CONTRIBUTION OF VOICE FUNDAMENTAL FREQUENCY AND FORMANTS TO THE IDENTIFICATION OF SPEAKER'S GENDER

Siu-Fung Poon & Manwa L. Ng

Speech Science Lab., Division of Speech & Hearing Sciences, University of Hong Kong, Hong Kong
poonsiufung@gmail.com; manwa@hku.hk

ABSTRACT

Identification of gender from speech sounds has been found to rely on speakers' voice fundamental frequency (F_0) and formant frequencies. The present study aimed at examining the contribution of F_0 and formants to the correct detection of speaker's gender. Results revealed that F_0 is the primary cue for gender perception and listeners showed a higher accuracy in identifying male's than female's voices.

Keywords: gender identification, fundamental frequency, formants, Cantonese

1. INTRODUCTION

According to the source-filter theory, human acoustic speech output is a product of the sound source (vocal fold vibrations during phonation) and the filter (vocal tract resonances). As an important attribute of sound source, voice fundamental frequency (F_0) of any speech signal is the physical measure of the rate of vocal fold vibration, and it is perceptually correlated with the perceived pitch [5]. During speech production, the source signal is modified by vocal tract resonance, resulting in some frequencies being amplified (formants), while other frequencies suppressed [5].

The anatomical differences between adult males and females with respect to source (vocal folds) and filter (vocal tract) have been well documented. According to Titze [9], male vocal folds are generally larger than female ones by approximately 60%, yielding a lower F_0 in males. In addition, the male vocal tract is about 15% longer than females', corresponding to the generally lower formants in male speech [1].

Given such sexual dimorphism in human speech characteristics, researchers have attempted to investigate the role of F_0 and formant frequencies in the perception of speakers' gender based on perceptual testing (e.g. [3, 5, 8, 10]). In Coleman [2], Gelfer and Mikos [4], and Whiteside [10], stimuli were divided into four types: (1) male F_0 paired with male formants; (2) female F_0 paired with female formants; (3) male F_0 paired with female formants; and (4) female F_0 paired with male formants. Findings in these studies were consistent: when F_0 and formant frequencies pointed to the same gender (1 & 2), the correct gender identification rate was higher in the male case than in the female one. However, when F_0 and formant frequencies conflicted with each other (i.e. 3 & 4), discrepant findings regarding gender identification were observed. Coleman [2] found that such stimuli tended to be perceived as males. Gelfer and Mikos [4] and Whiteside [10] suggested that listeners relied on F_0 more than formants for speakers' gender identification tasks.

Despite the interesting findings reported by Coleman [2], questions are raised regarding the methodology. In particular, connected speech produced by normal speakers using laryngeal vibrators (an equipment with which voice F_0 can be manipulated) as sound source was used in Coleman's study. Gelfer and Mikos [4] commented that the connected speech materials could have provided additional cues other than F_0 and formant frequencies, such as intonation and stress pattern, to aid gender identification. In addition, the flat intonation contour associated with the laryngeal vibrators may have favored the perception of maleness.

Unlike Coleman [2], Gelfer and Mikos [4] and Whiteside [10] made use of synthesized vowels for their perceptual judgment tasks, and both studies yielded similar conclusions: listeners tended to make use of F_0 more than formants for gender identification. Yet, Gelfer and Mikos used speech stimuli that appeared to be better designed. Whiteside synthesized the speech stimuli using averaged formant data extracted from vowel segments of sentences produced by six speakers. The use of short synthesized vowels (100 ms for long vowels and 50 ms for short vowels) in the study with varying F_0 contours could have affected listeners' judgment of gender [10]. Gelfer and Mikos modified the way speech stimuli were synthesized. They used individual formant data

extracted from the sustained vowels to synthesize one set of speech stimuli. Moreover, the synthesized vowels had a longer duration (3 seconds) and flat F_0 contour.

In contrast, based on five vowels produced by a single male speaker, Smith and Patterson [8] synthesized 245 speech stimuli over a wide range of F_0-formants combinations for their perceptual experiment. Listeners judged the size and age/sex (i.e. man, woman, boy, or girl) for each stimulus. They concluded that both F_0 and formant frequencies contributed to the perception of speaker's sex and age. Such a design allowed researchers to better demonstrate the changes of speakers' gender perception along variations of F_0 and formants.

Studies of English-speaking populations have demonstrated evidence supporting the roles of F_0 and formant frequencies in speaker's gender perception, but there is no consensus on the relative importance of the two acoustic cues to gender identification. In addition, studies on the Cantonese-speaking population are lacking. There is no information regarding if and how language affects the way speakers' gender is perceived by listeners. The main purpose of the present study is to find out *the relative contributions of F_0 and formant frequencies to perception of speakers' gender in Cantonese-speaking population*. In the study, speech stimuli were synthesized in a way similar to that used by Smith and Patterson [8].

2. METHOD

2.1. Participants

Ten male and 10 female adult native Cantonese speakers (age: *M* = 21.36 years, S.D. = 1.19, range = 20-26 years) were recruited as speakers. All speakers reported having no speech, language and hearing problems. For the perceptual experiment, 11 male and 17 female adult native Cantonese speakers (age: *M* = 19.79 years, S.D. = 1.03, range = 19-24 years) were recruited. All participants were university students who had no known speech, language and hearing problems.

2.2. Speech tasks and recording procedure

The recordings took place in a sound treated room of the Speech Science Laboratory of University of Hong Kong. During the experiment, each speaker was instructed to sustain the syllable /a/ at the high-level tone for approximately five seconds at a comfortable loudness level. The speech samples were recorded by using a high-quality microphone (SM58, Shure) via a preamplification unit (PreMobile USB, M-Audio). During the recording, a mouth-to-microphone distance of approximately 10 cm was maintained. A brief practice period was provided to participants to familiarize themselves with the recording environment. Audio signals were digitized with a sampling rate of 20 kHz and quantization rate of 16 bits/sample by using the Praat software. The digitized signals were stored in a computer for later analyses.

2.3. Acoustic analysis

A three-second segment was extracted from the medial portion of each recorded vowel and then analyzed by using Praat for F_0, formant frequencies (F1 to F3) and formant bandwidths (B1 to B3). One male and one female speaker who had the mean formant frequency (averaged across F1 to F3) closest to the corresponding group mean formant frequencies were selected. Their data were used for the formant scaling across gender and synthesis of stimuli for subsequent perceptual experiment. The ratios of mean F1, F2 and F3 between these two speakers were calculated respectively. Then a composite formant frequency scale factor (female/male) was calculated by averaging the three ratios, which was found to be 1.20.

2.4. Stimuli synthesis

Formant data of the male and female speakers were used as the basis for creating two sets of synthesized vowels. Vowels of one male and one female were used to synthesize the stimuli because it is not known if gender of the original vowels can affect results of gender perception. Using Praat, the male speakers' F1 to F3 were multiplied by 10 scale factors from 1 to 1.20 and F_0 was scaled to 10 values within 100 to 250Hz, creating 100 synthesized vowels (10 formant frequency values x 10 F_0 values) each of 3-second long. The vowels were then duplicated to form a set of 200 stimuli (the "male stimuli"). Similarly, another set of 200 stimuli (the "female stimuli") was synthesized using the female vowel, except that the formant frequencies were multiplied by 10 scale factors from 1 to 0.83 (i.e. 1/1.20). The scaling procedure aimed to simulate F_0 and formant frequencies that are male-appropriate, female-appropriate or gender-ambiguous. Upon completion of such process, 400 stimuli with different combinations of F_0 and formant frequencies were prepared.

2.5. Perceptual experiment

The two sets of synthesized vowels were presented to the listeners in two separate sessions. In each session, the listeners were seated in groups in a sound-treated room and listened to the stimuli presented in a randomized order via high-quality loudspeakers. For each stimulus, they were instructed to judge whether the speaker was a male or female. In the case of ambiguity, they were asked to guess. Upon listening to a stimulus, the listeners circled the gender they perceived on an answer sheet provided. An inter-stimuli pause of about five seconds was introduced to provide sufficient time for the listeners to complete the judgment task.

2.6. Intra-listener reliability

All stimuli were presented twice and perception results obtained from the first and second presentations were used to calculate the intra-listener reliability. An average reliability of 83.13% was found, indicating that listeners' perception was consistent and reliable.

3. RESULTS

Results of gender identification of male and female stimuli are shown in Figs. 1 and 2 respectively. Specifically, Figs. 1 and 2 respectively show the percentages of male stimuli being perceived as male voice and female stimuli being perceived as female voice over all F_0-formant combinations (thus the percent correct gender identification). For both male and female stimuli, a clear cutoff along the F_0 axis can be identified, but not along the formant axis. By using extrapolation, the cutoff F_0 for 75% correct gender identification was found to be 162.01 Hz and 204.97 Hz for male and female stimuli respectively. In general, both figures reveal a trend that the rate of the stimuli being perceived as male (male identification) decreased with increasing F_0. The opposite pattern is observed for female identification: female identification increased with increasing F_0. Yet, identification rate of either gender showed little or inconsistent change as formants changed. The identification contours appear to be smooth towards the two ends of F_0 (i.e., near 100 Hz and 250 Hz) as compared to medial F_0, indicating that at high or low F_0, gender identification did not change much with formants; whereas when F_0 was set to the middle range, the gender identification rate fluctuated with formant frequencies.

Figure 1: Percent male identification of male stimuli over different fundamental frequency (F_0)-formant combinations.

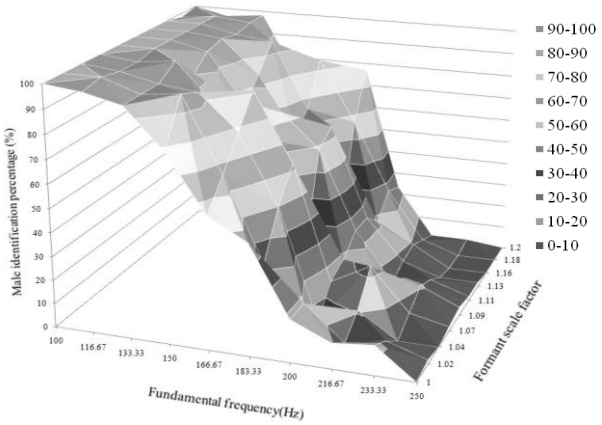
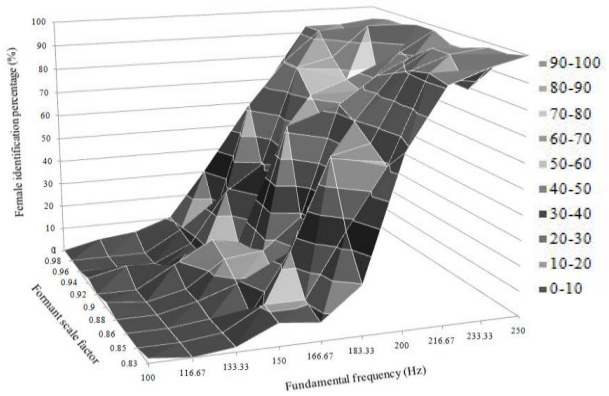


Figure 2: Percent female identification of female stimuli over different fundamental frequency (F_0)-formant combinations.



4. DISCUSSION

The present study attempted to examine the contribution of F_0 and formant frequencies to gender identification. Results showed that the chance of stimuli being perceived as a male voice decreased with increased F_0 (see Figs. 1 and 2). Yet, the effect of formants on gender perception of appeared to be smaller, as indicated by the lack of clear contour in the figures. This implies that listeners mainly depend on F_0, but not formants, to perceive speaker's gender. This is more obvious when F_0 is high or low. According to Fig. 1, for the stimuli perceived as male's voices, listeners showed highly consistent gender judgment at both ends of F_0 (male identification rate ranged from 98.21% - 100% and 0% - 3.57% when F_0 was 100 Hz and 250 Hz, respectively). Cues from F_0 are strong enough for stimuli to be perceived as a particular gender, regardless of formants. It

follows that F_0 is the primary cue for gender perception, with formants being supplementary cues. This finding is consistent with those reported in the studies of Coleman [2], Gelfer and Mikos [4], and Whiteside [10]. Apart from this, the 75% identification cutoff F_0 for male and female stimuli were found to be 162.01 Hz and 204.97 Hz respectively. This implies that when F_0 ranges from 100 Hz to 162.01 Hz and from 204.97 Hz to 250 Hz, listeners can reliably and correctly identify the speakers as males and females respectively. Conversely, when F0 falls within the range of 162.01 Hz to 204.97 Hz, listeners cannot judge speakers' gender reliably. In this case, listeners may make use of suprasegmental cues or just random strategy to judge speakers' gender.

However, the conclusion of F_0 being the major cue for gender perception is inconsistent with Smith and Patterson [8]. They concluded that both F_0 and formants contributed to perception of gender and age. However, the range of formants used for synthesizing stimuli extended to children's range in their study, and only seven data points along this large range of formants were used to synthesize the stimuli. The limited data points and larger range of formants might have over-simplified the possible effect of formants on gender perception. Since both age and gender are investigated together in their study, it is not known whether formants will still be as important if only adult males and females are studied.

The present results also indicate that perception of male stimuli was more accurate than that of female stimuli, when both F_0 and formant cues are not conflicting with each other. This is consistent with findings reported by Coleman [2], Gelfer and Mikos [4], and Whiteside [10]. Male stimuli (F_0 = 100 Hz / male formant) yielded 100% correct identification rate, whereas the female stimuli (F_0 = 250 Hz / female formant) only had 94.64%. As explained by Owren, Berkowitz and Bachorowski [7], low F_0 with low formant frequencies were the distinctive cues for adult males but the F_0-formant frequency combinations for adult females were more diverse. Thus, listeners identify male voices more accurately than female voices. However, the asymmetrical identification results may reveal the limitation of using synthesized vowels as stimuli. In natural speech, suprasegmental cues (e.g. breathiness) also affect the perception of gender. According to Klatt and Klatt [6], female voices are generally more breathy than male voices. When this difference is removed as in synthesized vowels, perception of male gender may be favored.

Concerning the stimuli in the current study, the relationship between male and female formant frequencies may have been simplified. When synthesizing the stimuli, F1, F2 and F3 are multiplied by the same scale factor. Fant [3] pointed out that this simple scaling method cannot accurately demonstrate male-female formant relationship because males have a larger ratio of pharyngeal length to mouth cavity length and larger larynx than females. The use of simple scaling may partially contribute to the unclear effect of formants on gender identification in this study. To improve, further research may scale the formant frequencies independently or a more sophisticated source-filter synthesizer is needed.

5. REFERENCES

[1] Bachorowski, J.A., Owren M.J. 1999. Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *J. Acoust. Soc. Am.* 106, 1054-1063.

[2] Coleman, R.O. 1976. A comparison of the contribution of two voice quality characteristics to the perception of maleness and femaleness in the voice. *J. Speech. Hear. Res.* 19, 168-180.

[3] Fant, G. 1966. A note on vocal tract size factors and non-uniform F-pattern scaling. *http://www.speech.kth.se/qpsr*

[4] Gelfer, M.P., Mikos, V.A. 2005. The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *J. Voice* 19, 544-554.

[5] Kent, R.D., Read, C. 1992. *The Acoustic Analysis of Speech*. San Diego, CA: Singular Publishing Group, Inc.

[6] Klatt, D.H., Klatt, L.C. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.* 87, 820-857.

[7] Owren M.J., Berkowitz, M., Bachorowski, J.A. 2007. Listeners judge talker sex more efficiently from male than from female vowels. *Percept. Psychoph.* 69, 930-941.

[8] Smith, D.R.R., Patterson, R.D. 2005. The interaction of glottal-pulse rate and vocal-tract length in judgments of speaker size, sex and age. *J. Acoust. Soc. Am.* 118, 3177-3186.

[9] Titze, I.R. 1989. Physiologic and acoustic differences between male and female voices. *J. Acoust. Soc. Am.* 85, 1699-1707.

[10] Whiteside, S.P. 1998. The identification of a speaker's sex from synthesized vowels. *Percept. Motor. Skill.* 87, 595-600.