TAL2018, Sixth International Symposium on Tonal Aspects of Languages
18-20 June 2018, Berlin, Germany

Voice Gender Effect on Tone Categorization and Pitch Perception

Wei Lai

Department of Linguistics, University of Pennsylvania, United States

weilai@sas.upenn.edu

Abstract

Current research on Cantonese tone perception showed that listeners were able to adjust their categorization boundaries between level tones in F_0 space depending on the voice gender [1]. The present study followed up on the voice gender normalization effect by replicating the tone identification experiment with an improved design, and further investigating whether the voice gender would affect the low-level pitch sensation in a parallel manner at a prelexical stage. Results from tone identification and pitch perception by Cantonese listeners showed that stimuli presented in male voices would not only give rise to identification of a higher tone, but also perception of a higher pitch, compared to stimuli of the same F_0 in female voices. Moreover, the magnitude of this effect on tone identification was significantly correlated with that on pitch perception for individual Cantonese listeners. However, pitch comparison behaviors of English listeners did not show obviously different patterns under different voice gender conditions. Implications were discussed with regards to the integration of indexical cues in phonological categorization and prelexical processing, and its interaction with individual language experience.

Index Terms: tone identification, pitch perception, F_0 normalization, gender voice, Cantonese

1. Introduction

Fundamental frequency (F_0) varies considerably across speakers. The interspeaker variation in F_0 is most extremely reflected by the gender difference: Male adults tend to have a low average F_0 (100-125 Hz) and narrow F_0 range (70-200 Hz) while female adults tend to have a high average F_0 (180-220 Hz) and wide F_0 range (140-400 Hz) [2]. For a tone language that uses F_0 to contrast lexical meanings, interspeaker F_0 variation may blur the acoustic boundary between tone categories and give rise to perceptual ambiguity. For example, a phonologically high tone produced by a male could be equivalent in F_0 to a low tone produced by a female [3].

In order to resolve this ambiguity, listeners would implement a process of speaker F_0 normalization before tone categorization, by estimating the location of an F_0 in the speaker's F_0 range based on external or contextual speech materials [4, 5, 6, 7, 8, 9]. [7] demonstrated the role of F_0 normalization in the perception of three level tones (high, mid, low) in Cantonese, by showing that the accuracy of tone identification increased from 48.6% to 80.3% when the presentation of isolated level tones was blocked by talker rather than mixed across talkers. [8, 9] reported that raising or lowering the F_0 on the sentential context would change the perception of an identical stimulus from a mid-tone word to a low-tone or a high-tone word in Cantonese.

Moreover, listeners' ability to locate a F_0 can be still more effective, such that context or prior knowledge of a speaker's F_0 is not mandatory. For example, listeners are still able to recognize the tone on the first few syllables they hear from an unknown speaker, before much of the contextual materials becomes available. [3] showed that Mandarin listeners could distinguish tones starting with a relatively high pitch (High, Falling) from those starting with a low pitch (Low, Rising) with only a fricative plus the first six glottal pulses of the vowel truncated from a /sa/ syllable. This finding indicates that listeners must somehow be able to implement a degree of speaker F_0 normalization merely based on voice cues, when contextual materials are not available for F_0 range estimation.

The above observation is echoed by another line of research on voice-dependent identification of F_0 location in range [10, 11]. [10] showed that even with very brief voice samples of 500 ms isolated vowels, English-speaking listeners could reliably locate its F_0 in the speaker's F_0 range, without external context or prior knowledge to the speaker's voice. [11] replicated this finding with both English-speaking listeners and Mandarin-speaking listeners, and further found that listeners' decisions about the location of F_0 in range were dependent on the sex of the speaker. [11] propose that voice cues are exploited in an indirect way for normalization purposes: Listeners made inference about the sex of the speaker based on voice cues, and then implemented F_0 normalization with their experience-based knowledge of cross-sex F_0 ranges, rather than individual F_0 ranges of the presented speakers.

The point that the information of speaker gender can be revealed from speech is supported by substantial acoustic and perception studies whose results indicate that male and female voices are acoustically distinct [12, 13, 14, 13, 15, 16] and perceptually distinguishable [11, 17, 18, 19, 20]. Acoustically, female voices tend to show greater open quotient by H1-H2 [12], steeper spectral tilt by H1-A3 [13, 14], and more aspiration noise by H1-A3 and H1-A1 [13], as well as higher vowel formant frequencies [15, 16]. It is reported that female and male voices are distinguishable even merely with materials of brief non-speech vocalizations [14, 21, 22]. Among the above cues, formant frequencies [19], glottal opening [17], aspiration noise [18] and spectral slope [20] are also frequently reported to be useful for the perception of speaker gender, along with F_0.

2. Background

Incorporation of speaker gender in speech normalization has been well established at the segemental level in the perception of fricatives [23] as well as vowels [24], by both auditory and visual modalities. [1] extended this vein of research to the perception of Cantonese level tones and found that listeners were capable of compensating for the F_0 difference between gender based on the voice in tone categorization.

Cantonese has six contrastive lexical tones (high-level, high-rising, mid-level, low-falling, low-rising, low-level) and three allotones of level tones on syllables ending with /p t k/. The three level tones, high (e.g., /ji55/, "doctor"), mid (e.g., /ji33/, "meaning"), and low (e.g., /ji22/, "son"), mainly contrast

10.21437/TAL.2018-21

in pitch height, while duration is not very relevant in the production and perception of these tones [25]. [1] reported that the level tone categorization along an F_0 continuum splits as a function of the voice gender on the tone-bearing unit: As shown in Fig 1, the overall distribution for each of the three responses (high/mid/low) shifts towards the low F_0 end under the male voice condition and towards the high F_0 end under the female voice condition. In other words, listeners tend to hear a higher tone for stimuli presented in a male voice compared to stimuli presented in a female voice with equalized F_0.

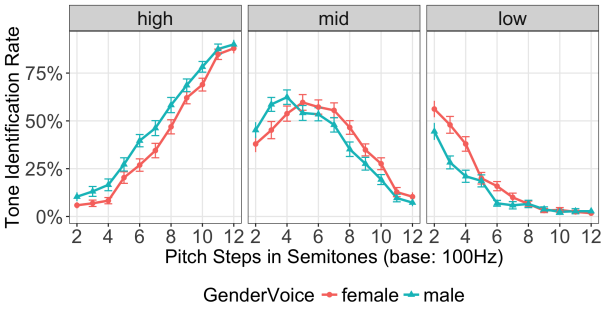

Figure 1: *Cantonese tone identification in gender voices [1]*

However, [1] has a few limitations. One concern is that stimuli in female voices with a low F_0 are not as ideal in naturalness as female voices with a high pitch and male voices overall, according to the naturalness ratings. Another concern is that the results of [1] might be interfering with an on-going merger between the mid tone and the low tone among some young Cantonese speakers [26]. Besides, the participants of [1] are constricted to the population of middle school students, and it is not clear whether the observed pattern can be replicated to listeners from broader age groups.

The first goal of the present study is to evaluate whether the voice gender effect on tone identification reported in [1] still holds after exclusion of the low-F_0 stimuli associated with perceived unnaturalness and confusion from tone mergers. A second and more principal goal is to investigate whether the voice gender also could lay an effect on sensation at a prelexical level, i.e., the perception of pitch height *per se*, and whether the magnitude of voice gender effect on pitch perception (if any) correlates with that on tone perception. The third goal is to explore whether listeners' inclination of integrating speaker gender in tone and pitch perception is shaped by their experience with tone languages, by comparing their behavior in perception with the behavior of English-speaking listeners.

3. Methods

This experiment first partially replicated the tone identification task in [1] on a group of native Cantonese listeners, and then explicitly asked them for meta-linguistic feedback with regards to whether they had been thinking of the speaking gender. After that, a pitch comparison task was conducted in which listeners were presented with two adjacent F_0 peaks and asked to identify which peak has a higher F_0 maximum. Another group of native English listeners were recruited to participate in the pitch comparison task only, in order to evaluate the voice gender effect on the pitch perception of non-tone language listeners.

3.1. Gender voices

This experiment uses the same 2 prototypical male voices and 2 prototypical female voices as were used in [1]. These voices were selected from a pool of 20 voices elicited from 10 male speakers and 10 female speakers who are either Cantonese or Mandarin speakers, by recording each of them produce a "yi" (/ji/) sound with a high tone. These sounds were then superimposed with three levels of pitch height and were rated on gender prototypicality and naturalness by 20 native Cantonese listeners. 2 voices from each gender that achieved the highest rating in either gender prototypicality or naturalness averaged across pitch levels were selected. For more details of the process and results of the voice selection process, as well as the acoustic properties of the selected voices, see [1].

3.2. Manipulation

For tone identification, the 4 voices were each superimposed with an 6-step pitch continuum from 7 semitones (149 Hz) to 12 semitones (200 Hz) with an interval of 1 st on the base of 100 Hz, using Linear Predictive Coding (LPC) algorithm in Praat. All the stimuli were then normalized to the duration of 0.45 second and the intensity of 60 dB.

For pitch perception, only 1 male voice and 1 female voice that achieved the highest rating for gender stereotypicality were used. The voices were concatenated either with each other or with a copy of itself, which results in stimuli of two adjacent "yi" sounds presented in four combinations of gender voices: female-male, male-female, female-female and male-male. Then, modeled on [27], the two concatenated "yi" sounds were each manipulated to bear a pitch peak in such a way that the bases of the peaks were kept 6 st (141.4 Hz) constantly, the maximum pitch on the first peak was kept 10 st (178.1 Hz) constantly, and the maximum pitch on the second peak was varied from 8 (158.7 Hz) st to 12 st (200 Hz) with an interval of 0.5 st (around 10 Hz). The duration and intensity of each peak are normalized to 0.5 second and 70 dB.

3.3. Participants

The Cantonese-speaking participants were recruited from Guangzhou, China, to attend the experiment online for a payment of 40 CYN. Among the 39 participants that had started the experiment, 29 of them have finished both the tone identification task and pitch the comparison task. They are 5 males and 24 females aged from 18 to 43 years old, all reporting themselves as native Cantonese speakers with a normal hearing.

Additionally, 39 English-speaking participants were recruited from the undergraduate population of University of Pennsylvania to attend the pitch perception task online to gain credits. They are 7 males and 32 females, aged from from 18 to 21, all reporting themselves to speak native English and has a normal hearing, without any experience of tone languages.

3.4. Procedure

The Cantonese listeners first completed a tone identification task with a two-alternative forced-choice paradigm. Isolated "yi" sounds in 4 voices at 6 F_0 steps were presented with 4 repetitions in a single block, and listeners were instructed to identify each word they hear as either a high-tone word (医) or a mid-tone word (意). They were allowed to hear the sound as many times as they wanted, and were instructed to respond by clicking on the button with the correct Chinese character on the screen.They were told beforehand that the number of responses

for each character may or may not be equal. There were 3 practice trials in the beginning of the block for familiarization purposes to which the responses were excluded in the data analysis. Each participant completed 6 step * 4 voice * 4 repetition = 96 trials. After the task, participants were asked to report whether they had been speculating speakers' gender from voices during tone identification, by choosing from "yes" "occasionally" or "no". They were also asked to report the number of voices they heard in total, as well as the number of male voices and female voices separately.

The Cantonese listeners then performed a pitch comparison task, in which they heard the stimuli of two pitch peaks superimposed on two adjacent "yi" sounds, and were instructed to identify which peak has a higher pitch maximum for each stimulus. The two adjacent "yi" sounds were in 4 combinations of voices, and the F_0 maximum on the second "yi" sound were manipulated into 9 steps. Therefore each participant completed 9 step * 4 voice combination * 5 repetition = 180 trials within a single block with the order of the items randomized. There were also 3 practice trials at the beginning of the block that were not counted towards their performance. The English listeners performed the same pitch perception task using identical materials and procedure with one exception: they were asked whether they had been speculating the speaker gender and the number of gender voices they heard after pitch perception, instead of before it. At the end of the experiment, participants from both groups were asked to report their background of music training by choosing from "professional", "amateur" and "no training".

4. Results

4.1. Tone and pitch perception by Cantonese listeners

Fig. 2 shows the percentage of high tone identification for stimuli along the pitch continuum under different voice gender conditions. The pattern resembles [1] in showing a split in the distribution of responses as a function of voice gender: It shifts towards the low F_0 end under the male voice condition and towards the high F_0 end under the female voice condition.

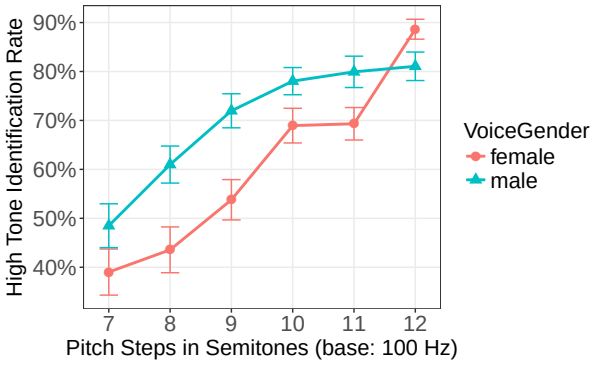

Figure 2: *Voice gender effect on Cantonese tone identification*

A generalized linear mixed-effect model was carried out, with Response (High/Mid) as the dependent value, VoiceGender and PitchStep as independent variables, PitchStep x VoiceGender as the interaction, and Participant and Voice as random factors. The result shows that Pitch ($\beta = 0.51$) and Gender ($\beta = 1.87$) as factors are each significant ($p < .001$), and their interaction is also significant ($p < .01$). This indicates that an increase in F_0 will raise the probability of a high tone response by 0.51 on the log odds, and presentation of a stimulus in a male voice would raise the probability of a high tone response by 1.87 on the log odds compared to the averaged probability across both genders. Simply put, stimuli in a male voice are more likely to be identified as a high-tone word than stimuli in a female voice with equalized F_0.

Fig. 3 shows the mean probability of reporting a higher second F_0 peak in the pitch comparison task as a function maximum pitch on the second peak, conditioned by voice gender combinations.

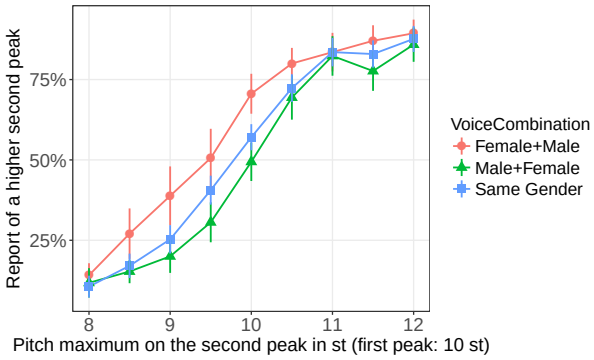

Figure 3: *Voice gender effect on pitch height perception*

Clearly, a female-male voice combination causes more perception of a higher second F_0 peak whereas a male-female voice combination causes less perception of a higher second peak at F_0 each step. The higher second peak report for stimuli concatenated of the same gender voice (averaged across male-male and female-female) falls in between. This indicates that a male voice would lead to perception of a higher pitch compared to a female voice, such that a male voice coming before a female voice leads to perception of a higher first peak whereas one coming after female voice leads to perception of a higher second peak. A mixed-effect logistic regression model that uses VoiceCombination and PitchStep as independent variables to predict Response, with Participant listed as the random factor, shows that PitchStep ($\beta = 0.11$) and VoiceCombination ($\beta = 0.61$) are each significant ($p < .001$).

Moreover, the magnitude of the voice gender effect on tone identification and on pitch perception were correlated at the in-

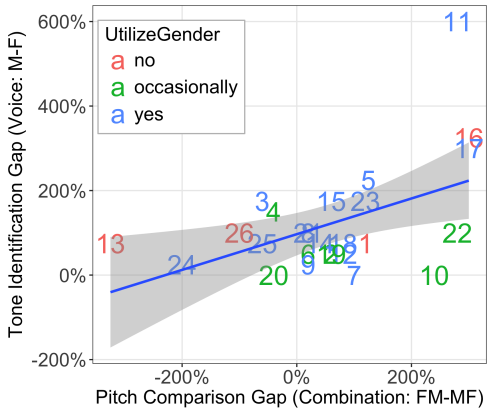

Figure 4: *Correlation between the sizes of voice gender effect on tone and pitch perception, and self-reported gender awareness*

dividual level. For each participant, Fig. 4 shows the magnitude of the voice gender effect on tone identification on the y-axis, as indexed by the difference in high tone identification rate between male-voice and female-voice conditions, and the magnitude of the voice gender effect on pitch perception on the x-axis, as indexed by the difference in second peak report between Female-Male and Male-Female conditions. The fitting line indicates a positive correlation between the two effect sizes, which turns out to be statistically significant ($R = 0.42$, $p = 0.02$).

Fig. 4 also shows the individual responses to the question whether participants had been thinking of the speaker's gender by the color code. Somewhat surprisingly, 25 participants out of 29 had reported "yes" or "occasionally", and only 4 participants (i.e., 1, 15, 25, 27) reported that they never thought of speaker gender at all. This indicates that an overall high level of gender awareness involved is in tone identification. On the other hand, whether the individual behavior in the tone or pitch perception is affected by voice gender or not does not seem to be predictable from their self-reported speaker gender awareness.

4.2. Comparison with English listeners in pitch perception

Fig. 5 shows the mean probability of reporting a higher second peak by English listeners at the 9 steps of F_0 maximum on the second peak. Contrary to the pattern in Fig. 5, the aggregate responses do not split as a function of voice gender combinations; Instead, a general overlap is observed across voice gender conditions. According to a mixed-effect logistic regression model with VoiceCombination and PitchStep as independent variables, Response as dependent variables, and Participant as the random factor, PitchStep is a significant predictor for Response ($\beta = 0.15$, $p < .001$), and the factor of VoiceCombination ($\beta = -0.18$, $p = 0.497$) is also marginally significant.

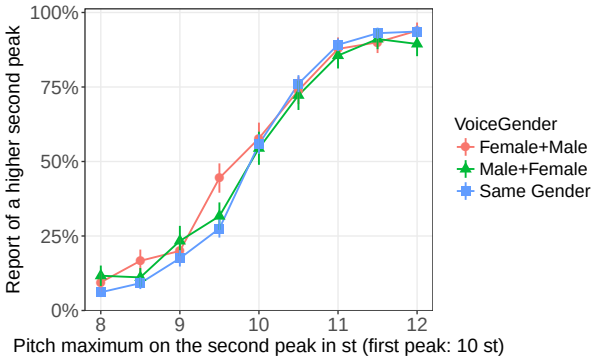

Figure 5: *Pitch perception by voice gender conditions*

The English speakers also showed an overall high level of awareness for the speaker gender by their metalinguitic reports. The majority of them reported that they had been thinking about the speaker's gender during the task by 25% "yes" responses and 45% "ocassionally" responses. The additive proportion of the two positive responses is lower than that of Cantonese listeners, but still within a comparable range. Note that the comparison on the outcomes of pitch perception and gender awareness report between language groups is not perfect: The English listeners were inquired of the speaker gender for materials from a non-speech task (pitch perception) with exposure to foreign voices and sounds, while the Cantonese speakers were inquired for materials from a speech task (tone identification) with exposure to native sounds and voices. The asymmetric perception behaviors shown across language groups should be partially attributed to the different task set-ups for the two groups of participants.

The degree to which individuals attend to F_0 in pitch perception is also reported to be affected by their musicality [28], so the music background of participants were also investigated and compared. Table 1 lists the distribution of participants from the two language groups on three musicality levels ("professional", "amateur", "no training"). According to 1, most Cantonese participants (71%) reported themselves to have never received any music training, while most English participants (61%) reported that they had amateur training in music before. This raised a concern that the different pitch perception behaviors reflected in Fig. 3 and Fig. 5 might also be caused by a bias of musicality in participant sampling.

Table 1: *Distribution of participants in music training*

Music training	**Cantonese listeners**	**English listeners**
Professional	7%	0%
Amateur	21%	61%
No training	72%	39%

5. Discussion and Conclusions

This study presents combined results of tone identification and pitch perception from Cantonese speakers, which shows that the voice gender would lay an effect on tone categorization and pitch perception in a parallel manner: A male-sounding voice would induce perceptual bias towards either a higher tone or a higher pitch compared to a female-sounding voice of the same F_0. This result verified the robustness of the voice gender effect on tone categorization currently reported in [1], by replicating the same tone boundary shift triggered by voice gender with a slightly different experiment design and another group of participants. Besides, the gap in tone identification correlates nicely with the gap in pitch perception responses between different gender voice conditions for individual listeners, which favors a view in which the experience of speech perception might loop back to shape the prelexical sensation of raw acoustic inputs.

The study also presents results of pitch perception from English listeners, which does not show an obvious difference as a function of voice gender conditions. This observation shed lights on an intriguing possibility that listeners' sensation bias for nonspeech materials might be affected by their language-specific knowledge. In this particular case, English listeners might not need to use voice gender cues for F_0 normalization as frequently as Cantonese listeners in phonological categorization, such that the voice gender cues interfere less with the F_0 perception for English listeners than for Cantonese listeners. However, this possibility is not fully confirmed by the present study, due to limitations such as musicality bias in participant sampling and different degrees of familiarity with the voices and sounds of the stimuli. Exploration on this possibility under more careful control is worth pursuing by further studies.

6. Acknowledgements

The author is very grateful to Professor Mark Liberman, Professor Jianjing Kuang and Professor Meredith Tamminga, for their helpful discussion and feedback.

7. References

[1] W. Lai, "Auditory-visual integration of talker gender in cantonese tone perception," *Proc. Interspeech 2017*, pp. 664–668, 2017.

[2] M. Biemans, *Gender variation in voice quality*. Netherlands Graduate School of Linguistics, 2000.

[3] C.-Y. Lee, "Identifying isolated, multispeaker mandarin tones from brief acoustic input: A perceptual and acoustic study," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1125–1137, 2009.

[4] J. Leather, "Speaker normalization in perception of lexical tone." *Journal of Phonetics*, 1983.

[5] R. A. Fox and Y.-Y. Qi, "Context effects in the perception of lexical tone/," *Journal of Chinese Linguistics*, pp. 261–284, 1990.

[6] C. B. Moore and A. Jongman, "Speaker normalization in the perception of mandarin chinese tones," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1864–1877, 1997.

[7] P. C. Wong and R. L. Diehl, "Perceptual normalization for inter- and intratalker variation in cantonese level tones," *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 2, pp. 413–421, 2003.

[8] A. L. Francis, V. Ciocca, N. K. Y. Wong, W. H. Y. Leung, and P. C. Y. Chu, "Extrinsic context affects perceptual normalization of lexical tone," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1712–1726, 2006.

[9] G. Peng, C. Zhang, H.-Y. Zheng, J. W. Minett, and W. S.-Y. Wang, "The effect of intertalker variations on acoustic–perceptual mapping in cantonese and mandarin tone systems," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 2, pp. 579–595, 2012.

[10] D. N. Honorof and D. H. Whalen, "Perception of pitch location within a speaker's f0 range," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2193–2200, 2005.

[11] J. Bishop and P. Keating, "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex," *The Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1100–1112, 2012.

[12] E. Pépiot, "Voice, speech and gender:. male-female acoustic differences and cross-language variation in english and french speakers," *Corela. Cognition, représentation, langage*, no. HS-16, 2015.

[13] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *The Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 1064–1077, 1999.

[14] D. G. Childers and K. Wu, "Gender recognition from speech. part ii: Fine analysis," *The Journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1841–1856, 1991.

[15] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *The Journal of the acoustical society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[16] R. O. Coleman, "Male and female voice quality and its relationship to vowel formant frequencies," *Journal of speech and hearing research*, vol. 14, no. 3, pp. 565–577, 1971.

[17] ——, "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice," *Journal of Speech and Hearing Research*, vol. 19, no. 1, pp. 168–180, 1976.

[18] 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*, vol. 87, no. 2, pp. 820–857, 1990.

[19] Y. Lavner, I. Gath, and J. Rosenhouse, "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," *Speech Communication*, vol. 30, no. 1, pp. 9–26, 2000.

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

[21] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *The journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[22] J. Kreiman, "Listening to voices: theory and practice in voice perception research," *Talker variability in speech processing*, pp. 85–108, 1997.

[23] E. A. Strand and K. Johnson, "Gradient and visual speaker normalization in the perception of fricatives." in *KONVENS*, 1996, pp. 14–26.

[24] K. Johnson, E. A. Strand, and M. D'Imperio, "Auditory–visual integration of talker gender in vowel perception," *Journal of Phonetics*, vol. 27, no. 4, pp. 359–384, 1999.

[25] P. C. Wong and R. L. Diehl, "The effect of reduced tonal space in parkinsonian speech on the perception of cantonese tones," *The Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 1246–1246, 1999.

[26] P. P.-K. Mok and P. W.-Y. Wong, "Perception of the merging tones in hong kong cantonese: Preliminary data on monosyllables," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[27] J. Kuang and M. Liberman, "The effect of spectral slope on pitch perception." in *INTERSPEECH*, 2015, pp. 354–358.

[28] J. Kuang, "The effect of musicality on cue selection in pitch perception," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3818–3818, 2017.