker normalization presumably involves talker voice processing at an earlier stage, it is possible that native speech contexts may benefit talker voice processing as well as the mapping of a talker's tone space, which subsequently facilitates the perception of the location of a target speech sound within the tone space.

B. Influence of talker typicality on lexical tone perception

Previous studies have found that the perception of Cantonese level tones produced by unfamiliar talkers was influenced by the typicality of a talker's *F*0 range relative to the population mean *F*0 range (Peng *et al.*, 2012; Zhang and Chen, 2016; Zhang *et al.*, 2012). The influence of talker typicality was most prominent when the tones were presented in isolation without any extra contextual cues. In nonspeech contexts with only auditory pitch cues, a small but significant influence of talker typicality was still observed, which was similar to when the tones were presented in isolation. But in anomalous and meaningful speech contexts where there were phonological cues for talker adaptation, listeners demonstrated high accuracy in tone identification across all talkers, and the influence of talker typicality disappeared. Based on these results, it is conjectured that a set of tone templates/representations shaped by the *F*0 range of typical talkers in each gender are present in the brain of Cantonese listeners (Zhang and Chen, 2016). When listening to tones produced by an unfamiliar talker without useful contextual cues for talker adaptation, Cantonese listeners resort to these tone templates/representations. However, if an unfamiliar talker's *F*0 range is less typical, these tone templates/representations lead to perceptual biases instead of accurate tone percepts. In such cases, it is proposed that native speech contexts with cues to a talker's tone space are critical for tone normalization (Zhang and Chen, 2016). When native speech contexts are available, listeners primarily perceive tones according to the distribution of a talker's tone space, and are no longer influenced by those tone templates/representations. It is speculated that these tone templates/representations are formed from the accumulated auditory exposure to exemplars of lexical tones produced by a large sample of talkers, especially typical ones, in a listener's past auditory experience (Zhang and Chen, 2016).

In the current study, the performance of controls largely replicated the previous results (Zhang and Chen, 2016), showing a significant influence of talker typicality on tone perception in the nonspeech context, and to a smaller extent in the reversed speech context. This result implies that controls may have been influenced by tone templates/representations in nonspeech and reversed speech contexts. It is intriguing that the influence of talker typicality was not found in amusics in the nonspeech context and reduced in the reversed speech context. This result seems to suggest that the set of tone templates/representations shaped by typical talkers' *F*0 range might be formed differently or less robustly in the brain of amusics. Since amusia is an innate pitch processing disorder, it is possible that the impoverished pitch processing ability could affect the robustness of tone templates/representations or their *F*0 characteristics formed in the brain. This possibility could explain why the effects of talker typicality were null or reduced in amusics. Future studies could investigate the influence of talker typicality on tone perception in amusia.

V. CONCLUSION

To conclude, we found that Cantonese-speaking amusics were impaired in the perceptual normalization of lexical tones through the distribution of *F*0 cues in anomalous and meaningful speech contexts, which involves the adjustment of the perceived *F*0 location of a tone within a talker's shifted tone space. The impairment in tone normalization, though severe, appears not to be disabling, since amusics preserved some tone normalization abilities. This result implies that the perceptual normalization system in the amusic brain might be at least partially intact. Amusics also showed null or less influence of talker typicality on tone perception compared to controls in nonspeech and reversed speech context conditions, which suggests that the set of tone templates/representations shaped by typical talkers' *F*0 range might be formed differently or less robustly in the amusical brain. These results extended the findings of previous studies and shed some light on the nature of deficits of amusic in linguistic pitch processing and the mechanisms of talker normalization.

This study also has a few limitations that wait to be addressed in future studies. First, future studies should include a condition without any context (i.e., isolated target speech signals), which could provide a baseline for examining the context effect on tone perception in the comparison of the performance of amusics and controls. The isolation condition would also enable researchers to better examine the influence of talker typicality on tone perception and compare the performance of amusics and controls in this regard, since the influence of talker typicality was strongest when listeners perceived target speech signals in isolation (Zhang and Chen, 2016). Second, while the current study confirmed that Cantonese-speaking amusics were impaired in utilizing phonetic context for the normalization of three Cantonese level tones, it remains an open question how this finding translates to their difficulty in daily speech recognition. The current findings imply that amusics may be less efficient at

adapting to talker variation in speech comprehension and less accurate in recognizing speech utterances produced by unfamiliar talkers. While there is some evidence that the speech processing deficit of amusics goes beyond pitch processing, affecting vowel perception and speech comprehension (Liu *et al.*, 2015a; Zhang *et al.*, 2017b), how speech recognition on a broader scale (e.g., consonants and vowels) and in a more naturalistic listening setting is affected by amusia remains to be examined in future studies.

ACKNOWLEDGMENTS

This work was supported by grants from the Research Grants Council of Hong Kong (ECS: Grant No. 25603916), the National Natural Science Foundation of China (NSFC: Grant No. 11504400), and the PolyU Start-up Fund for New Recruits (Grant No. 1-ZE4Y). We thank the Editor and two anonymous reviewers for constructive comments. We thank Yike Yang for help with data collection, and James Porteous for proofreading earlier versions of this paper.

[1]Note that *F*0 range is a complex notion including "*F*0 span" and "*F*0 level" (Ladd, 1996).

[2]Throughout the paper, a tone is annotated using Chao's tone letters, which are in the range of 1–5, with 5 referring to the highest pitch and 1 referring to the lowest pitch (Chao, 1930). Each tone is described using two letters, which indicate the pitch at the beginning and end of a syllable.

[3]See supplementary material at https://doi.org/10.1121/1.5049147 for supplementary Tables I and II.

[4]Note that talkers may differ in the typicality of a number of acoustic cues other than *F*0, such as phonation type (e.g., creaky voice and breathy voice) and formant frequencies. But these differences in typicality are not the focus of this study.

Ayotte, J., Peretz, I., and Hyde, K. (**2002**). "Congenital amusia: A group study of adults afflicted with a music-specific disorder," Brain **125**(2), 238–251.

Bishop, J., and Keating, P. (**2012**). "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex," J. Acoust. Soc. Am. **131**(2), 1100–1112.

Boersma, P., and Weenink, D. (**2014**). Praat: Doing phonetics by computer. (Version 6.0.12) [computer program], available at http://www.praat.org (Last viewed September 4, 2014).

Chao, Y.-R. (**1930**). "A system of tone letters," Le Maître Phonétique **45**, 24–27.

Chen, F., and Peng, G. (**2016**). "Context effect in the categorical perception of Mandarin tones," J. Signal Process. Sys. **82**(2), 253–261.

Fowler, C. A. (**2006**). "Compensation for coarticulation reflects gesture perception, not spectral contrast," Percept. Psychophys. **68**(2), 161–177.

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., and Chu, P. C. Y. (**2006**). "Extrinsic context affects perceptual normalization of lexical tone," J. Acoust. Soc. Am. **119**(3), 1712–1726.

Gerstman, L. (**1968**). "Classification of self-normalized vowels," IEEE Trans. Audio Electroacoust. **16**(1), 78–80.

Holt, L. L., and Kluender, K. R. (**2000**). "General auditory processes contribute to perceptual accommodation of coarticulation," Phonetica **57** (2–4), 170–180.

Honorof, D. N., and Whalen, D. H. (**2005**). "Perception of pitch location within a speaker's F0 range," J. Acoust. Soc. Am. **117**(4), 2193–2200.

Huang, J., and Holt, L. L. (**2009**). "General perceptual contributions to lexical tone normalization," J. Acoust. Soc. Am. **125**(6), 3983–3994.

Huang, J., and Holt, L. L. (**2011**). "Evidence for the central origin of lexical tone normalization (L)," J. Acoust. Soc. Am. **129**(3), 1145–1148.

Huang, W. T., Liu, C., Dong, Q., and Nan, Y. (**2015**). "Categorical perception of lexical tones in Mandarin-speaking congenital amusics," Frontiers Psychol **6**, 829.

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., and Yang, Y. (**2010**). "Processing melodic contour and speech intonation in congenital amusics with Mandarin Chinese," Neuropsychol. **48**(9), 2630–2639.

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., and Yang, Y. (**2012**). "Impaired categorical perception of lexical tones in Mandarin-speaking congenital amusics," Memory Cognit. **40**(7), 1109–1121.

Johnson, K. (**2005**). "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell Publishing, Hoboken, NJ), pp. 363–389.

Ladd, R. D. (**1996**). *Intonational Phonology* (Cambridge University Press, London).

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Laing, E. J. C., Liu, R., Lotto, A. J., and Holt, L. L. (**2012**). "Tuned with a tune: Talker normalization via general auditory processes," Front. Psychol. **3**, 1–9.

Lee, T., Lo, W. K., Ching, P. C., and Meng, H. (**2002**). "Spoken language resources for Cantonese speech processing," Speech Commun. **36**(3), 327–342.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**(6), 431–461.

Liu, F., Jiang, C., Thompson, W. F., Xu, Y., Yang, Y., and Stewart, L. (**2012a**). "The mechanism of speech processing in congenital amusia: Evidence from Mandarin speakers," PLoS One **7**(2), e30374.

Liu, F., Jiang, C., Wang, B., Xu, Y., and Patel, A. D. (**2015a**). "A music perception disorder (congenital amusia) influences speech comprehension," Neuropsychologia **66**, 111–118.

Liu, F., Maggu, A. R., Lau, J. C. Y., and Wong, P. C. M. (**2015b**). "Brainstem encoding of speech and musical stimuli in congenital amusia: Evidence from Cantonese speakers," Frontiers Human Neurosci. **8**, 1029.

Liu, F., Patel, A. D., Fourcin, A., and Stewart, L. (**2010**). "Intonation processing in congenital amusia: Discrimination, identification and imitation," Brain **133**(6), 1682–1693.

Liu, F., Xu, Y., Patel, A. D., Francart, T., and Jiang, C. (**2012b**). "Differential recognition of pitch patterns in discrete and gliding stimuli in congenital amusia: Evidence from Mandarin speakers," Brain Cognit. **79** (3), 209–215.

Moore, C. B., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**(3), 1864–1877.

Nan, Y., Sun, Y., and Peretz, I. (**2010**). "Congenital amusia in speakers of a tone language: Association with lexical tone agnosia," Brain **133**(9), 2635–2642.

Patel, A. D., Wong, M., Foxton, J., Lochy, A., and Peretz, I. (**2008**). "Speech intonation perception deficits in musical tone deafness (congenital amusia)," Music Perception **25**(4), 357–368.

Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. W., and Wang, W. S.-Y. (**2012**). "The effect of inter-talker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems," J. Speech Lang. Hear. Res. **55**(2), 579–595.

Peretz, I., Ayotte, J., Zatorre, R. J., Mehler, J., Ahad, P., Penhune, V. B., and Jutras, B. (**2002**). "Congenital amusia: A disorder of fine-grained pitch discrimination," Neuron **33**(2), 185–191.

Peretz, I., Gosselin, N., Tillmann, B., Cuddy L. L., Gagnon, B., Trimmer C. G., Paquette, S., and Bouchard, B. (**2008**). "On-line identification of congenital amusia," Music Perception **25**(4), 331–343.

Peretz, I., and Vuvan, D. T. (**2017**). "Prevalence of congenital amusia," European J. Human Genetics **25**(5), 625–630.

Perrachione T. K., Pierrehumbert, J. B., and Wong, P. C. M. (**2009**). "Differential neural contributions to native- and foreign-language talker identification," J. Exp. Psychol. **35**(6), 1950–1960.

Pfeifer, J., and Hamann, S. (**2015**). "Revising the diagnosis of congenital amusia with the Montreal battery of evaluation of amusia," Frontiers Human Neurosci. **9**, 161.

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

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

Sjerps, M. J., Mitterer, H., and McQueen, J. M. (**2011**). "Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics," Neuropsychologia **49**(14), 3831–3846.

Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**(4), 1086–1100.

Vuvan, D. T., Nunes-Silva, M., and Peretz, I. (**2015**). "Meta-analytic evidence for the non-modularity of pitch processing in congenital amusia," Cortex **69**, 186–200.
Wang, J., Zhang, C., Wan, S., and Peng, G. (**2017**). "Is congenital amusia a disconnection syndrome? A study combining tract- and network-based analysis," Frontiers Human Neurosci. **11**, 473.
Wang, X., and Peng, G. (**2014**). "Phonological processing in Mandarin speakers with congenital amusia," J. Acoust. Soc. Am. **136**(6), 3360–3370.
Wong, P. C. M., Ciocca, V., Chan, A. H. D., Ha, L. Y. Y., Tan, L.-H., and Peretz, I. (**2012**). "Effects of culture on musical pitch perception," PLoS One **7**(4), e33424.
Wong, P. C. M., and Diehl, R. L. (**2003**). "Perceptual normalization for inter- and intratalker variation in Cantonese level tones," J. Speech, Lang., Hear. Res. **46**(2), 413–421.
Zhang, C. (**2018**). "Online adjustment of phonetic expectation of lexical tones to accommodate speaker variation: A combined behavioural and ERP study," Lang., Cognit., Neurosci. **33**(2), 175–195.
Zhang, C., and Chen, S. (**2016**). "Toward an integrative model of talker normalization," J. Exp. Psychol. **42**(8), 1252–1268.
Zhang, C., Peng, G., Shao, J., and Wang, W. S. Y. (**2017a**). "Neural bases of congenital amusia in tonal language speakers," Neuropsychologia **97**(July 2016), 18–28.
Zhang, C., Peng, G., and Wang, W. S.-Y. (**2012**). "Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones," J. Acoust. Soc. Am. **132**(2), 1088–1099.
Zhang, C., Peng, G., and Wang, W. S.-Y. (**2013**). "Achieving constancy in spoken word identification: Time course of talker normalization," Brain Lang. **126**(2), 193–202.
Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., Peng, G., and Wang, W. S.-Y. (**2016**). "Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study," NeuroImage **124**, 536–549.
Zhang, C., and Shao, J. (**2018**). "Normal pre-attentive and impaired attentive processing of lexical tones in Cantonese-speaking congenital amusics," Sci. Reports **8**, 8420.
Zhang, C., Shao, J., and Huang, X. (**2017b**). "Deficits of congenital amusia beyond pitch: Evidence from impaired categorical perception of vowels in Cantonese-speaking congenital amusics," PLoS One **12**(8), e0183151.