Toward an integrative model of talker normalization

Caicai Zhang[a,b,*], Si Chen[a]

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

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

* Corresponding author at:

Caicai Zhang: Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China. Tel: (+852) 3400 8465. E-mail address: caicai.zhang@polyu.edu.hk.

Abstract

Successful speech perception requires accurate mapping of speech signals to linguistic categories despite talker variation in signals. Although factors like intrinsic and context cues have been identified, a full understanding of talker normalization remains to be achieved. In particular, it is important to examine the co-contribution of intrinsic, extrinsic and other cues in an integrative way. In Experiment 1, we examined the effect of intrinsic cues and typicality of a talker's F0 range relative to population F0 range on word identification in isolation. In Experiment 2, we compared the effects of four contexts to identify those that consistently facilitate talker normalization. We found that without contexts, word identification accuracy was low and variable depending on talker typicality. Contexts improved performance across all talkers regardless of typicality. But only meaningless and meaningful speech contexts with cues to a talker's acoustic-phonological space showed consistent effects. We proposed a new model, integrating talker typicality, talker familiarity and context. Whereas speech signals from familiar or typical talkers may be accurately identified standing alone, a context with cues to a talker's acoustic-phonological space is necessary in the case of unfamiliar and atypical talkers. It is thus the first model that integrates memory and context effects.

Keywords: Talker normalization, intrinsic normalization, extrinsic normalization, talker typicality, lexical tone

A fundamental question in speech perception is how listeners manage to recognize speech sounds produced by different talkers. Due to differences between talkers such as pitch, voice quality and vocal tract length, the same speech sound produced by different talkers is acoustically different (Koenig, 2000; Morris, McCrea, & Herring, 2008; Peterson & Barney, 1952; Peng, 2006; Remez, Fellowes, & Rubin, 1997). Talker variation has been widely observed in the acoustics of segmentals such as vowels (e.g., Peterson & Barney, 1952) and consonants (e.g., Koenig, 2000; Morris et al., 2008), and suprasegmentals such as lexical tones (e.g., Peng, 2006). Talker differences increase variation within the same linguistic category and create overlap between different categories, leading to complexities in the mapping of speech signals to linguistic categories (e.g., Peng, Zhang, Zheng, Minett, & Wang, 2012).

Despite the large amount of talker variation in speech signals, listeners manage to accurately perceive speech sounds produced by different talkers. The process of accurately mapping speech signals to linguistic categories despite talker variation is referred to as talker normalization in the current study. Although a full understanding of talker normalization remains to be achieved, several mechanisms have been proposed in the literature, each focusing on different types of cues in the speech signal.

One mechanism is to reduce or 'normalize' acoustic variation of speech sounds by rescaling/transforming the acoustic cues relative to *intrinsic* acoustic cues within the same speech signal (e.g., Syrdal & Gopal, 1986; Fujisaki & Kawashimi, 1968; Monahan & Idsardi, 2010; Johnson, 1990; Slawson, 1968). For example, whereas the absolute frequencies of the first two vowel formants (F1 and F2), which are critical for vowel perception, vary dramatically due to talker variation (e.g., Peterson & Barney, 1952), other intrinsic cues such as the frequency of the third formant (F3) and fundamental frequency (F0) also vary but in a way indicative of a specific talker's voice characteristics. Syrdal & Gopal (1986) found that rescaling the F1 and F2 frequencies relative to the frequency of F3 and F0 in bark scale reduced talker variation, and increased the accuracy of binary vowel classification in each dimension (high/low, front/back). The F1-F0 dimension is found to represent vowel height, and the F3-F2 dimension represents vowel frontness/backness. Because of its emphasis on intrinsic talker cues, this mechanism is called intrinsic normalization.

A second mechanism emphasizes the cues from *outside* the speech signal, i.e., a context (e.g., Holt, 2005, 2006a, 2006b; Holt, Lotto, & Kluender, 1996; Kingston, Kawahara, Mash, & Chambless, 2011; Laing et al., 2012; Mann, 1980; Mann & Repp, 1981; Gerstman, 1968; Huang & Holt, 2012; Ladefoged & Broadbent, 1957; Nearey, 1989; Nearey & Assmann, 1986; Shankweiler, Strange, & Verbrugge, 1977; Sjerps et al., 2011a, 2011b, 2012; Chen & Peng, 2015; Francis, Ciocca, Wong, Leung, & Chu, 2006; Huang & Holt, 2009; Leather, 1983; Lin & Wang, 1984; Moore & Jongman, 1997; Wong & Diehl, 2003; Zhang et al., 2012, 2013). With context carrying information of extremes of a phonological space (e.g., /a/, /i/, /u/ of the vowel space), listeners can build a talker-specific acoustic-phonological space, which serves as a frame/reference to estimate the location/identity of a target speech signal (e.g., Joos, 1948). For example, Ladefoged and Broadbent (1957) found that the identification of synthetic /bVt/ stimuli depended on the relative F1 distance between the target vowel and the carrier sentence. The same target vowel was more frequently identified as /ɪ/ when the carrier sentence had a high F1, and more frequently identified as /ɛ/ when the carrier sentence had a low F1. The effect of context is contrastive, since the shift of contextual F1 cues elicited a contrastive change in the perception of the target vowel. This indicates that the location/identity of the target vowel is evaluated according to the distribution of formant frequencies in the context. When a talker's acoustic-phonological space is shifted to a region with higher F1 values, the target vowel is perceived as having lower F1, and vice versa. Because of its emphasis on extrinsic cues, this mechanism is referred to as extrinsic normalization.

Though both intrinsic and extrinsic normalization have gained support in the literature, the approach to look at individual cues separately (intrinsic or extrinsic cues) while ignoring other cues is not without problems. In order to solve the task of accurately mapping speech signals to linguistic categories, listeners must make use of all cues available, intrinsic and extrinsic, to assist talker normalization. It is therefore important to take an approach that integrates different cues to understand their co-contribution to talker normalization, which is the aim of the current study.

We use Cantonese level tones to investigate this question in the current study. There are three lexically contrastive level tones in Cantonese, high level tone (e.g., 醫 /ji55[1]/ 'a doctor'), mid level tone (e.g., 意 /ji33/ 'meaning') and low level tone (二 /ji22/ 'second') (Chao, 1947). We chose Cantonese level tones for two reasons. Firstly, there is great talker variation in the acoustics of Cantonese level tones, such that the F0 of these tones produced by talkers with different F0 ranges overlap substantially (Peng, 2006; Peng et al., 2012). It therefore allows us to examine how various factors, intrinsic, extrinsic and other cues, contribute to the normalization of lexical tones. Secondly, level tones are primarily distinguished by a simple phonetic cue, i.e., pitch height, while other cues such as the pitch contour and phonation type are largely similar. Cantonese level tones are therefore easy to control. Although many previous studies on talker normalization focused on vowel or consonant perception (Syrdal & Gopal, 1986; Fujisaki & Kawashimi, 1968; Monahan &

[1] Note that a tone is annotated using Chao's tone letters, which are in the range of 1-5, with 5 being the highest pitch and 1 being the lowest pitch (Chao, 1930). Each tone is described using two or more numbers, which indicate the pitch at the beginning and end of a syllable.

Idsardi, 2010; Johnson, 1990; Slawson, 1968; Joos, 1948; Sjerps, Mitterer, & McQueen, 2011a, 2011b, 2012), the mechanism of accommodating talker variation in the perception of speech sounds is probably general. Therefore findings generated from the investigation of lexical tones may also be informative for research on other types of speech sounds. In the text below, we will first review studies on intrinsic and extrinsic normalization with a focus on tone normalization, and then discuss the aim of the current study in detail.

Intrinsic Normalization

In order to accurately identify lexical tones, it is critical to be able to estimate a particular talker's F0 range (e.g., Wong, 1998; Wong and Diehl, 2003). This is because though the absolute F0 value of a tone produced by a talker can vary dramatically, it tends to be located close to the upper bound of a talker's F0 range if it is a high tone, and close to the lower bound if it is a low tone (Peng, 2006; Peng et al., 2012). However, so far the findings are mixed regarding whether a talker's F0 range can be estimated based on intrinsic cues.

Honorof and Whalen (2005) examined whether English listeners can identify where an isolated pitch sample is located within a speaker's pitch range (i.e., relative pitch height rating) based on intrinsic cues only. The authors found that the pitch location judgment correlated well with the actual location within each talker's pitch range, even though those talkers were unfamiliar to the listeners. This finding seemingly indicates that English listeners can judge pitch location with intrinsic cues. The authors speculated that intrinsic cues such as voice quality might have aided pitch location judgment.

However, Bishop and Keating (2012) failed to find a direct link between voice quality and pitch location judgment. Voice quality measures played a very limited role in the correlation with pitch location judgment, which was actually most strongly correlated with the absolute F0 value of the pitch stimuli. This suggests that listeners based their pitch location judgment primarily on the absolute F0 values instead of F0 range. Nonetheless, voice quality could influence pitch location judgment *indirectly* via gender judgment. It is found that listeners judged the pitch location differently depending on a speaker's gender, such that a pitch sample was rated as having higher pitch when it was perceived to be from a male speaker than a female speaker. Voice quality, which was significantly correlated with explicit speaker gender judgment, thus could influence pitch location judgment indirectly. In a nutshell, pitch location judgment was mostly strongly contributed to by the absolute F0 values, and the effect of voice quality was indirect and small. This also means that other cues, such as extrinsic cues, are probably important for talker normalization in pitch perception.

Extrinsic Normalization

There is plenty of evidence that listeners estimate the location/identity of a lexical tone according to the distribution of a talker's speaking F0 in a speech context (Chen & Peng, 2015; Francis, Ciocca, Wong, Leung, & Chu, 2006; Huang & Holt, 2009; Leather, 1983; Lin & Wang, 1984; Moore & Jongman, 1997; Wong & Diehl, 2003; Zhang et al., 2012, 2013).

Lin and Wang (1984) investigated how the relative F0 height of a single-word context affected the perception of Mandarin tones. An identical target word was perceived more frequently as having low tones when attached to a context produced with high F0, and more frequently as having high tones when attached to a context with low F0.

Leather (1983) explicitly examined the effect of speech context on the accommodation of between-talker variability. Talker-ambiguous Mandarin tone stimuli were embedded in speech utterances produced by two male talkers with different F0 ranges. Identical stimuli were perceived as different tones in a contrastive way, depending on which talker, high-pitch talker or low-pitch talker, was perceived to 'produce' them.

Moore and Jongman (1997) also found that speech contexts with a talker's F0 information affected the identification of talker-ambiguous tone stimuli in a contrastive way. The same word was more frequently identified as having a low tone if the context was produced by a talker with high mean F0, and more frequently as having a high tone in the context produced by a talker with low mean F0.

The aforementioned studies examined tone normalization in Mandarin. Wong and Diehl (2003) investigated this question in Cantonese. The authors also reported a contrastive context effect, i.e., raising and lowering the contextual F0 changed the perception of the target word with mid level tone to words to low level tone and high level tone in a contrastive way. Moreover, within the carrier sentence, words adjacent to the target word had a greater effect than those far away (Wong and Diehl, 2003).

Although the context effect has been widely found, it remains unclear whether the listeners actually estimate a talker's acoustic-phonological space from the context. A number of studies have found that nonspeech contexts matched with speech contexts in the long-term average spectrum elicit similar normalization effects as speech contexts do (Huang & Holt, 2009; 2012; Holt et al., 1996; Holt, 2005, 2006a, 2006b; Lotto & Kluender, 1998; Lotto, Sullivan, & Holt, 2003; Watkins, 1991; Watkins & Makin, 1996) or even stronger effects sometimes (e.g., Laing et al., 2012). Because nonspeech contexts only contain auditory cues, it is argued that listeners rely on a general auditory mechanism in talker normalization, rather than a talker's acoustic-phonological space built from the context.

Holt et al. (1996) found that a nonspeech context synthesized using frequency-modulated sine-wave glides that modeled the center frequency of formants in a speech context elicited a contrastive effect as the speech context did. However, it is possible that the nonspeech context may have been processed in the speech mode. The authors argued against this possibility, showing that listeners failed to label the glides as /b/ or /d/

consistently. Moreover, it is found that Japanese quail trained to peck in response to /da/ to /ga/ sounds showed a contrastive shift in the responses according to the preceding speech contexts as human listeners do (Lotto, Kluender, & Holt, 1997). Since it is unlikely for the Japanese quail to process the sounds in the human speech mode, the authors claimed that the results favored the broad generality of perceptual processes underlying the context effect.

As far as lexical tone perception is concerned, Huang and Holt (2009) found that nonspeech contexts (harmonic tones and pure tone contexts), which modeled the mean F0 of speech contexts, elicited a similar effect on the identification of tone stimuli ranging from Mandarin Tone 1 (high level tone) to Tone 2 (high rising tone). For both speech and nonspeech contexts, the effect was contrastive, such that the contexts with raised F0 led to more Tone 2 responses and vice versa (the mean F0 of Tone 2 was lower than that of Tone 1). Similarly contrastive effects of speech and nonspeech contexts led the authors to suggest that the general perceptual mechanism is engaged in lexical tone normalization.

However, it has also been argued that similar response patterns do not imply identical perceptual processes underlying the effect of speech and nonspeech contexts (Fowler, 2006; Viswanathan, Magnuson, & Fowler, 2013). While the effect of nonspeech context may be mediated by general auditory processing, affecting speech perception via general contrasts of auditory cues between the context and the target speech sound, the effect of speech context is likely mediated by more specific, speech-related processing. In support of this claim, it has been found that nonspeech contexts often show weaker or null effects on lexical tone normalization.

Chen and Peng (2015) compared the effect of speech and nonspeech contexts on the categorization of tone stimuli ranging from Mandarin Tone 1 to Tone 2. Inconsistent with the findings of Huang and Holt (2009), Chen and colleague found that only F0 shift in the speech context had an effect on the categorization of Mandarin tones.

Francis et al. (2006) compared the normalization effects of a speech context and a 'nonspeech' context generated with a 'hummed' neutral vocal tract (/ə/ like) on the categorization of Cantonese level tones. Whereas shifting the F0 of a speech context upward and downward changed the perception of a target word carrying Cantonese mid level tone to words with low level tone and high level tone contrastively, the neutral vocal tract context showed no such effect – the perception of the target word remained unchanged irrespective of the F0 height of the context. However, it should be noted that the 'nonspeech' status of the neutral vocal tract context is debatable, because it contained at least some formant structure. Despite its speech-likeness, the neutral vocal tract context did not affect tone normalization as speech contexts did.

In two studies, Zhang et al. (2012, 2013) further confirmed that the effects of speech and nonspeech context were unequal on the perception of Cantonese level tones, with a more careful design of the nonspeech context (using a triangle wave to generate the nonspeech context). The findings of Francis et al. (2006) were largely replicated. The speech context showed a contrastive effect on the perception of the target word, consistently for the target words produced by four talkers with different F0 ranges. However, the effect of the nonspeech contexts was small, almost negligible, and not consistent on the perception of the target word produced by all four talkers.

To summarize, the effect of speech context on tone normalization has been consistently found in different studies, whereas the effect of nonspeech context is sometimes smaller or even absent. It remains unclear whether talker normalization relies on a general perceptual mechanism or the estimation of a talker's acoustic-phonological space.

The Current Study

As discussed above, intrinsic and extrinsic normalization mechanisms have each gained support in the literature. But there are also some unresolved issues within each mechanism. More importantly, the approach to look at individual cues separately (intrinsic or extrinsic cues) is not without problems.

Firstly, in order to solve the task of accurately and quickly mapping speech signals to linguistic categories, listeners must make use of all cues available, intrinsic and extrinsic, to assist talker normalization. However, previous models of talker normalization tend to focus on one source of cues (intrinsic or extrinsic), while ignoring the other cues. The existing evidence calls for an approach that integrates different cues in talker normalization. Such an integrative approach is especially important, because isolating a component from the whole and studying it could change its behavior from when it functions within the whole. Putting back isolated components together does not guarantee an accurate view of the whole picture. For example, in studying intrinsic cues and extrinsic cues in isolation, we could have overlooked the relationship between the two when they function together in talker normalization. Where intrinsic cues fall short of fully accommodating talker variation, extrinsic cues might be especially useful. In order to fully understand talker normalization, it is thus important to study it from a more integrative perspective.

Secondly, many previous studies on extrinsic normalization focused on whether speech and nonspeech contexts elicited similarly contrastive effects. Though this question is important, more emphasis should be placed on finding a mechanism that *sufficiently* enables listeners to accurately map speech signals with talker variation to linguistic categories. By ‘sufficient’, we mean a consistent context effect in all conditions. Given that null effects of nonspeech contexts have been reported in some studies (cf. Chen & Peng, 2015; Francis et al., 2006; Zhang et al., 2012; 2013), it might be the case that the general contrast of auditory cues between the context and the target speech sound is *not* sufficient for accommodating talker variation. Whether the estimation of a particular talker's acoustic-phonological space is sufficient requires further investigation, especially via a comparison of two contexts with and without cues to a talker's acoustic-phonological space.

Thirdly, previous models assume that the same normalization mechanism is used in the perception of speech sounds produced by different talkers, without considering the typicality of a talker's voice characteristics. Bishop and Keating (2012) found that an F0 stimulus was rated as having higher pitch when it was from a male speaker than a female speaker. The authors speculated that English listeners have expectations of F0 ranges for general male and female talkers respectively, which affected the pitch rating. Within each gender, it is conceivable that some talkers are more typical with F0 range closer to that of general male/female talkers than other talkers. This raises the question of how the typicality of a talker's F0 within each gender affects speech perception, and most importantly, whether listeners would adopt different mechanisms in accommodating speech signals of typical and atypical talkers. However, the effect of talker typicality has not been investigated before.

In this study we aim to address the aforementioned issues and propose a new model to integrate different cues. In Experiment 1, we examined intrinsic normalization, i.e., whether Cantonese listeners can accurately identify isolated words carrying Cantonese level tones based on intrinsic cues only. We also examined whether the identification accuracy is correlated with the typicality of a talker's F0 range in terms of its distance from an estimated population F0 range, such that level tones produced by typical talkers would be more likely to be correctly identified. In Experiment 2, we examined extrinsic normalization, by comparing four types of context with incrementally more cues (nonspeech, reversed speech, meaningless speech and meaningful speech) to find those that provide sufficient cues for fully accommodating talker variation. In particular, we included two contexts, a reversed speech context without cues to a talker's acoustic-phonological space, and a normal speech context with cues to a talker's acoustic-phonological space for comparison. Lastly, based on our results, we proposed a new model that integrates various cues to account for talker normalization.

Experiment 1

Methods

Experimental design. As mentioned earlier, it remains unclear whether listeners can estimate an unfamiliar talker's pitch range based on intrinsic cues only (Bishop and Keating, 2012; Honorof and Whalen; 2005). Moreover, this question has not been tested on speakers of a tone language, who have to estimate a particular talker's F0 range in order to judge the location/identity of a tone. In this experiment, we aim to examine whether Cantonese listeners can accurately identify *isolated words* carrying Cantonese mid level tone produced by unfamiliar talkers. If listeners can estimate the F0 location within an unfamiliar talker's F0 range with only intrinsic cues (cf. Honorof and Whalen, 2005), we expect the word with mid level tone to be accurately recognized. We also aim to examine whether the identification performance can be explained by the typicality of a talker's F0 range relative to the population F0 range, which was estimated from a large Cantonese speech corpus (Lee, Lo, Ching, & Meng, 2002). To this end, we selected four native Cantonese speakers (2 female, 2 male) with large differences in their F0 ranges, covering a wide range of F0 distribution.

Bishop and Keating (2012) speculated that English listeners may have expectations of gender-specific F0 ranges, e.g., female speakers with a higher F0 range than male speakers, which affect the pitch location rating. However, the effect of gender-specific F0 ranges has not been examined in the correlation with the identification performance. It is also possible that listeners might rely on a non-gender-specific F0 range. In this experiment, we also examine whether the perceptual performance is better explained by the typicality of a talker's F0 range relative to an estimated *gender-specific* population F0 range, or *non-gender-specific* population F0 range.

Participants. Sixteen native speakers of Cantonese (8 female, 8 male; mean age = 22.6 yr, SD = 2.6 yr) participated in this experiment. No subjects had hearing impairment or music training. The subjects were all students at the Chinese University of Hong Kong. All subjects gave informed consent in compliance with a protocol approved by the Survey and Behavioral Research Ethics Committee of The Chinese University of Hong Kong.

Stimuli. Four native Cantonese speakers with different F0 ranges (2 male speakers: M01 and M02; 2 female speakers: F01 and F02) were recruited to record the stimuli. Each of the four talkers was asked to read aloud six words carrying six Cantonese unchecked tones and repeat each word six times. A speaker's F0 range was estimated from the production of two words, 醫 (/ji55/ 'a doctor'), which carries the highest tone in Cantonese, and 兒 (/ji21/ 'a son'), which carries the lowest tone. The upper F0 range was measured from the maximal F0 of six repetitions of 醫 /ji55/, and the lowest tone was obtained from the minimal F0 of six repetitions of 兒 /ji21/[2]. Figure 1A shows the F0 range of the four talkers (F01: 212~331 Hz; F02: 198~280 Hz; M01: 128~190 Hz; M02: 91~122 Hz).

The word with mid level tone – 意 (/ji33/ 'meaning') – was used as the sole listening materials. The reason for using only one word with mid level tone, instead of all words with six tones, is to prevent listeners from estimating a talker's F0 range from extrinsic cues, i.e., the presentation of high and low tones from the same talker. The word with mid level tone

[2] In order to avoid extreme F0 values that might be outliers, we chose the F0 value located at the 90% highest and lowest end of all F0 measurements of six repetitions of /ji55/ and /ji21/ respectively.

was selected, because mid level tone is most ambiguous, which can be misidentified as either high level tone or low level tone depending on talker variation (Peng et al., 2012; Zhang et al., 2012). For each talker, one clear token of the word with mid level tone was selected. The target word produced by four talkers was normalized in duration to 450 ms and in average intensity level to 60 dB using Praat (Boersma & Weenink, 2012). The F0, F1, and F2 of the test words are shown in Table 1.

Filler items were included to maintain listeners' interest in the speech stimuli. We ensured that words with the highest and lowest tone were excluded in order to minimize the information of a talker's F0 range revealed by filler words. One filler item was a different repetition of the target word (意 /ji33/ 'meaning') and the other filler item was the syllable /ji/ carrying low level tone (二 /ji22/ 'second'). Each filler item was produced by one talker in each gender. The four filler items were normalized in duration and intensity accordingly. For the filler item with mid level tone, its mean F0 was very similar to that of the target word (only 1.5~1.7 Hz difference); as for the filler item with low level tone, its mean F0 was lower than the target word by 1.5~3.3 semitones. The ratio of test words and filler items was 1:1.

The population F0 range was estimated for male and female Cantonese talkers separately from a large-scale speech corpus, which contains read speech materials from 68 native Cantonese speakers, with half of the speakers in each gender (Lee et al., 2002). The upper and lower F0 range was measured from the average F0 of words carrying the highest tone and lowest tone produced by all female and male speakers respectively. Only sentence-initial words were considered in order to minimize the effect of intonation. The estimated mean F0 range of female talkers is approximately 220~290 Hz, and male talkers is approximately 110~160 Hz. We also estimated a non-gender-specific population F0 range (165~225 Hz), by averaging the estimated mean F0 range of female and male talkers. We then calculate the distance of the four talker's F0 range from the gender-specific and non-gender-specific population range in semitones respectively.

Procedure. Target and filler words produced by four talkers were mixed and randomly presented in a sub-block and each sub-block was repeated seven times. The task was three-alternative forced choice identification. Subjects were instructed to identify the heard word (target or filler) as any of the three Cantonese words, 醫 (/ji55/ 'a doctor'), 意 (/ji33/ 'meaning'), and 二 (/ji22/ 'second') by pressing labeled buttons on a computer keyboard as soon as possible.

Data analysis. The results were analyzed in terms of identification accuracy and perceptual height scores. The reason that we did not use percent phoneme identification analysis, as many previous studies did (e.g., Holt, 2005, 2006a, 2006b; Huang & Holt, 2009; 2012), is because our task is three-alternative forced choice, which is less straightforward to analyze using percent phoneme identification. For identification accuracy, we analyzed the percent of correct responses (i.e., mid level tone responses) to a target word produced by each talker. The perceptual height analysis was to code each response (correct and incorrect) according to the perceptual height of the selected tone, and calculate the mean height score of all responses to the target word produced by each talker. A high level tone response was coded as '6', a mid level tone response as '3', and a low level tone response as '1', for the reason that the high level tone is 3 semitones higher than the mid level tone, which is again 2 semitones higher than the low level tone, according to previous descriptions (Chao, 1947). If the average perceptual height score was close to '1', it indicates that the target word was primarily identified as having the low level tone; if it was close to '6', it indicates that the target word was primarily identified as having the high level tone.

Results

Figure 2A shows the identification accuracy and Figure 2B-C show the perceptual height scores plotted as a function of the distance of each talker's F0 range relative to the estimated gender-specific population F0 range.

We first tested whether the identification accuracy was better than chance (0.33) for the target word produced by all four talkers. If subjects can estimate an unfamiliar talker's F0 range based on intrinsic cues only, they are expected to be able to correctly identify the target word above chance across all four talkers, irrespective of talker typicality in F0 range. The data were analyzed using a generalized mixed-effects model for each talker respectively, with the responses to each trial (correct and incorrect) as the input and subjects as a random effect. Results show that only the target word produced by the female talker F02 was accurately recognized above chance (mean = 0.62, SD = 0.26, z = 4.0, $p < 0.001$). It indicates that intrinsic cues alone may not be enough for listeners to fully accommodate talker variation in the perception of Cantonese mid level tone.

We then examined whether the perceptual performance can be explained by the typicality of a talker's F0 range relative to the estimated population range. Two linear regression analyses were carried out on the perceptual height scores, one analysis on *gender-specific* population range and the other analysis on *non-gender-specific* population range. The purpose is to examine to what extent the target word's likelihood of being misidentified as having *high level tone* or *low level tone* is related to a talker's F0 range being *higher* or *lower* than the gender-specific or non-gender-specific population range. For the linear regression analysis on *gender-specific* population range, the dependent variable was perceptual height scores and two independent variables were the distance of a talker's upper and lower F0 range relative to the gender-specific population range (two distance values per talker). The linear regression model reached significance and accounted for 68.3% of the variance in perceptual height scores (adjusted $R^2 = 0.672$, $p < 0.001$). Typicality of a talker's lower F0 range contributed most strongly to the perceptual heights scores (t = 3.853, $p < 0.001$), while typicality of a talker's upper F0 range was also significant but the effect was less strong (t = 2.039, $p = 0.045$). For the linear regression analysis on the *non-gender-specific* population range, the model also reached significance, but accounted for overall less variance in perceptual height scores than gender-specific population range (adjusted $R^2 = 0.549$, $p < 0.001$). Typicality of both lower and upper F0 ranges in terms of non-gender-specific population range contributed significantly to the perceptual heights scores (*p*s < 0.001).

When subjects were included as a random effect, the mixed effects model with gender-specific population range also outperformed the model with non-gender-specific population range, according to the model selection criterion – Akaike Information Criterion (AIC; gender-specific model AIC = 197.39; non-gender-specific model AIC = 219.32; the smaller the AIC, the better a model). Again, the model with gender-specific population range accounted for more variance than the model with non-gender-specific population range. Following the methods described by Nakagawa and Schielzeth (2013), we calculated marginal R^2 that describes the proportion of variance explained by fixed effects, and conditional R^2 that describes the proportion of variance explained by fixed and random effects. The model on gender-specific population range accounted for 67.5% of variance with fixed effects considered and 72.9% of variance with both fixed and random effects considered, whereas the model on non-gender-specific population range accounted for 55.6% of variance with fixed effects considered and 56.9% of variance with fixed and random effects considered. This indicates that the perceptual height scores of the target word were most heavily influenced by talker typicality in terms of *gender-specific* population range. The higher a talker's F0 range relative to gender-specific population range, the higher the perceptual height scores, meaning that the target word is more likely to be identified as having high level tone, and vice versa.

Discussion

In this experiment, we examined whether Cantonese listeners can accurately identify isolated words carrying Cantonese mid level tone produced by four unfamiliar talkers with different F0 ranges. If listeners can estimate the F0 location within an unfamiliar talker's F0 range with intrinsic cues alone (cf. Honorof and Whalen, 2005), the level tone produced by all four talkers is expected to be identified accurately. On the contrary, we found that only the target word produced by one talker was accurately identified above chance. This indicates that intrinsic cues *alone* may not be enough for fully accommodating talker variation, a finding consistent with the finding of Bishop and Keating (2012). Moreover, we compared the effects of *gender-specific* and *non-gender-specific* population F0 range and found that the gender-specific population F0 range contributes more to the perception of Cantonese level tones. Although the model with non-gender-specific population range also reached significance, it accounted for less variance of perceptual height scores. It confirms that the perception of isolated words carrying Cantonese level tone is primarily influenced by talker typicality relative to the *gender-specific* population F0 range.

We found that target words produced by typical talkers were more likely to be correctly recognized, whereas target words produced by less typical talkers were biased to the perception of neighboring high/low level tones, depending on whether the talker's F0 range is higher or lower than the population F0 range. This finding may imply that Cantonese listeners have built-in tone templates that are shaped by typical talkers' F0 range. The target words produced by typical talkers presumably match the built-in tone templates and were thus correctly recognized, whereas the target words produced by less typical talkers deviated from the tone templates and were thus misrecognized. It appears that the tone templates are gender-specific, with the tone templates for general female talkers carrying higher F0 than those for general male talkers (cf. Bishop and Keating, 2012). It is not totally clear how such tone templates are built and stored. But it is likely that the templates are built from the auditory experience of verbally interacting with many female and male talkers in a speech community. The F0 of typical female and male talkers close to the center of the population F0 distribution, who probably occur most frequently, determines the F0 of the tone templates. This indicates that the representation of lexical tones is not totally abstract or deprived of phonetic details, but at least preserves the F0 information of typical female and male talkers.

We found that the lower bound of gender-specific F0 range is a better predictor of perceptual height scores than the upper bound. This result is intriguing, because the between-talker variability is usually greater at the upper bound than at the lower bound (Bishop & Keating, 2012; Keating & Kuo, 2012), meaning that the upper bound might be more indicative of between-talker differences. However, the upper bound also varies more within a talker. For example, a talker's upper F0 range can be inflated by factors such as emotional status (e.g., Protopapas & Lieberman, 1997). Therefore, the lower F0 range might be a more reliable indicator of a talker's speaking F0.

This experiment is a first attempt to examine the effect of talker typicality on speech perception and leaves open a few questions for future investigation. Firstly, only four talkers were examined in this experiment. Future studies should include more talkers covering a broader range of the population F0 distribution, to zoom in on the effect of talker typicality. Secondly, this experiment only considered one syllable that contrasts six Cantonese tones. Future studies may increase the phonological coverage with more syllables to further examine the effect of talker typicality.

Experiment 2

Methods

Experimental design. Experiment 1 showed that intrinsic cues *alone* might not be enough for fully accommodating talker variation in the perception of Cantonese level tones. Identification accuracy varied depending on the typicality of a talker's F0 range relative to the gender-specific population F0 range. It thus calls for the investigation of the effect of other cues, especially extrinsic cues, on talker normalization. To this end, we examined extrinsic normalization in Experiment 2, with a context (nonspeech, reversed speech, meaningless speech and meaningful speech) × F0 shift (raised, unshifted and lowered) × talker (F03, F04, M03 and M04) design.

We compared the effects of four types of contexts with incrementally more cues – nonspeech, reversed speech, meaningless speech and meaningful speech context (see Table 2), to find a context sufficient for accommodating talker variation in perception of Cantonese level tones. By 'sufficient', we mean that a context elicits above-chance speech identification accuracy across all talkers, irrespective of the typicality of a talker's F0 range. The nonspeech context contained only auditory cues of pitch for general auditory contrast between the context and a target word (cf. Huang & Holt, 2009; Holt et al., 1996). The reversed speech context contained some phonetic cues such as vowels, apart from auditory cues, but the phonetic cues can hardly be associated with native consonants, vowels, lexical tones or syllables in Cantonese. Thus this context likely allows listeners to estimate the range of general phonetic variation of a talker, but such phonetic variation is not associated with phonological cues in Cantonese. The meaningless and meaningful contexts, which were composed of native Cantonese syllables, contained cues to estimate a talker's acoustic-phonological space. The meaningful context further contained valid and coherent semantic and syntactic information. These four contexts were tested in two sub-experiments (see Table 2), in order to control the experiment length. Nonspeech, reversed speech and meaningful speech contexts were tested in Experiment 2A, and nonspeech, meaningless and meaningful speech contexts were tested in Experiment 2B.

Participants. Sixteen native speakers of Cantonese (9F, 7M; age = 21±1.2 yr) participated in Experiment 2A and another group of 16 native speakers of Cantonese (8F, 8M; age = 21.3 ± 2.0 yr) participated in Experiment 2B with monetary compensation. All subjects reported normal hearing and no musical training. The subjects were students at the Chinese University of Hong Kong. Informed written consent was obtained in compliance with the experiment protocol approved by the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee.

Stimuli. Stimuli were recorded from four native Cantonese speakers with different F0 ranges (2 male speakers: M03 and M04; 2 female speakers: F03 and F04), who were different from the talkers of Experiment 1. The F0 range of these four talkers is shown in Figure 1B. From each talker, a meaningful sentence, 呢個字係意 /li55 ko33 tsi22 hɐi22 ji33/ (‘This character is meaning’) and a meaningless sentence, 呢錯視幣意 /li55 tsʰo33 si22 pɐi22 ji33/ (‘This mistake sees money meaning’) were recorded. The meaningful sentence was semantically neutral, to minimize the effect of semantic expectation on the identity of the target word. The meaningless sentence was a semantically meaningless combination of real morphemes in Cantonese. In both sentences, the final word carrying mid level tone was the target word. Meaningful and meaningless contexts were matched in rhymes and tones.

Each talker was asked to read aloud these sentences six times. One typical token of the target word close to the average F0 of all six repetitions was selected for each talker. Each target word was normalized in duration to 450 ms and in average intensity level to 55 dB using Praat. The F0, F1 and F2 of the target word are shown in Table 3.

One clearly produced meaningful context and meaningless context roughly matched in the mean, minimal and maximal of F0 were selected for each talker. The average intensity level of all the contexts was normalized to 55 dB, identical to the target word. Following the design of previous studies (Zhang et al., 2012; 2013), the overall F0 trajectory of each context was lowered by 3 semitones, kept unshifted and raised by 3 semitones, giving rise to three contextual F0 heights. The F0 manipulation has very little, if any, influence on the formant trajectories of the context. Acoustic characteristics of the meaningful speech context with the three F0 heights are shown in Table 3.

In addition, four filler sentences were included. The reason for including filler sentences is to increase context variability so that listeners would not lose interest in the contexts and only attend to the target word. Unless it is known that the extrinsic normalization process is automatic and mandatory even without attention, such a control seems necessary. A meaningful context, 請留心聽_ /ʦʰiŋ25 ləu21 sɐm55 tʰiŋ55 _/ 'Please carefully listen to _' was recorded from two of the above four talkers, and a second context, 我以家讀_ /ŋo23 ji21 ka55 tuk2 _/ 'Now I will read _' from the other two talkers. Two meaningless contexts 頂留金青_ /tiŋ25 ləu21 kɐm55 ʦiŋ55 _/ 'top maintains gold green _' and 我時花俗_ /ŋo23 si21 fa55 ʦuk2 _/ 'I time flower convention _' were recorded accordingly. The target words were syllable /ji/ with either mid level tone or low level tone produced by the same four talkers. One clearly produced meaningless and meaningful filler sentence was also selected. The average intensity level of filler contexts was normalized to 55 dB, similar to the test contexts. The F0 of filler contexts was not manipulated in order to keep the experiment short. The ratio of test and filler sentences was 3:1.

Reversed speech contexts were then generated by time-reversing the meaningful speech contexts with the three F0 heights – lowered F0, unshifted F0 and raised F0 using Praat. Reversed speech contexts were also generated for filler meaningful contexts.

Nonspeech contexts were generated using a triangle wave that has a different harmonic structure but roughly similar acoustic complexities (i.e., containing only odd number of harmonics) as speech sounds. The triangle wave context modeled the F0 and intensity profiles of meaningful speech contexts with three F0 heights. The average intensity level of the nonspeech contexts was set to be 75 dB, higher than the speech contexts, because nonspeech contexts sound softer than speech contexts (cf. Zhang et al., 2012; 2013). Nonspeech contexts were also generated for filler contexts.

Lastly, the target word was attached to the end of four types of contexts after a jittered interval of 300-500 ms for each talker. The F0 contour of four types of contexts and the target word from one talker are shown as an example in Figure 3.

Procedure. Stimulus presentation was blocked by the context, with each block comprising only trials from one context condition. Within a block, 16 trials ((3 test sentences + 1 filler sentence) × 4 talkers) were presented randomly in a sub-block and each sub-block was repeated nine times. Subjects were instructed to identify the target word (test or filler) as any of the three Cantonese words, 醫 (/ji55/ 'doctor'), 意 (/ji33/ 'meaning'), and 二 (/ji22/ 'two') by pressing labeled buttons on a computer keyboard within two seconds.

The way that trials with three contextual F0 heights from four different talkers were mixed and randomly presented in a sub-block means that listeners have to re-normalize talker variation from trial to trial. To normalize between-talker variation (i.e., four talkers) and within-talker variation (i.e., thee contextual F0 heights from the same talker), listeners have to obtain a most updated estimate of a talker's F0 range from an immediate context. Including filler sentences added more variation into the stimuli, which perhaps also drives listeners to re-normalize from trial to trial.

In Experiment 2A, subjects listened to nonspeech, reversed speech and meaningful speech contexts. In Experiment 2B, subjects listened to nonspeech, meaningless speech and meaningful speech contexts. The presentation order of three blocks was counterbalanced across the subjects as much as possible. Before each experiment, one practice block with meaningful speech contexts from two extra talkers not used in the experiment was given to the subjects to familiarize them with the procedure.

Data analysis. The results were analyzed in terms of identification accuracy and perceptual height scores. As mentioned before, we did not use percent phoneme identification analysis as previous studies did (e.g., Holt, 2005, 2006a, 2006b; Huang & Holt, 2009; 2012), because our task is three-alternative forced choice instead of two-alternative forced choice, which is less straightforward to analyze using percent phoneme identification. For identification accuracy, the rate of expected responses was analyzed for each talker, each context F0 height, and each type of context. Given the contrastive context effect (e.g., Ladefoged & Broadbent, 1957; Francis et al., 2006; Zhang et al., 2012, 2013), the target word was expected to be identified as having high level tone (/ji55/, 'doctor') in the lowered F0 condition, as having mid level tone (/ji33/, 'meaning') in the unshifted F0 condition, and as having low level tone (/ji22/, 'two') in the raised F0 condition. A response was deemed as correct if it was the expected response and incorrect if not. The perceptual height analysis was to code each response (correct and incorrect) according to the perceptual height of the selected tone (6 for high level tone, 3 for mid level tone, and 1 for low level tone), and calculate the mean height score of all responses to each target word. If the average perceptual height score was close to ‘1’, it indicates that the target word was primarily identified as having the low level tone; if it was close to ‘6’, it indicates that the target word was primarily identified as having the high level tone.

Results

Experiment 2A. As mentioned above, three contexts were examined in Experiment 2A, nonspeech, reversed speech and meaningful speech contexts. Figure 4A shows the identification accuracy for each talker condition within each F0 shift and context condition.

We tested whether the identification accuracy was better than chance (0.33) across all talkers and all F0 shift conditions for each type of context. If a certain type of context facilitates talker normalization, the identification accuracy is expected to be significantly higher than the chance level (0.33) consistently across all talkers and all F0 shift conditions, irrespective of the typicality of a talker and the manipulation of F0 shift. To that end, the identification accuracy was analyzed using a generalized mixed-effects model for each talker (within each F0 shift and context condition), with the response to each trial (correct and incorrect) as the input and subjects as a random effect. The model compared the rate of correct responses with chance level accuracy.

For the nonspeech context, the identification accuracy was only significantly higher than chance level in two conditions: female talker F03 in the unshifted F0 condition (mean = 0.66, SD = 0.48, z = 3.20, p = 0.001), and female talker F04 (mean = 0.78, SD = 0.42, z = 6.65, p < 0.001) in the raised F0 condition. As for the reversed speech context, more conditions reached significance, but still not all conditions were significant. The identification accuracy was significantly higher than chance in six conditions: female talker F03 (mean = 0.72, SD = 0.45, z = 2.58, p = 0.01) and male talker M04 (mean = 0.53, SD = 0.50, z = 2.21, p = 0.027) in the unshifted F0 condition, and all four talkers F03 (mean = 0.75, SD = 0.43, z = 6.11, *p* < 0.001), F04 (mean = 0.84, SD = 0.37, z = 2.69, *p* = 0.007), M03 (mean = 0.82, SD = 0.39, z = 3.69, *p* < 0.001) and M04 (mean = 0.79, SD = 0.41, z = 4.84, *p* < 0.001) in the raised F0 condition. It is only in the meaningful speech context that the identification accuracy was significantly higher than the chance level in all conditions (*p*s < 0.05). It indicates that only the meaningful speech contexts consistently facilitate talker normalization, irrespective of talker typicality and F0 shift manipulations.

Experiment 2B. As mentioned above, nonspeech, meaningless speech and meaningful speech contexts were examined in Experiment 2B. Figure 4B shows the identification accuracy for each talker, F0 shift and context condition.

We tested whether the identification accuracy was better than chance (0.33) across all talkers and all F0 shift conditions for each type of context. Similar to Experiment 2A, the responses were analyzed using a mixed-effects logistic regression model for each talker (within each F0 shift and context condition), with the response to each trial (correct and incorrect) as the input and subjects as a random effect.

For the nonspeech context, the identification accuracy were significantly higher than the chance level only in three conditions: female talker F03 and male talker M04 in the unshifted F0 condition (mean = 0.65, SD = 0.48, z = 4.68, *p* < 0.001; mean = 0.54, SD = 0.50, z = 3.01, *p* = 0.003), and female talker F04 in the raised F0 condition (mean = 0.67, SD = 0.47, z = 3.89, *p* < 0.001). As for the meaningless and meaningful speech contexts, the identification accuracy was significantly higher than the chance level in all conditions (*p*s < 0.01). It suggests that the nonspeech context and two speech contexts have unequal effects on the perception of Cantonese level tones. Whereas the two speech contexts consistently facilitate talker normalization irrespective of talker typicality and F0 shift manipulations, the effect of nonspeech context is not consistent on talker normalization.

Experiment 2A and 2B combined. Since the results of Experiment 2A and 2B were largely similar (in terms of the effects of nonspeech and meaningful speech contexts), the two experiments were grouped together for further analysis.

Firstly, the overall effects of four types of contexts were compared, using a mixed-effects logistic regression model with the response to each trial (correct and incorrect) as the input, context type (nonspeech, reversed speech, meaningless speech and meaningful speech), F0 shift (lowered F0, unshifted F0 and raised F0) and talker (F03, F04, M03, M04) as three fixed effects and subjects as a random effect. Two-way and three-way interactions were also included in the full model. Although both nonspeech and reversed speech contexts failed to elicit significantly higher-than-chance accuracy across all conditions, it appears that the overall effect of the reversed speech context was stronger than that of the nonspeech context (see Figure 4A). Moreover, although both meaningless and meaningful speech contexts elicited significantly higher-than-chance accuracy across all conditions, the overall effect of the meaningful context might be stronger than that of the meaningless context (see Figure 4B).

All the main and interaction effects are significant and reported in Table 4. Here we focused on reporting the comparison of four types of context. The main effect of context type reached significance by likelihood ratio tests ($\chi^2(3) = 2515.6$, $p < 0.001$; nonspeech: mean = 0.337, SD = 0.35; reversed speech: mean = 0.473, SD = 0.398; meaningless speech: mean = 0.771, SD = 0.263; meaningful speech: mean = 0.889, SD = 0.187). Pairwise comparisons show that the reversed, meaningless and meaningful speech contexts elicited significantly higher identification accuracy than the nonspeech context (*p*s < 0.01). The meaningless and meaningful speech contexts elicited significantly higher identification accuracy than the reversed speech context (*p*s < 0.001). No significant difference was found between meaningless and meaningful speech contexts (z = 0.071, p = 0.483). However, the difference between meaningful and meaningless speech contexts was actually significant (z = -10.36, p < 0.001) when the interaction effects were not included in the full model. This indicates that there might be significant differences between these two contexts, but this effect is subject to the influence of model specifications.

Secondly, a linear regression model was conducted to examine whether the perceptual performance in each context was correlated with talker typicality. Since the nonspeech and reversed speech context failed to elicit higher-than-chance accuracy in all conditions, it is worth examining whether the perceptual performance was actually influenced by the typicality of a talker's F0 range.

Similar to Experiment 1, a linear regression model was fit to the perceptual height scores of each context, with the distance of a talker's lower and upper F0 range from the gender-specific population F0 range as two predictors. The purpose is to examine to what extent the target word's likelihood of being identified as having high level tone or low level tone is related to the typicality of a talker's F0 range in a context. Only the model on the nonspeech context reached significance, which accounted for 9.6% of variance in the perceptual height scores (adjusted R^2 = 0.097, p < 0.001). If subjects were included as a random effect, again, only the linear mixed effects model on the nonspeech context was significant: the typicality of a talker's lower F0 range contributed significantly to perceptual heights scores in the nonspeech context (t(360.8)=4.475, p < 0.001), whereas the typicality of a talker's upper F0 range was not significant (t(360.8)=-1.089, p = 0.277). The typicality of a talker's lower and upper F0 range together accounted for 9.87% of variance with fixed effects considered and accounted for 29.7% of variance with both fixed and random effects considered (Nakagawa & Schielzeth, 2013). This indicates that perceptual performance in the nonspeech context was significantly influenced by talker typicality in F0 range.

To summarize, the magnitude of effects of the four contexts fell in this order: nonspeech context < reversed speech context < meaningless speech context ≤ meaningful speech context. Moreover, only perceptual performance in the nonspeech context was influenced by talker typicality.

Discussion

In Experiment 2, we compared the effects of four types of contexts – nonspeech, reversed speech, meaningless and meaningful speech contexts, in order to find a mechanism sufficient for accommodating talker variation in the perception of Cantonese level tones. We discuss the four types of contexts one by one in the text below.

As mentioned before, the effect of nonspeech context is primarily mediated by a general auditory mechanism, affecting speech perception via the general contrast of auditory cues across the context and target word (Huang & Holt, 2009; 2012; Holt et al., 1996; Holt, 2005, 2006a, 2006b; Lotto & Kluender, 1998). We found that the nonspeech context failed to elicit significantly higher-than-chance accuracy in a number of conditions. Moreover, the perceptual performance was influenced by talker typicality in F0 range, despite the contextual F0 cues. This indicates that general contrast of auditory cues may not be sufficient for talker normalization in the perception of Cantonese level tones, which is consistent with the findings of previous studies (Chen & Peng, 2015; Francis et al., 2006; Zhang et al., 2012; 2013; Sjerps et al., 2012).

Within speech contexts, contexts with different cues could affect speech perception via different mechanisms. In particular, not all speech contexts necessarily contain cues to estimate a talker's acoustic-phonological space. The reversed speech context, for example, only contained cues of a talker's phonetic variation. One may argue that some phonological information could be gleaned from the reversed speech context. For example, listeners might be able to identify the monophthongs and level tones even when they were time-reversed, especially since the original context was primarily composed of words with monophthongs and level tones (呢個字係 /li55 ko33 tsi22 hɐi22/). But it should be noted that the reversed context deviates from Cantonese phonology in the following ways, which probably hinders the extraction of phonological information. Firstly, reversing the context creates a foreign syllable structure that differs from Cantonese (/ia his (n)o kin/). Note that consonant coda /s/ is not allowed in Cantonese syllables. Secondly, although level tones are phonologically described as 'level', they are often phonetically realized with a slightly falling F0 contour. Tonal coarticulation also modifies the phonetic realization of a level tone, such that a level tone could be produced with a rising/falling contour due to F0 transition between neighboring tones (e.g., Xu, 1997). Thus when reversed, it is difficult to identify the level tones as they are. Thirdly, time-reversing a speech utterance creates unnatural formant transitions between consonants that do not occur in natural human production. Given the above three factors, we estimate that it is difficult for the reversed speech context to be mapped to native speech sounds in Cantonese.

We found that the reversed speech context failed to elicit significantly higher-than-chance accuracy across all conditions, which indicates that the range of a talker's phonetic variation is not sufficient for talker normalization in the perception of Cantonese level tones. Although insufficient, the reversed speech context elicited higher identification accuracy than the nonspeech context. Moreover, the perceptual performance was not influenced by talker typicality any more, unlike the nonspeech context, which also indicates an effect of the contextual F0 cues. It appears that the reversed speech context has some effect, though insufficient, on talker normalization (especially in the raised F0 condition where the identification accuracy was significantly higher than chance across all four talkers, see Figure 4A).

One may argue that the reversed speech context is similar to a foreign speech context, in that both contain non-native speech sounds. However, Sjerps and Smiljanic (2013) found that foreign speech contexts have a similar effect on vowel perception as native speech contexts (see Wong, 1998 for a similar finding). Native English, Dutch and Spanish listeners compensated for F1 shift in the contexts in a contrastive way, irrespective of whether the context (English, Dutch and Spanish) was in their native language. Overall, shifting the F1 in a native/non-native context appears to change the identification rate by approximately 5~20% across all context types and listener groups (see Figure 6 of Sjerps and Smiljanic, 2013).

Why did foreign speech context affect vowel perception, whereas the reversed speech context failed to affect tone perception consistently? We believe several technological differences could have contributed to this discrepancy. Firstly, we compared the identification accuracy of each context against chance-level accuracy, according to the criterion that a context should elicit significantly higher-than-chance accuracy consistently across all talkers and all F0 shift conditions. It turns out that the reversed speech context only elicited significantly higher-than-chance accuracy in some conditions. But if we look at the overall identification accuracy in the reversed speech context (mean = 0.473, SD = 0.398), the magnitude of context effect is approximately 14% higher than the chance level accuracy, which actually overlaps with the amount of context effect (5~20%) reported by Sjerps and Smiljanic (2013). This means that the overall effect of the reversed speech context and foreign speech context might be largely similar. Secondly, as mentioned before, the reversed speech context contained unnatural acoustic properties that do not occur in human production such as time-reversed formant transitions, whereas a foreign speech context obeyed the constraints of natural speech production. Such unnatural acoustic properties could have impeded the context effect to some extent.

On the other hand, meaningless and meaningful contexts contained real morphemes with native consonants, vowels and lexical tones in Cantonese, which allowed listeners to estimate a talker's acoustic-phonological space. In addition, the meaningful context contained valid semantic and syntactic cues. We found that both meaningless and meaningful contexts elicited significantly higher-than-chance accuracy across all conditions, which indicates that a talker's acoustic-phonological space is sufficient for talker normalization in perception of Cantonese level tones. Moreover, the effect of meaningful speech context was even stronger than that of the meaningless speech context (the difference reached significance when interaction terms were not included in the model). This is probably because valid and coherent semantic and syntactic cues help to disambiguate some potentially ambiguous words, enabling listeners to obtain a more accurate estimation of a talker's acoustic-phonological space. For example, the word 字 /tsi22/ 'character' could be confused with words with high level tone (e.g., 知 /tsi55/ 'know') or mid level tone (e.g., 志 /tsi33/ 'aspiration'). But within the context 呢個字係 /li55 ko33 tsi22 hɐi22/ 'This character is', the interpretation of this sound is constrained by semantic and syntactical cues to the word 字 /tsi22/, which helps listeners to accurately identify it as having low level tone, and to estimate the lower bound of a talker's F0 range. This indicates that valid semantic and syntactic cues, though not necessary, can further facilitate talker normalization.

General discussion: A new model of talker normalization

Here we propose a first formulation of a new model that integrates different factors to explain the processes of talker normalization, based on our experimental results and other studies in the literature. Although we focused on lexical tones in the experiments, we believe that this model may apply to speech sounds in general. Figure 5 shows the model.

As mentioned earlier, previous models of talker normalization did not consider the effect of talker typicality. Another factor that has been ignored by previous models is talker familiarity, probably because most models focus on unfamiliar talkers. It has been found that speech perception is sensitive to talker familiarity (Nygaard & Pisoni, 1998; von Kriegstein & Giraud, 2004). Nygaard and Pisoni (1998) found that listeners showed better performance in the identification of isolated words and words in sentences, if the words were produced by familiar talkers whose voices were presented to the listeners in the training phase. This indicates that talker-specific vocal characteristics, once learned, could be stored in memory to assist speech recognition later on. Craik and Kirsner (1974) found that listeners were more accurate at detecting whether an auditory word was a repeated word or a new word, when the 'repeated' words were produced by the same talker than when the words switched to a new talker. This indicates that exemplars of words stored in memory preserved a talker's voice information implicitly. Altogether, these findings suggest that a learned talker's voice characteristics can be stored in memory (e.g., Craik & Kirsner, 1974; Goldinger, 1991; 1996; 1997; 1998; Hintzman et al., 1972; Johnson, 1997; 2007; Palmeri et al., 1993).

Based on the above discussion, we speculate that a space of learned talker models, i.e., voice characteristics of previously encountered talkers and maybe exemplars of speech sounds from those talkers, is stored in the memory of listeners. When a speech signal enters the auditory system, spectro-temporal characteristics of the signal are processed to extract the linguistic information (see the first arrow in Figure 5). Acoustic parameters carrying talker information such as F0, F3 and voice quality will be analyzed (Green, Tomiak, & Kuhl, 1997; Kaganovich, Francis, & Melara, 2006; Magnuson & Nusbaum, 2007; Mullennix &

Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Nusbaum & Magnuson, 1997; Nusbaum & Morin, 1992; Wong & Diehl, 2003; Wong, Nusbaum, & Small, 2004; Zhang, Pugh, Mencl, Molfese, Frost, Magnuson, Peng, & Wang, 2016). The analyzed talker characteristics are likely checked against learned talker models stored in memory to see if the characteristics match any familiar talkers. If the analyzed talker characteristics of the incoming speech signal match that of a familiar talker, whose acoustical-to-phonological mapping is already learned, the speech signal probably can be recognized accurately (Nygaard & Pisoni, 1998; von Kriegstein & Giraud, 2004).

But if the voice characteristics do not match any existing talker model, in which case the talker is unfamiliar, the listeners likely resort to the mental templates of speech sounds that are shaped by gender-specific population voice distribution, according to our results of Experiment 1. If an unfamiliar talker turns out to be a typical talker, the speech sound produced by that talker would match the mental templates of speech sounds, and thereby be recognized accurately. However, if an unfamiliar talker turns out to be atypical, whose production deviates from the mental templates, the speech signal would be incorrectly recognized. The less typical a talker's F0, the more likely that talker's speech signal would be misrecognized.

In the case of unfamiliar and atypical talkers, the speech signal would be misidentified temporarily, but it probably can be corrected when more information of a talker's acoustic-phonological space becomes available. For example, our findings show that a speech signal with the mid level tone can be misrecognized as carrying the high level tone, if its F0 is higher than the mental template of mid level tone. But the misrecognized speech signal can be re-evaluated, if the subsequent speech signal carries even higher F0. According to previous studies, it is preferable for a speech context to cover extremes of a phonological space (e.g., /a/, /i/, /u/ of a vowel space, or high and low tone of a tonal space), in order for listeners to accurately estimate a talker's phonological space (e.g., Wong & Diehl, 2003; Joos, 1948; Sjerps et al., 2011a, 2011b, 2012; Zhang et al., 2012). In Experiment 2, the meaningless and meaningful speech contexts (呢個字係 /li55 ko33 tsi22 hɐi22/; 呢錯視幣 /li55 tsʰo33 si22 pɐi22/) contained words with high level tone (/55/) and low level tone (22). It is likely that listeners may have achieved a quite accurate estimation of a talker's tonal space when they reached the third syllable of the context. We found that both meaningless and meaningful speech contexts are sufficient for talker normalization, but the meaningful speech context elicited even higher identification accuracy. This indicates that the dynamic estimation of a talker's acoustic-phonological space can be done purely by the comparison of phonological cues in neighboring speech signals, and semantic information is not necessary. But the presence of coherent semantic and syntactic content must make the estimation of a talker's acoustic-phonological space more accurate, probably by disambiguating ambiguous sounds in the context and correcting any inaccurate estimation of a talker's acoustic-phonological space. The estimated talker-specific acoustic-phonological space then serves as a reference for the recognition of incoming speech signals from the same talker, ensuring accurate perception despite the atypicality of this talker's F0.

For familiar or typical talkers, although the identification accuracy is already high without a context, it should be noted that the dynamic context process might still apply. When the speech signals of a familiar or typical talker match the stored talker models or speech templates, the context process might apply without leading to a change of the categorization of the speech signal. But when a familiar and typical talker speaks in an extreme voice (e.g., in extreme emotional states) (Protopapas & Lieberman, 1997), which sometimes happens, the speech signal would deviate from the stored talker models or speech templates. In that case, the dynamical context process would be important to update the estimation of a talker's acoustic-phonological space in order to ensure accurate perception.

The newly learned talker model (and continually updated via the context process) is probably added to the memory space of learned talker models. It is not totally clear how long a new talker model will stay in memory and how large the memory space is. Some studies indicate that the implicit memory traces of words with a talker's voice characteristics may be retained in memory for up to a day and for at least ten talker's voices. As mentioned earlier, listeners were more accurate at detecting repeated/new words, when the repeated words were produced by the same talker (Craik and Kirsner, 1974). Goldinger (1996) tested various lengths of time lag between the first and second time of word presentation, and found that the same-voice advantage in the recognition of repeated words disappeared if the lag was longer than a day, which indicates that implicit memory of voice information within words might be retained up to a day. Moreover, it is found that the detection accuracy was similar no matter whether subjects listened to word lists produced by two talkers, six talkers or ten talkers, indicating that at least ten talkers' word productions can be retained in memory implicitly. Interestingly, even when a word was repeated by two voices, if these two voices were perceptually similar to each other, it still elicited the same-voice advantage in detection accuracy. This suggests that the same-voice advantage is mediated by the acoustic characteristics of a talker (e.g., F3 and F0) stored in memory, rather than discrete talker labels (e.g., 'John').

According to the above discussion, the acoustic characteristics of at least ten talkers can be implicitly stored in the memory space for about a day. However, this is a very conservative estimate based only on words learned implicitly. The memory space must be larger and can retain talker models for a much longer time, if the speech signals are learned explicitly and with effort (e.g., explicitly associating the speech signals with a specific talker), and if the memory traces are reinforced by repetitive encounter of speech signals from the same talker, as in the case of highly familiar talkers.

We found that the mental templates of speech sounds are shaped by the F0 range of typical male and female talkers. How are such templates built? It is possible that talker-specific exemplars of speech sounds are a prerequisite to build the more abstract speech templates. As the number of talker-specific exemplars increases, cross-talker similarity in the phonetic form likely strengthens whereas talker-specific aspects of the exemplars decay. Thus the speech templates become gradually shaped by the phonetic form of typical male and general female talkers, who presumably occur more frequently in a speech community. It is even possible that each listener may develop individualized tone templates, for the reason that each listener has unique auditory experiences. If so, each listener would also perceive the same speech sounds slightly differently. But if the sample of talker-specific exemplars is large and overlaps across listeners, the speech templates might converge across listeners eventually. Since the subjects examined in this study were all adults living in Hong Kong, it is possible that the speech templates have converged among them. But the question of the formation and convergence of speech templates in individual listeners require further investigation.

Conclusion

To conclude, we presented a first formulation of a model that integrates three major factors – talker familiarity, talker typicality and context, to explain the processes of talker normalization. Such an integrative approach is important in order to fully understand talker normalization. Speech sounds produced by familiar or typical talkers can be recognized accurately without any context. In the case of unfamiliar and atypical talkers, a context with cues to estimate a specific talker's acoustic-phonological space is necessary for talker normalization.

The current model is a first step towards developing a more fully specified model of talker normalization, and will be refined on the basis of future results. It leaves open a couple of questions for future studies. Firstly, it is speculated that a space of learned talker models is stored in the memory of listeners. Future studies should examine this question, and the formation and convergence of speech templates in individual listeners. Secondly, future studies may use a gating paradigm to examine the minimal amount of context needed to accurately estimate a talker's acoustic-phonological space. Previous studies suggest that it is preferable for the context to cover the extremes of the phonological space (e.g., Joos, 1948; Sjerps et al., 2011a, 2011b, 2012; Zhang et al., 2012). But this question warrants further investigation. Thirdly, this model was proposed primarily based on studies on lexical tones. Future studies may investigate to what extent this model applies to the perception of vowels and consonants. Fourthly, future studies may examine the relative contribution of talker familiarity, talker typicality and context to the *probabilistic* mapping of speech signals to linguistic categories using computational modeling (cf. Kleinschmidt & Jaeger, 2015). In this model, we proposed these factors without examining their contribution to probabilistic mapping. Lastly, other cues contributing to talker normalization but not discussed in the current formulation of the model can be considered to be included based on future evidence.

Acknowledgements

This study was supported by grants from the National Natural Science Foundation of China (NSFC: 11504400), the Research Grant Council of Hong Kong (GRF: 14408914), and University's Start-up Fund for New Recruits. We thank the anonymous reviewers for comments that helped us to strengthen the paper. We thank Dr. James Magnuson for constructive comments on the model, and Dr. Steve Politzer-Ahles, Dr. Matthias Sjerps and Dr. Feng Gu for comments on the manuscript. We thank Prof. William S-Y. Wang, Dr. Gang Peng and other members of the Joint Research Centre for Language and Human Complexity for useful discussions, and Ms. Xiao Wang and Mr. Steve Ka-Hong Wong for help with data collection. The raw data were reported at the 21st International Congress on Acoustics, the 18th International Congress of Phonetic Sciences and in Zhang et al. (2013) (i.e., Experiment 2B).

References

Bishop J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *Journal of the Acoustical Society of America, 131*, 1-13.

Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer (Version 5.3.23) [Computer program], http://www.praat.org (Last viewed August 7, 2012).

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

Chao, Y.-R. (1947). *Cantonese Primer*. Harvard University Press.

Chen, F., & Peng, G. (2015). Context effect in the categorical perception of Mandarin tones. *Journal of Signal Processing Systems*, 1–9.

Craik, F., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, 26:274-284.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception and Psychophysics, 68*, 161-177.

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *Journal of the Acoustical Society of America, 119*, 1712-1726.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics, 16*, 73-77.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions of Audio Electroacoustics, AU-16*, 78-80.

Goldinger, S. D. (1991). On the nature of talker variability effects on serial recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 152-162.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166-1183.

Goldinger, S. D. (1997). Word and voices: Perception and production in an episodic lexicon. In Johnson, K., and Mullennix, J. W. (Eds.), *Talker Variability in Speech Processing* (pp. 33-66). San Diego: Academic Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.

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

Hintzman, D. L., Block, R., & Inskeep, N. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behavior, 11*, 741-749.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science, 16*, 305-312.

Holt, L. L. (2006a). Speech categorization in context: Joint effects of nonspeech and speech precursors. *Journal of the Acoustical Society of America, 119*, 4016-4026.

Holt, L. L. (2006b). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America, 120*, 2801-2817.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (1996). Perceptual compensation for vowel undershoot may be explained by general perceptual principles. The 131st Meeting of the Acoustical Society of America, Indianapolis.

Honorof, D. N., & Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *Journal of the Acoustical Society of America, 117*, 2193-2200.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *Journal of the Acoustical Society of America, 125*, 3983-3994.

Huang, J., & Holt, L. L. (2012). Listening for the norm: Adaptive coding in speech categorization. *Frontiers in Psychology, 3*, 10.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America, 88*, 642–54.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson, K., and Mullennix, J. W. (Eds.), *Talker Variability in Speech Processing* (pp 145-166). San Diego: Academic Press.

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. In Sole, M. J., Beddor, P., and Ohala, M., (Eds.), *Experimental Approaches to Phonology: In Honor of John Ohala* (pp. 25-40). Oxford University Press.

Joos, M. (1948). *Acoustic Phonetics*. Baltimore: Linguistic Society of America.

Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research, 1114*(1), 161–172.

Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *Journal of the Acoustical Society of America, 132*, 1050-1060.

Kingston, J, Kawahara, S., Mash, D., & Chambless, D. (2011). Auditory contrast versus compensation for coarticulation: Data from Japanese and English listeners. *Language and Speech, 54*, 499-525.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148-203.

Koenig, L. L. (2000). Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds. *Journal of Speech, Language, and Hearing Research, 43*(5), 1211–1228.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98-104.

Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology, 3*, 1–9.

Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics, 11*, 373-382.

Lee, T., Lo, W. K., Ching, P. C., & Meng, H. (2002). Spoken language resources for Cantonese speech processing. *Speech Communication, 36*, 327-342.

Lin, T., & Wang, W. S-Y. (1984). Shengdiao ganzhi wenti (Tone perception). *Zhongguo Yuyan Xuebao, 2*, 59-69.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects of speech perception: effect of preceding liquid on stop consonant identification. *Perception and Psychophysics, 60*, 602–619.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). *Journal of the Acoustical Society of America, 102*, 1134–1140.

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification (L). *Journal of the Acoustical Society of America, 113*, 53-56.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 391–409.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception and Psychophysics, 28*, 407-412.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69*, 548–558.

Monahan, P. J., & Idsardi, W. J. (2010). Auditory sensitivity to formant ratios: Toward an account of vowel normalization. *Language and Cognitive Processes, 25*, 808-839.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America, 102*, 1864-1877.

Morris, R. J., McCrea, C. R., & Herring, K. D. (2008). Voice onset time differences between adult males and females: Isolated syllables. *Journal of Phonetics, 36*(2), 308–317.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*, 379–390.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*(1), 365–378.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution 4*(2), 133-142.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*, 2088–113.

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America, 80*, 1297–1308.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In Johnson, K., & Mullennix, J. W. (Eds.), *Talker Variability in Speech Processing* (pp. 109-132). San Diego: Academic Press.

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

Nygaard, L., & Pisoni, D. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60*, 355-376.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 309-328.

Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: A syllabus-based comparative study of Mandarin and Cantonese. *Journal of Chinese Linguistics, 34*, 135–154.

Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. W., & Wang, W. S-Y. (2012). The effect of inter-talker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems. *Journal of Speech, Language, and Hearing Research, 55*, 579-595.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175-184.

Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America, 101*, 2267-2277.

Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America, 101*, 2267-2277.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance, 23*(3), 651–666.

Shankweiler, D., Strange, W., & Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In Shaw, R., & Bransford, J., (Eds.) *Perceiving, Acting, and Knowing* (pp. 315-345). Hillsdale, NJ: Erlbaum.

Sjerps, M. J., & Smiljanic, R. (2013). Compensation for vocal tract characteristics across native and non-native languages. *Journal of Phonetics, 41*, 145–155.

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

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011b). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception and Psychophysics, 73*, 1195-1215.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2012). Hemispheric differences in the effects of context on vowel perception. *Brain and Language, 120*, 401-405.

Slawson, A. (1968). Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *Journal of the Acoustical Society of America, 43*, 87.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79*, 1086-1100.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2013). Similar response patterns do not imply identical origins: An energetic masking account of nonspeech effects in compensation for coarticulation. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 1181-1192.

Von Kriegstein, K., & Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage, 22*, 948-955.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America, 90*, 2942-2955.

Watkins, A. J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America, 99*, 3749-3757.

Wong, P. C. M. (1998). Speaker normalization in the perception of Cantonese level tones. Master's thesis, University of Texas at Austin unpublished.

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

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16*, 1173–1184.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*, 61–83.

Zhang, C., Peng, G., & Wang, W. S-Y. (2012). Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones. *Journal of the Acoustical Society of America, 132*, 1088-1099.

Zhang, C., Peng, G., Wang, & W. S-Y. (2013). Achieving constancy in spoken word identification: Time course of talker normalization. *Brain and Language, 126*, 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.

Table 1. Mean and standard deviation (SD) of F0, F1 and F2 of the target words produced by four talkers in Experiment 1.

Talker	**F0 (SD)**	**F1 (SD)**	**F2 (SD)**
Female F01	257.0 (14.6)	307.8 (20.0)	3167.3 (66.6)
Female F02	225.5 (5.1)	286.8 (21.9)	2708.6 (25.6)
Male M01	156.8 (3.6)	253.5 (28.8)	2267.0 (30.9)
Male M02	104.8 (2.8)	241.8 (13.5)	2498.0 (72.4)

Table 2. An overview of the four context conditions in Experiment 2A and 2B.

Context	Type of cue	Experiment
Nonspeech	Auditory only	2A, 2B
Reversed speech	Auditory+Phonetic	2A
Meaningless speech (呢錯視幣_. /li55 tsʰo33 si22 pɐi22 _/. 'This mistake sees money _.')	Auditory+Phonetic+Phonological	2B
Meaningful speech (呢個字係_. /li55 ko33 tsi22 hɐi22 _/. 'This character is _.')	Auditory+Phonetic+Phonological +Semantic+Syntactic	2A, 2B

Table 3. Mean, minimal and maximal F0 of the meaningful speech context, and the mean and SD of the F0, F1 and F2 of the target words produced by four talkers in Experiment 2. The other three types of contexts (meaningless speech, reversed speech and nonspeech) were matched with the speech context in terms of the mean, minimal, maximal F0.

Talker	**Context**			**Target**		
	Lowered: Mean F0 (min, max)	Unshifted: Mean F0 (min, max)	Raised: Mean F0 (min, max)	F0 (SD)	F1 (SD)	F2 (SD)
Female F03	198.2 (151.9, 275.9)	236.8 (190.2, 314.4)	280.5 (233.7, 357.7)	234.0 (4.9)	285.0 (15.9)	3026.4 (32.5)
Female F04	173.9 (116.0, 265.6)	208.1 (148.7, 300.1)	246.3 (193.4, 337.8)	206.8 (3.0)	304.3 (40.8)	2608.4 (30.7)
Male M03	124.5 (84.8, 184.9)	148.4 (108.1, 208.7)	174.6 (134.7, 234.6)	143.6 (5.6)	281.5 (24.5)	2586.3 (186.1)
Male M04	96.8 (75.0, 137.7)	113.8 (88.8, 157.1)	134.5 (110.0, 177.5)	114.8 (1.7)	265.4 (20.1)	2426.2 (9.0)

Table 4. A summary of the results of statistical analyses of Experiment 2A and 2B combined.

	Chi square	**DF**	***p* value**
Main effects			
Context	2516.6	3	$p < 0.001$
F0 shift	487.67	2	$p < 0.001$
Talker	19.593	3	$p < 0.001$
Two-way interaction effects			
Context by F0 shift	245.88	6	$p < 0.001$
Context by Talker	596.07	6	$p < 0.001$
F0 shift by Talker	25.869	9	$p = 0.002$
Three-way interaction effect			
Context by F0 shift by Talker	160.92	18	$p < 0.001$

Figure 1. F0 range of four talkers. (A) Experiment 1. (B) Experiment 2. The dotted line indicates the estimated population F0 range for female and male talkers.

Figure 2. Results of Experiment 1. (A) Accuracy of word identification for the four talkers. The dotted line indicates the chance-level accuracy (0.33). (B) Perceptual height score plotted as a function of the distance of each talker’s upper F0 range from the gender-specific population F0 range. (C) Perceptual height score plotted as a function of the distance of each talker’s lower F0 range from the gender-specific population F0 range. The dotted lines indicate the perceptual height score of high level tone ‘6’, mid level tone ‘3’ and low level tone ‘1’.

Figure 3. The F0 trajectory of contexts with raised, unshifted and lowered F0, followed by the target word superimposed on the spectrogram. (A) Nonspeech context. (B) Reversed speech context. (C) Meaningless speech context. (D) Meaningful speech context. The red line represents the raised F0 context, the black line represents the unshifted F0 context and the blue line represents the lowered F0 context.

Figure 4. Results of Experiment 2. (A) Rate of expected responses for the nonspeech, reversed speech and meaningful speech context in Experiment 2A. (B) Rate of expected responses for the nonspeech, meaningless speech and meaningful speech context for Experiment 2B. Dotted lines indicate the chance level accuracy (0.33). Asterisks refer to the significance level of generalized mixed-effects models comparing the identification accuracy with chance-level accuracy (see the text for details): *, $p<0.05$, **, $p<0.01$; ***, $p<0.001$.

Figure 5. An integrative model of talker normalization.

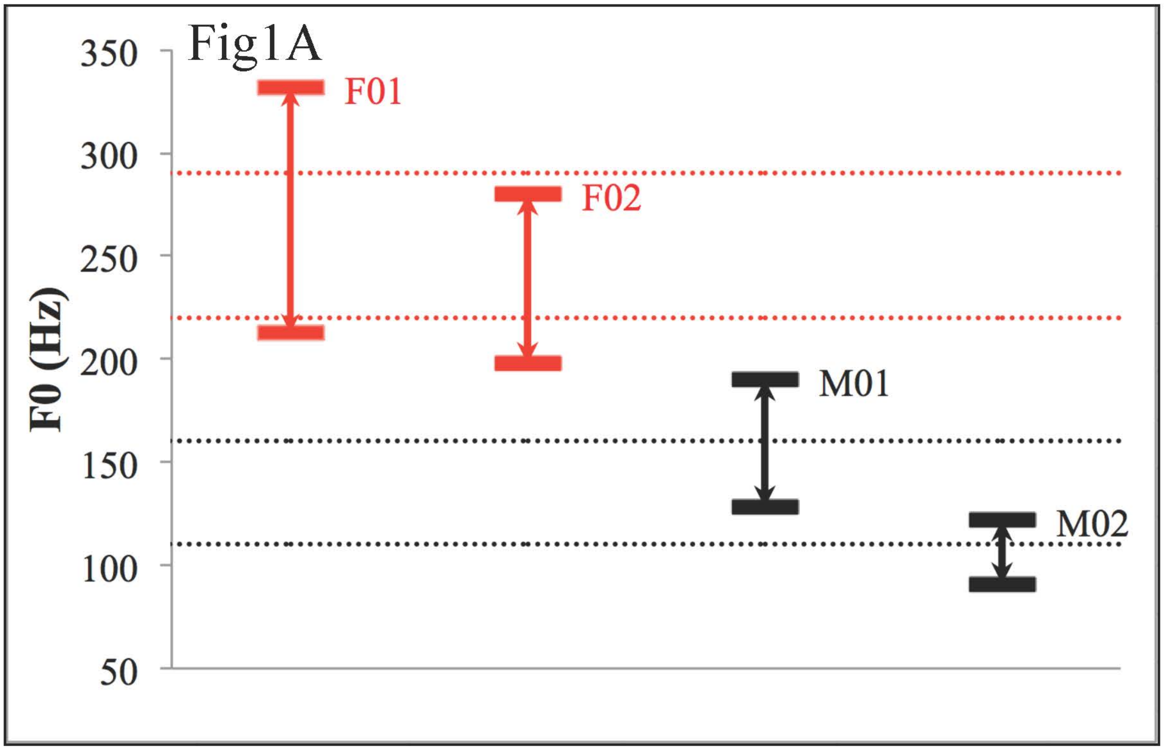


Fig1A
350
300
250
200
150
100
50
F0 (Hz)
F01
F02
M01
M02

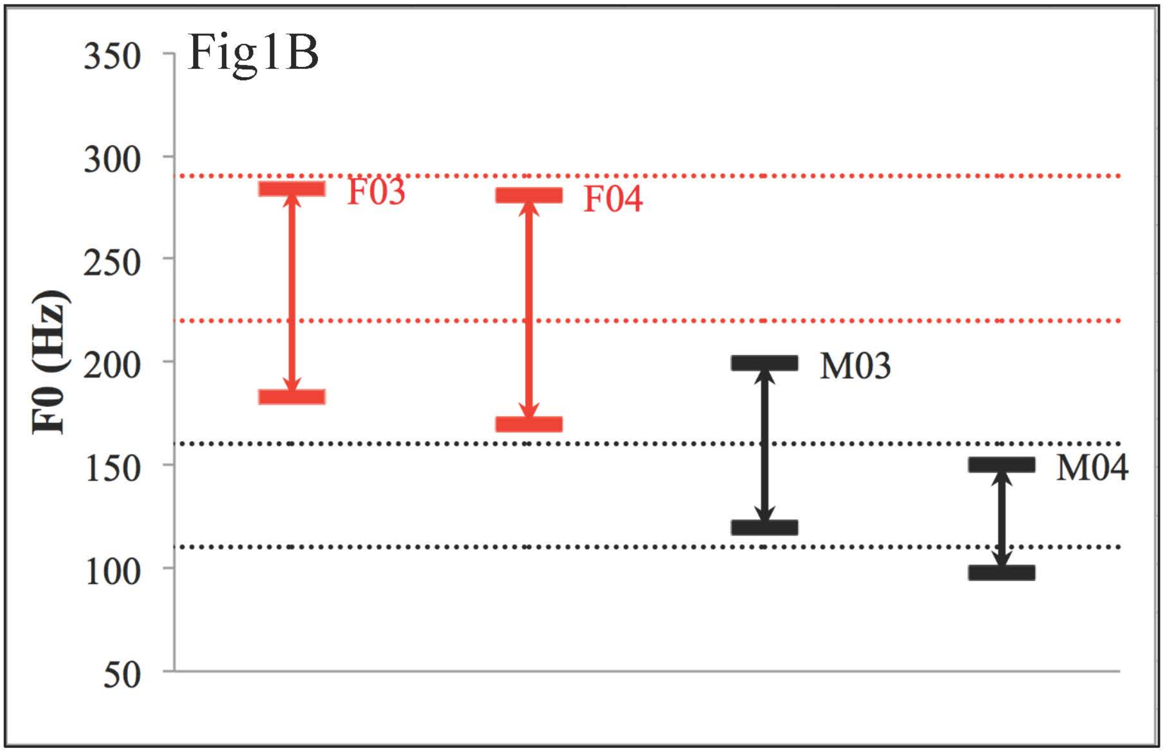


Fig1B
350
300
250
200
150
100
50
F0 (Hz)
F03
F04
M03
M04

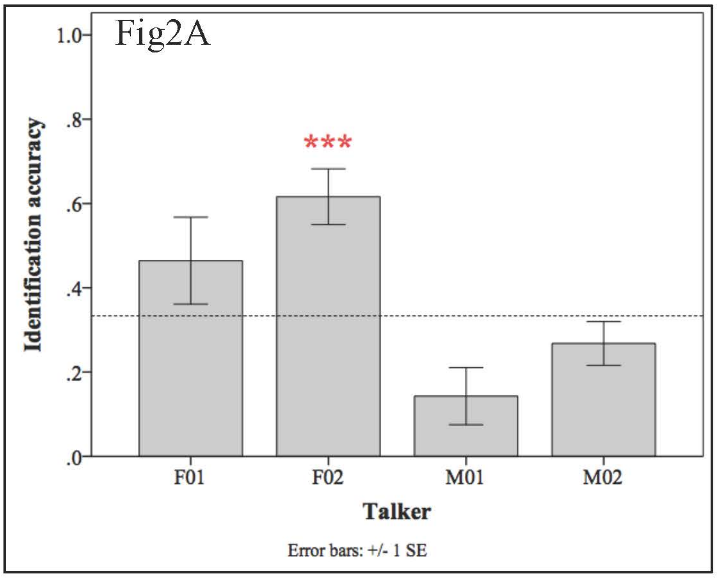


Fig2A
1.0
.8
.6
.4
.2
.0
Identification accuracy
***
F01
F02
M01
M02
Talker
Error bars: +/- 1 SE

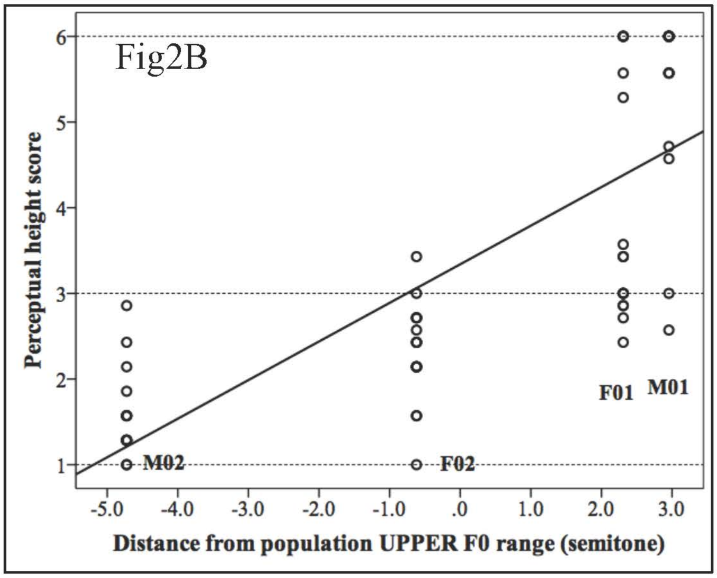


Fig2B
6
5
4
3
2
1
Perceptual height score
M02
F02
F01
M01
-5.0
-4.0
-3.0
-2.0
-1.0
.0
1.0
2.0
3.0
Distance from population UPPER F0 range (semitone)

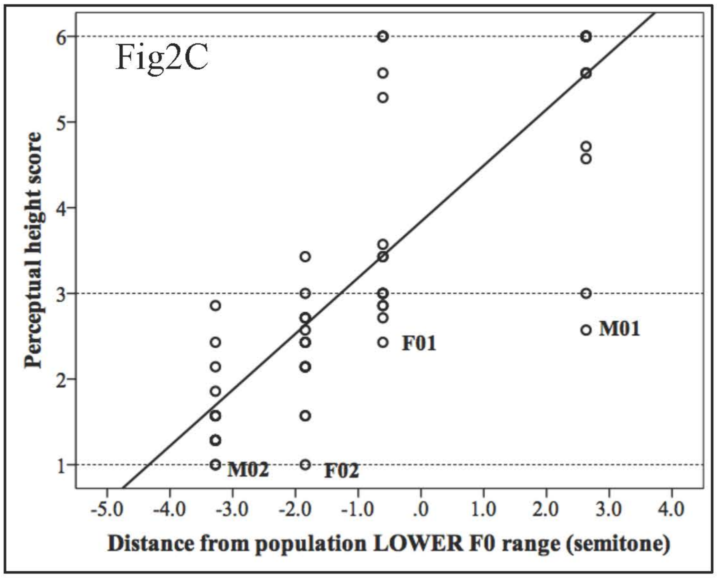


Fig2C
6
5
4
3
2
1
Perceptual height score
F01
M01
M02
F02
-5.0
-4.0
-3.0
-2.0
-1.0
.0
1.0
2.0
3.0
4.0
Distance from population LOWER F0 range (semitone)

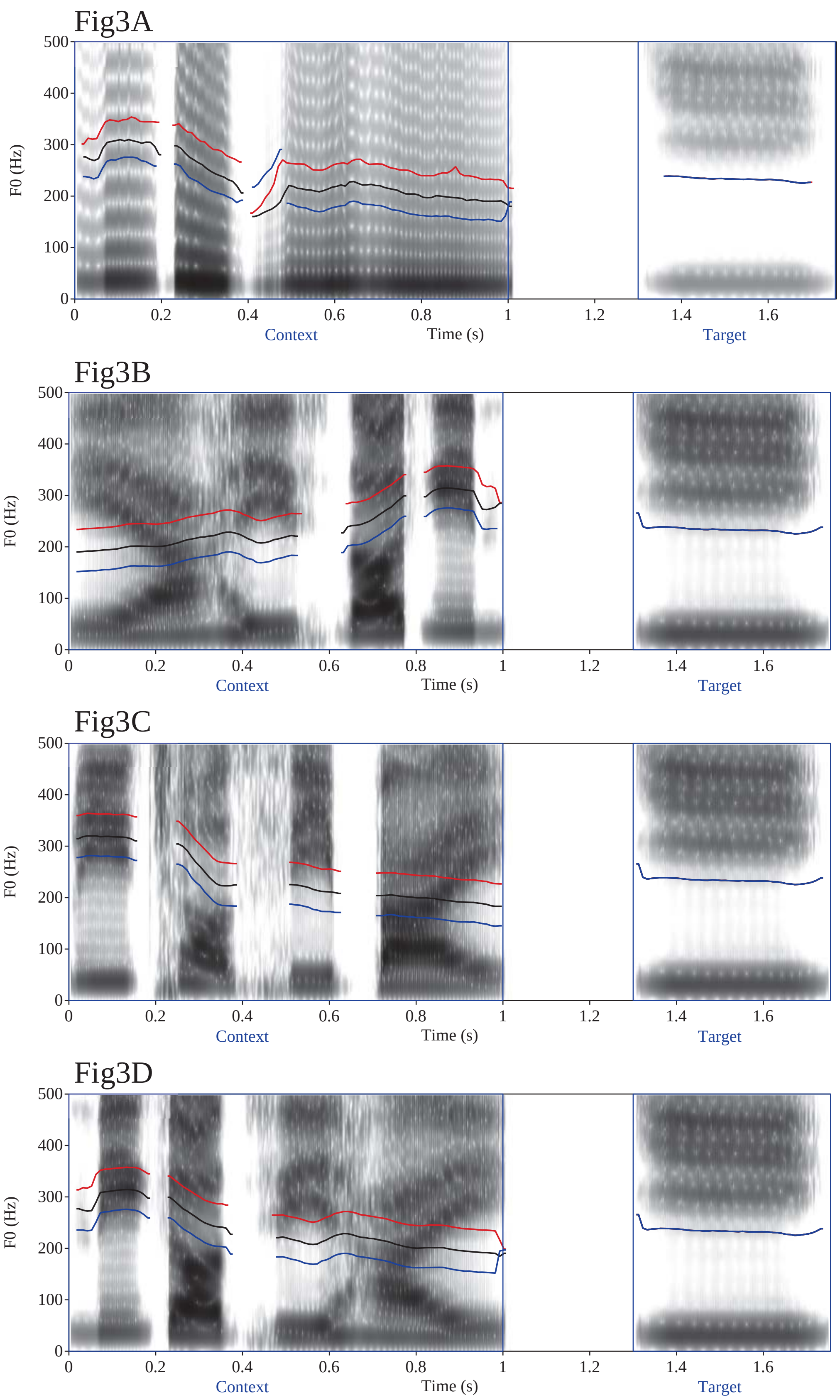


Fig3A
Fig3B
Fig3C
Fig3D
F0 (Hz)
0
100
200
300
400
500
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Time (s)
Context
Target

Fig4A

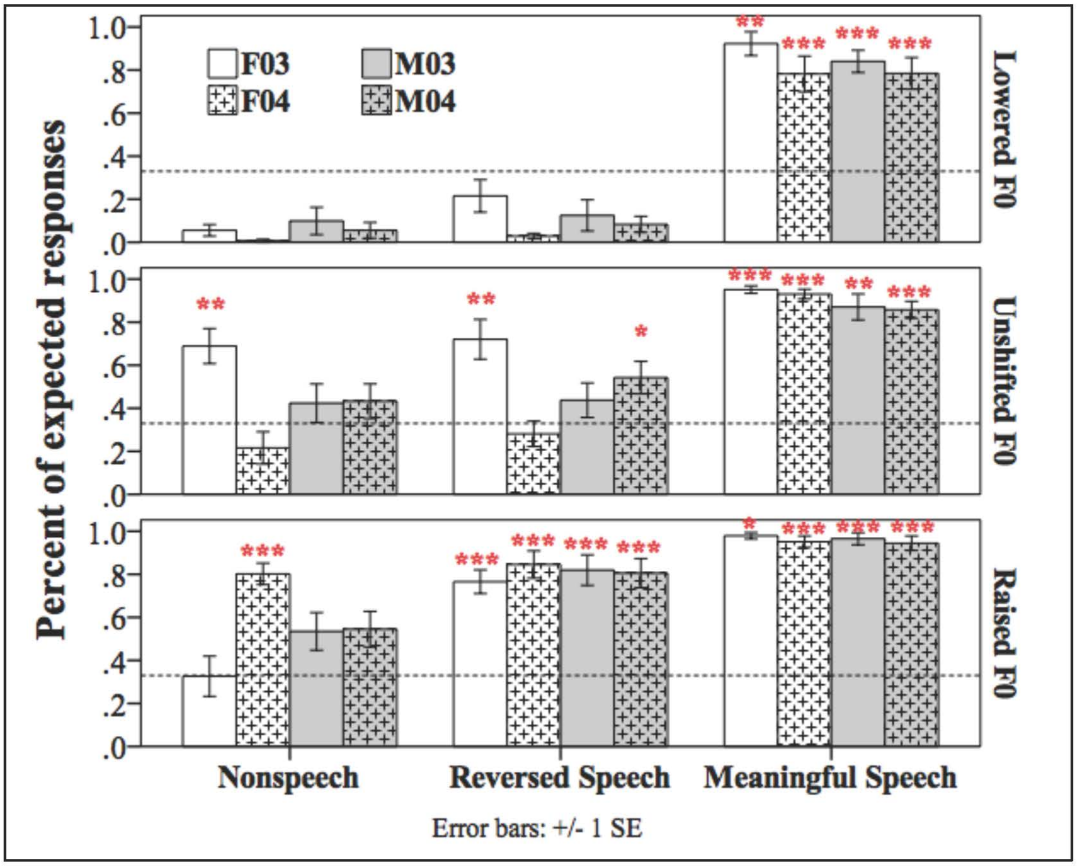

F03
M03
F04
M04
Percent of expected responses
1.0
.8
.6
.4
.2
.0
Lowered F0
Unshifted F0
Raised F0
Nonspeech
Reversed Speech
Meaningful Speech
Error bars: +/- 1 SE

Fig4B

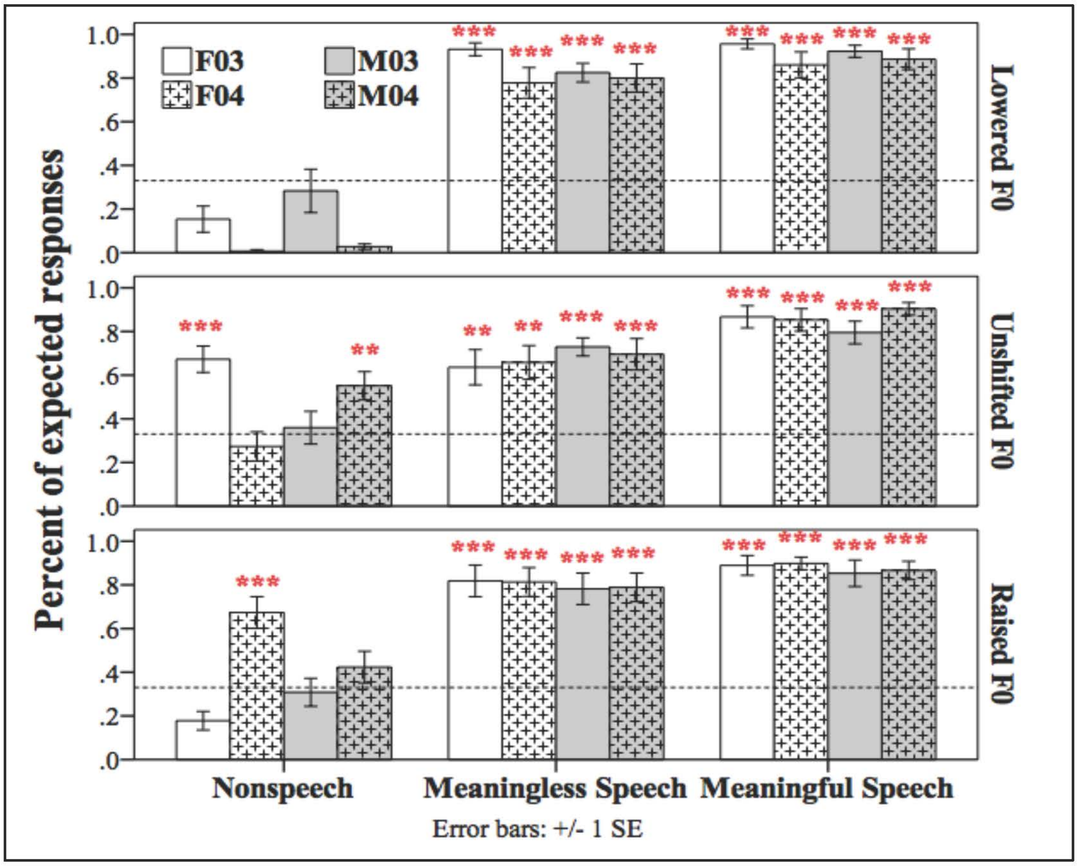
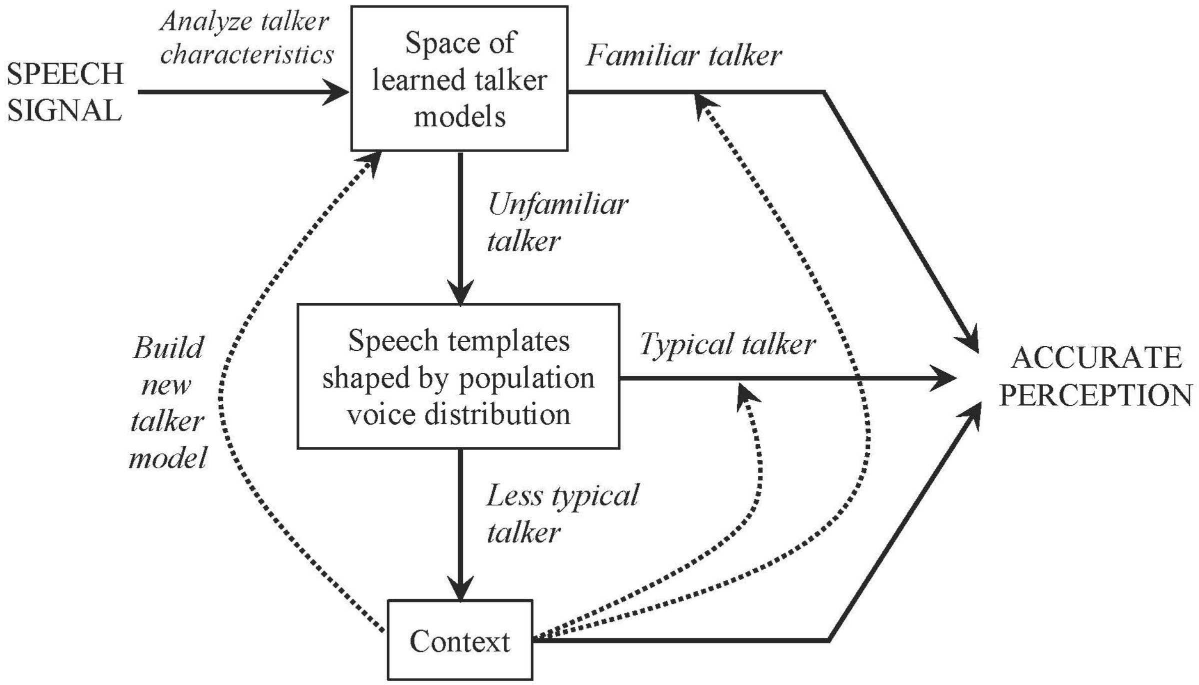

F03
M03
F04
M04
Percent of expected responses
1.00
.80
.60
.40
.20
.00
Lowered F0
Unshifted F0
Raised F0
Nonspeech
Meaningless Speech
Meaningful Speech
Error bars: +/- 1 SE

Fig5


SPEECH SIGNAL
Analyze talker characteristics
Space of learned talker models
Familiar talker
Unfamiliar talker
Speech templates shaped by population voice distribution
Typical talker
ACCURATE PERCEPTION
Build new talker model
Less typical talker
Context