INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Auditory-visual integration of talker gender in Cantonese tone perception

Wei Lai

Department of Linguistics, University of Pennsylvania
weilai@sas.upenn.edu

Abstract

This study investigated the auditory-visual integration of talker gender in the perception of tone variances. Two experiments were conducted to evaluate how listeners use the information of talker gender to adjust their expectation towards speakers' pitch range and uncover intended tonal targets in Cantonese tone perception. Results from an audio-only tone identification task showed that tone categorization along the same pitch continuum shifted under different conditions of voice gender. Listeners generally heard a tone of lower pitch when the word was produced by a female voice, while they heard a tone of higher pitch when the word was produced at the same pitch level by a male voice. Results from an audio-visual tone identification task showed that tone categorization along the same pitch continuum shifted under different conditions of face gender, despite the fact that the photos of different genders were disguised for the same set of stimuli in identical voices with identical pitch heights. These findings show that gender normalization plays a role in uncovering linguistic pitch targets, and lend support to a hypothesis according to which listeners make use of socially constructed stereotypes to facilitate their basic phonological categorization in speech perception and processing.

Index Terms: speech perception, normalization, gender, tone

1. Introduction

A tone language uses f_0 to contrast word meaning. However, inter and intratalker variation in tone production may contribute to acoustic overlap among tone categories. For one thing, tone categorization is based on the perceived relative pitch height with reference to a specific talker's f_0. But talkers in a speech community have different f_0, partly due to physiological differences in the vocal apparatus. Consequentially, different tone categories can overlap in f_0 space, e.g., a high tone spoken by a male talker might have an f_0 value similar to a low tone of a female talker. Further more, f_0 for a particular tone can also vary within the same talker across the day, and change as a function of mood, emotion, etc.

The variation in f_0 of lexical tones might give rise to lexical ambiguity in perception [1-4]. One crucial process that helps reduce this ambiguity is tone normalization of inter and intratalker variability based on the information about a talker's f_0, such as speech materials or context that contains cues to a talker's f_0. [1] found that the identification accuracy for the three Cantonese lexical tones was significantly higher when the presentation of tokens was blocked by talker rather than mixed across talkers. [2, 3] showed that raising or lowering the f_0 of speech context changed the perception of identical stimuli from a mid tone to a low or a high tone in Cantonese tone perception. [3] further pointed out that the tone identification accuracy was slightly increased when the tone was embedded in a contoured f_0 context rather than a flattened f_0 context, which suggested that the f_0 range affected tone normalization separately from the f_0 mean. Context effect on tone perception was also found in other tone languages such as Mandarin [5-7].

It still remains mysterious, though, how listeners locate a tone relative to an estimation of speaker f_0 range even before the speaker's speech information is available. In other words, how do listeners identify the tone on the first syllable they hear from a speaker without reference to his or her speech? Evidently, one reliable cue people would use in estimating a talker's f_0 range is gender. It has been universally accepted that men on average have lower f_0 than women, partly due to their larger larynxes [8, 9]. The association between f_0 and perceived talker gender allows listeners to compensate for the f_0 difference between genders and categorize varied f_0 values as the same tone by imposing a high or low f_0 baseline in speech perception depending on whether the speaker looks or sounds like a female or a male.

Aside from the dominant difference in f_0, male and female speech was also found to differ in quite a few other ways. Compared to males, females have been found to have higher vowel formant frequencies as a result of their smaller vocal track [10, 11], slower speech rate [12], more breathiness in voice quality [13, 14], etc.

The integration of talker gender in speech perception has already been studied at the segmental level. [15] showed that in the perception of a /s/ - /ʃ/ continuum, listeners appeared to categorize the synthetic fricatives differently depending on the perceived gender of the voice producing the rest of the word. A female voice leads the listener to expect higher sibilance frequency since females' sibilance is usually centered around a higher frequency than males'. This gender effect on tone calibration has also been found triggered by visual gender stereotype, i.e., tokens associated with the same voice but faces of different genders, similar to a "McGurk effect" paradigm [16]. This visual gender effect has been replicated on vowel perception using a vowel continuum spanning from /u/ to /a/ with the F1 value altered for each step in the continuum [17]. Furthermore, [18] found that the gender effect on vowel perception could be triggered even without auditory or visual inputs; but rather, simply suggestions about the gender of an illusory talker could affect the gender-ambiguous vowels' perception. The above results indicate that listeners would use socially constructed beliefs to assist their basic phonological categorization of speech signals.

This paper investigates whether or how listeners integrate talker gender in lexical tone perception by creating expectations towards the talkers' f_0 range and implementing a degree of gender normalization. Rich level tones in Cantonese provide an optimal window to probe the mechanism of tone normalization. In Cantonese, there are three level tones, two rising tones and a falling tone. The three level tones mainly contrast in pitch height: high level tone (e.g., /ji55/ 医 "a doctor"), mid level tone

http://dx.doi.org/10.21437/Interspeech.2017-1069

(e.g., /ji33/ 意 “meaning”), and low level tone (e.g., /ji22/ 儿 “son”). Two experiments were conducted to test how listeners integrate the information of talker gender revealed by auditory and visual cues in Cantonese level tone perception. In the first experiment, we evaluate whether voices of different genders will lead to different patterns of tone categorization along the same pitch continuum. In the second experiment, we evaluate whether faces of different genders will lead to different patterns of tone categorization along the same pitch continuum superimposed to the same voice. Our general hypothesis is that the tone categorization will shift to the low end of the pitch continuum when the talker is perceived to be a male, either by auditory or visual cues; on the contrary, when the talker is perceived to be a female, the categorization will shift to the high pitch end along the continuum.

2. Experiment I

2.1. Voice selection

A pretest was first conducted to select prototypical male and female voices, and the selected voices would then be used to construct stimuli for tone identification experiments. 20 Voices were recorded from 10 male speakers and 10 female speakers, with each speaker producing a “yi” sound with a high tone. Among the 20 voice contributors, 8 natively speak Cantonese and 12 speak Mandarin. This should not be problematic since, in both languages, the same high-tone “yi” sound is associated with real words, and the high tone in Mandarin is typically considered to have the same f_0 target (55) as the high tone in Cantonese. All materials were recorded in a professional recording booth at the University of Pennsylvania. The speakers repeated the sound several times, and only a well-articulated token with stable and smooth formant and pitch trajectories was selected for each speaker.

Then the gender stereotypicality of these voices was rated for voice selection. Since pitch height would affect listeners’ expectation towards talker gender, we resynthesized the 20 sounds with 3 levels of pitch height superimposed: 2 semitones (St), 7 St and 12 St on the base of 100 Hz, using the Linear Predictive Coding (LPC) algorithm in Praat [19]. The three pitch values are the endpoints and the midpoint of the pitch continuum adopted in the tone identification task (see 2.2).

Each of the 60 sounds (20 voices * 3 pitch levels) were rated by 20 listeners for both gender stereotypicality and naturalness. The listeners, 12 females and 8 males, between age 16 and 27, all reported that they speak Cantonese as their native language. The sounds were presented over headphones using a Praat MFC (Multiple Forced Choice) experiment interface. These listeners could listen to each sound for as many times as they wanted, and then rated the voice they heard on a 9-point scale from “most feminine” to “most masculine”, with a score of 1 being “most male-sounding”, 9 being “most female-sounding” and 5 being “completely neutral in gender”. These sounds were also rated for naturalness on a 5-point scale.

2 prototypical female voices and 2 prototypical male voices were selected based on the rated gender stereotypicality and naturalness. Figure 1 shows the rated masculinity/femininity for each voice at different pitch levels. On one hand, we can see that an increase in pitch indeed raises the amount of perceived femininity for all the four voices; on the other hand, male voices and female voices of the same pitch height can still differ in their gender-stereotypicality rating, and the gap becomes bigger for higher pitches. As for naturalness, the four voices generally have a score above 3 in all conditions, except for the fact that one of the female voices received a score below three (2.0) at the pitch height of 2 Hz. Table 1 shows the formants of the 4 selected “yi” sounds. As expected, female speakers generally have higher vowel formants than male speakers do.

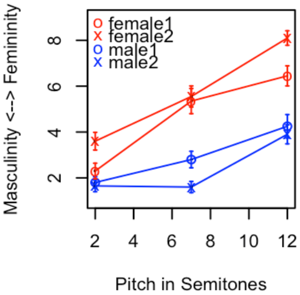


Figure 1: *Mean and standard error of gender-rating scores for the selected 4 voices at 3 pitch height (2 St, 7 St, 12 St)*

Table 1. *The formants of the 4 selected “yi” sounds (Hz)*

	Female1	*Female2*	*Male1*	*Male2*
F1	359	401	272	311
F2	2652	2922	2305	2203
F3	3454	3639	3344	2958

2.2. Pitch manipulation

Each of the 4 selected voices was superimposed with an 11-step pitch continuum from 2 semitones to 12 semitones on the base of 100 Hz, using Linear Predictive Coding (LPC) algorithm in Praat. Since duration was reported to play a role in the tonal contrasts between Cantonese level tones in production and perception (Wong and Diehl, 1999), all the stimuli in this experiment were normalized to the duration of 0.45 seconds and the intensity of 60 dB.

2.3. Procedure

A three-alternative forced-choice tone identification task was used to test listeners’ tonal categorization of pitch values under different gender-voice conditions. Subjects were instructed to identify the word they heard as any of the three Cantonese words, “医”(55), “意”(33) and “儿”(22) after hearing each sound. Subjects were allowed to hear the sound as many times as they wanted, and were asked to respond by clicking on the button with the correct Chinese character on the screen.

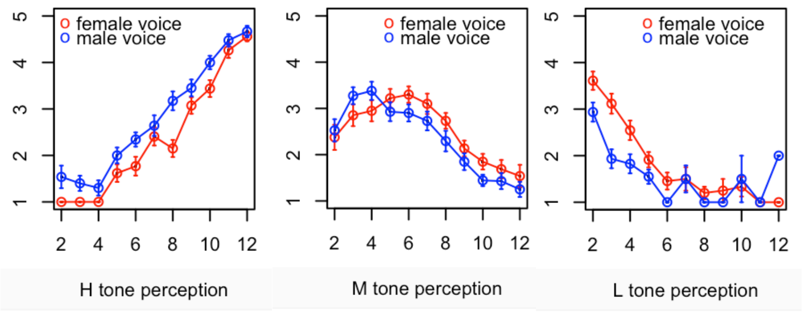
The task was repeated 5 times, with each repetition presented in a separate block. The order of items was randomized within each block. In total there were 4 voices * 11 steps * 5 blocks = 220 tokens. The whole procedure usually took 30 minutes.

2.4. Subjects

21 participants (15 females and 6 males), between age 15 and 24, were recruited to participate in the tone identification task. 4 of them were recruited from the student population at University of Pennsylvania, and 17 of them were recruited from the student population at Xiangxian Middle School, Guangdong Province in China mainland. All of them reported to speak Cantonese as their primary language. None of the participants reported having hearing issues.

2.5. Results

Figure 2 contains three graphs that respectively show how many times (out of a total of 5) high, mid and low tones were



H tone perception

M tone perception

L tone perception

Figure 2: *Mean and SE for times of high, mid and low responses per listener under different voice gender conditions (x-axis: pitch; y-axis: times of response)*

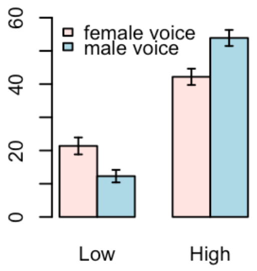


Figure 3: *Mean and SE for total numbers of high and low responses per listener under different voice gender conditions*

identified along the 11-step pitch continuum under different voice gender conditions, averaged across 21 participants. In all three graphs, a split is observed between different voice gender conditions in the distribution of responses along the continuum, with the distribution shifted towards the low end under the male voice condition and towards the high end under the female voice condition. Figure 3 shows the total number of high and low responses under different voice gender conditions averaged across participants. Under the same pitch height condition (from 2 St to 11 St), items in female voices were perceived to have more low tones and fewer high tones (21 and 42) than items in male voices (12 and 54).

Two logistic regression models were conducted to evaluate this voice gender effect on tone identification. In the first model, the dependent value is a [-H] judgment, and the independent variables are Gender and Pitch-step. The models also contain the interaction Pitch-step x Gender. The second model is the same as the first one, except that its dependent value is a [-L] judgment. The coefficient estimate for Pitch-step is 0.58 in Model 1 and -0.53 in Model 2, representing the change in the log odds of a [-H]/[-L] response associated with one step of pitch change (2 St). The coefficient estimate for Gender is 1.2 in Model 1 and -0.8 in Model 2, representing the effect of male gender on the log odds of a [-H]/[-L] response relative to the geometric mean of both genders. In both models the two variables, Pitch and Gender, are each significant (p<.001). The interaction between variables is significant in Model 1 (p<.05) but insignificant in Model 2.

3. Experiment II

3.1. Face selection

A face-selection pretest was conducted to select prototypical female faces and prototypical male faces as experimental materials from a pool of 20 photos of different Chinese faces (10 females, 10 males) downloaded from the internet. 10 Chinese participants (4 males, 6 females) rated all 20 photos on a 9-point scale from “most feminine” to “most masculine”, with a score of 1 being “most male-looking”, 9 being “most female-looking” and 5 being “completely neutral in gender”. These faces were also rated for “attractiveness” on a 5-point scale. Differing from the gradient scores of voice gender rating, scores of face gender rating are clearly distributed around the extreme values, with few intermediate responses in gender perception.

3 pairs of prototypical female and male faces were selected based on the above ratings, of which one pair was used to construct the critical stimuli, and the other two pairs were used as fillers. The gender rating scores for the stimulus photos are: 1.5 (male) and 8.5 (female); and the averaged scores for the filler photos are 1.3 (male) and 8.3 (female). The attractiveness scores for the 6 photos are all around 3, meaning that the results of this experiment should not be affected by attractiveness of the faces.

3.2. Voice selection and manipulation

We selected a gender-neutral voice to construct stimuli and two pairs of prototypical female and male voices to construct fillers based on the the gender-stereotypicality rating scores reported in 2.1. The gender-neutral voice we selected was respectively rated 2.7, 3.3 and 5.5 at the pitch height of 2 St, 7 St and 12 St. For the filler voices we used the four voices selected in 2.1, with their gender rating scores in Figure 1.

The stimulus voice was superimposed with an 11-step pitch continuum from 2 St to 12 St on the base of 100 Hz, using the Linear Predictive Coding (LPC) algorithm in Praat. The filler voices were superimposed with a 6-step pitch continuum from 2 St to 12 St with a pitch change of 2 St at each step. All the resynthesized sounds were normalized to 0.45 s and 60 dB.

3.3. Procedure

The task was also a three-way forced choice identification in which subjects were instructed to identify the target word as any of the three Cantonese words “医”, “意” and “儿” after hearing each sound. However, different from experiment 1, a photo was presented on a PowerPoint slide associated with each sound and subjects were told to look at the photo that shows the speaker of that sound before making a choice. The subjects were also instructed to listen to the sound as many times as they wanted, and to mark their response in a textbox below each photo.

The stimuli and fillers were presented 5 times in five blocks. In each block, the set of 9-step stimuli with a gender-neutral voice occurred twice, once patterned with a female face and a second time with a male face. Meanwhile, the two pairs of gender-stereotypical voices were matched one-to-one with the two pairs of gender stereotypical faces, with the constraint that the matched voice and face should be of the same gender. The combinations of faces and voices are shown in Table 2.

Within each block, the items appear in a pseudorandom order with the constraints of a) the first three items are fillers; b) two target stimuli don’t occur next to each other; and c) stimuli with extreme pitch values (i.e., a female face patterned with a low pitch and a male face patterned with a high pitch) occur among the second half of the tokens. In total there are 11 (steps of stimuli) * 2 (voice-face pairs) * 5 (blocks) + 6 (steps of fillers) * 4 (voice-face pairs) * 5 (blocks) = 230 items.

Table 2. *The combination of faces and voices in Exp 2*

Tokens	*Visual female tokens*	*Visual male tokens*
Fillers	F Face 1 – F Voice 1	M Face 1 – M Voice 1
	F Face 2 – F Voice 2	M Face 2 – M Voice 2
Stimuli	F Face3 – Neutral Voice	M Face 3 – Neutral Voice

3.4. Subjects

8 subjects (6 females and 2 males), between the ages of 16 and 22, were recruited to participate in the task. 3 of them were recruited from the student population at the University of Pennsylvania, and 5 of them were recruited from the student population at Xiangxian Middle School, Guangdong Province in China mainland. All of them reported to speak Cantonese as their primary language. None of the participants reported having hearing issues. One subject failed to mark his responses for all the stimulus items, and thus his data was excluded from the experiment.

3.5. Preliminary result

Figure 4 shows how many times (out of a total of 5) high, mid and low tones were identified along the 11-step pitch continuum under different face gender conditions, averaged by subjects. Although the split between different face gender conditions in the distribution of responses is not totally distinct, it can still be observed that the distribution is shifted towards the high pitch end when a female face was presented and is shifted towards the low pitch end when a male face was presented for the identification of all three tones.

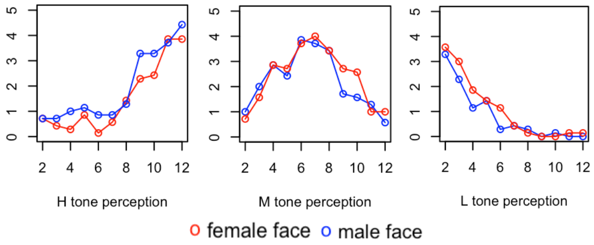


Figure 4: *Mean for times of high, mid and low responses per listener under different voice gender conditions (x-axis: pitch; y-axis: times of response)*

Figure 5 shows the total number of high and low responses under different face gender conditions per subject. For the identical set of speech tokens, items associated with a female face were perceived to have more low tones and less high tones (12 and 17) than items associated with a male face (10 and 21).

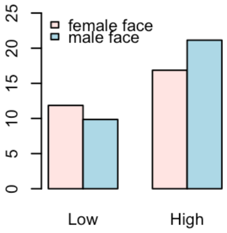


Figure 5: *Mean for total numbers of high and low responses per listener under different voice gender conditions*

4. Discussion

Results from both experiments support our hypothesis of the effect of talker gender on tone perception. If the listeners hear or see a female, they form an expectation that the talker has a high pitch and use that expectation to implement tone normalization. Thereby, they will more frequently perceive tones of lower pitch because the f_0 baseline is assumed to be higher. On the contrary, if they hear or see a male, they will implement tone normalization by a low pitch baseline that they expected of the speaker and will perceive tones of higher pitch more frequently.

A mystery that still remains to be solved is that the gender effect on tone perception revealed by this paper is much smaller than expected. The f_0 difference between male and female in the real world can be very large, even sometimes with the possibility that the minimum f_0 of a female speaker is similar to the maximum f_0 of a male speaker. Accordingly, the gap between the two distributions of high/mid/low responses for different talker genders should be larger than just the one or two steps that we found in this paper.

We partly attribute this discrepancy to the limitation of experimental settings. One of the limitations is that the gender stereotypicality listeners perceived does not remain stable along the continuum, due to the effect from the change of pitch height, as shown in Figure 1. This allows more degrees of freedom in the normalization besides the dichotomy of male and female. Another limitation is the lack of f_0 range information. Previous studies show that the use of a wider pitch range will contribute to the perception of femininity, even where the overall pitch is low [20], but listeners in this paper cannot use this information in gender perception because the pitch ranges for male-voice stimuli and female-voice stimuli are identical.

In spite of these limitations, the results show that socially constructed stereotypes may play a role in the basic phonological categorization of the f_0 signal. This finding challenges “gender-free” conceptions of human speech perception and language processing, and introduces a new set of complications to the present conception of cognitive processing.

5. Conclusions

This study explores how listeners integrate the information of talker gender in the perception of pitch height. Two experiments were conducted to evaluate potential gender normalization based on auditory and visual cues in Cantonese level tone identification. In the first experiment, when the same set of pitch continuum was superimposed on female-voice speech and male-voice speech, listeners more frequently heard a tone of lower pitch for items in a female voice compared to its male-voice counterpart at the same pitch height. In the second experiment, when the same set of pitch continuum superimposed on the same voice is presented with a photo of a female speaker or a male speaker, listeners more frequently heard a tone of lower pitch for items presented with a female face than its male-face counterpart in the same voice at the same pitch height. These results support the hypothesis that listeners make use of talker gender to form expectations towards talkers' f_0 range and implement a degree of normalization in tone perception by imposing a high or low f_0 baseline depending on whether the speaker looks or sounds like a female or a male.

6. Acknowledgement

The author would like to thank Professor Meredith Tamminga, Professor Abby Walker and Lacey Arnold Wade for their helpful comments and feedback.

7. References

[1] P. C. Wong and L. D. Randy L. “Perceptual normalization for inter-and intratalker variation in Cantonese level tones,” *Journal of Speech, Language, and Hearing Research* 46.2: 413-421. 2003.

[2] A. L. Francis, V. Ciocca, N. K. Y. Wong, W. H. Y. Leung & P. C. Y. Chu. “Extrinsic context affects perceptual normalization of lexical tone,” *The Journal of the Acoustical Society of America*, 119(3), 1712-1726. 2006.

[3] C. Zhang, G. Peng and W. S. Wang. “Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones,” *The Journal of the Acoustical Society of America*, 132(2), 1088-1099. 2002.

[4] G. Peng, C. Zhang, H. Zheng, J. W. Minett and W. S. Wang. "The Effect of Intertalker Variations on Acoustic–Perceptual Mapping in Cantonese and Mandarin Tone Systems," *Journal of Speech, Language, and Hearing Research* 55, no. 2: 579-595. 2002.

[5] J. Leather. “Speaker normalization in perception of lexical tone,” Journal of Phonetics, 11, 373–382. 1983.

[6] R. Fox and Y. Y. Qi. “Context effects in the perception of lexical tone,” *Journal of Chinese Linguistics*, 18, 261–283. 1990.

[7] Y. Xu. “Production and perception of coarticulated tones,” *Journal of the Acoustical Society of America*, 95, 2240–2253. 1994.

[8] R. O. Coleman. "Speaker identification in the absence of inter-subject differences in glottal source characteristics," *The Journal of the Acoustical Society of America* 53, no. 6: 1741-1743. 1973.

[9] C. D. Aronovitch. “The voice of personality: stereotyped judgements and their relation to voice quality and sex of speaker,” *Journal of Social Psychology*, 99, 207–220. 1976.

[10] G. E. Peterson and H. L. Barney. “Control methods used in a study of the vowels,” *The Journal of the acoustical society of America*, 24(2), 175-184. 1952.

[11] R. O. Coleman. “A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice,” *Journal of Speech, Language, and Hearing Research*, 19(1), 168-180. 1976.

[12] D. Byrd. "Preliminary results on speaker- dependent variation in the TIMIT database," *The Journal of the Acoustical Society of America* 92, no. 1: 593-596. 1992.

[13] C. G. Henton and R. A. W. Bladon. "Breathiness in normal female speech: Inefficiency versus desirability." *Language & Communication* 5, no. 3: 221-227. 1985.

[14] D. H. Klatt and L. C. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *The Journal of the Acoustical Society of America* 87, no. 2: 820-857. 1990.

[15] E. A. Strand and K. Johnson. “Gradient and Visual Speaker Normalization in the Perception of Fricatives,” In *KONVENS*, pp. 14-26. 1996.

[16] P. Bertelson, V. Jean, and D. G. Béatrice. “Visual recalibration of auditory speech identification: a McGurk aftereffect.” *Psychological Science* 14, no. 6: 592-597. 2003.

[17] K. Johnson, E. A. Strand, and M. D'Imperio. “Auditory–visual integration of talker gender in vowel perception,” *Journal of Phonetics* 27, no. 4: 359-384. 1999.

[18] E. A. Strand. “Uncovering the role of gender stereotypes in speech perception,” *Journal of language and social psychology*, 18(1), 86-100. 1999.

[19] P. Boersma and D. Weenink. Praat-doing phonetics by computer, online resource: http://www.praat.org.

[20] L. Terango. "Pitch and duration characteristics of the oral reading of males on a masculinity-femininity dimension," *Journal of Speech, Language, and Hearing Research* 9, no. 4: 590-595. 1966.