Proceedings of the Institute of Acoustics

THE IDENTIFICATION OF GENDER FROM SYNTHETIC SPEECH

S. P. Whiteside[1] & J. Goodall[2]

1 Department of Human Communication Sciences, University of Sheffield, UK S10 2TA.
2 Department of Speech and Language Therapy, Kendray Hospital, Barnsley, S70 3RD.

ABSTRACT

This paper presents findings on some of the acoustic cues that listeners use to make decisions about the gender of an adult speaker from synthetic speech. Four young adult male and adult female speakers from South Yorkshire, England were chosen for the initial part of this study. They were selected on the basis of criteria which included age, accent and normal speech and hearing. Subsequently recordings of the eight speakers were played to 14 young adult female listeners. They were asked to rate the gender of the speakers. On the basis of this one male and one female speaker were selected to provide data for analysis, synthesis and perception test procedures. Acoustic parameters of the male speaker's data were obtained using a KAY Computerised Speech Lab (CSL) model 4300. These parameters were used as reference values to manipulate the female speaker's acoustic patterns using an LPC parameter manipulation/synthesis program (model 4304, version 2.01) with an editing facility. The female speaker's vowels were edited with either a lowered fundamental frequency and/or lowered formant frequencies. Listeners were asked to judge whether they heard an adult male or and adult female speaker. Results showed that primarily the perception of gender is determined by fundamental frequency. However in some cases the formant frequencies served as cues in the perception of gender particularly when a vowel had high second and third formant frequencies. These findings are presented and discussed in light of previous research.

1. INTRODUCTION

Women are usually perceived as having higher pitched voices than men and this appears to be related to their higher fundamental frequency values [1]. There is widespread evidence from acoustic studies to suggest that F0 is a salient and robust cue that listeners use to judge speaker sex [2, 3, 4, 5]. However, the perceptual salience of the vocal tract resonances should not be underestimated when judging a speaker's gender [6, 7, 8]. This is further supported by Childers and Wu [9] and Wu & Childers [10] who found that formant characteristics contained enough information for accurate automatic gender recognition in the absence of fundamental frequency. That listeners are able to perceive the "pitch" of whispered vowels to judge for speaker sex has been demonstrated by Thomas [11]. From the whispered vowels of a male and female speaker listeners were able to determine the "pitch" (p. 469 [11]) of whispered vowels by comparison with a pure tone. Results showed that the perceived "pitch" of the sets of vowels corresponded almost exactly to the frequencies of F2. The role of F2 as a robust acoustic cue is further substantiated by Childers and Wu [9] who found that it surpassed even F0 as a single feature in the automatic recognition of gender.

2. METHOD

2.1 Subjects

Selection of reference male and female speakers

Four adult male and four adult female subjects were chosen. They were between 20 and 25 years of age. They had lived in South Yorkshire all their lives. They were non-smokers and had no history of any speech or language problems.

The eight speakers were asked to read a set of nonsense words which were recorded. These stimuli were then played to fourteen listeners who were asked to rate gender of the natural stimuli using a scale of 1 to 7 where 1 represented 'maleness' and 7 represented 'femaleness'. The male speaker who had the lowest score and the female speaker who had the highest score were selected as the reference speakers.

Reference male and female speakers

Both the male and female reference speakers were 22 years of age.

2.2 Speech material

The speech stimuli used for the study were based on those used in the Peterson & Barney data [12]. The nine monosyllables used represented a CVC structure of the form /h V d/. The list of words representing this structure were: 'heed', 'hid', head', 'had', 'hawed', 'hod', 'who'd', 'hood' and 'heard'. See Tables 1 and 2 for the vowels represented. The nine syllables were randomised five times to form five word lists. Extra stimuli were added to each list to avoid practice effects.

2.3 Recording procedures

The speakers were presented with the lists of words and asked to read them in their normal speaking voice. All speakers were recorded in an anechoic chamber using a Sony Professional Walkman (WM D6C) and a Sony (ECM 909) microphone.

2.4 Analysis

Ninety stimuli (forty-five stimuli for each speaker) were digitized onto a Kay Computerized Speech Lab (CSL) model 4300 using a sampling rate of 10kHz. The steady state of each vowel in the /hVd/ syllable was edited out for analysis. Fundamental frequency (f0) (using the fundamental frequency contour) and the first four formants (using LPC analysis) were measured at three distinct points along the steady state - the onset, midpoint and offset. This gave five sets of measurements for each of the nine vowels.

2.5 Synthesis

Editing the fundamental frequency (f0) values of the female stimuli

All forty five stimuli representing the female speaker were transferred from the CSL 4300 system to the KAY LPC Parameter Manipulation/ Synthesis Program (5365). Fundamental frequency and formant frequencies were derived through the 5365 module for each stimulus. The fundamental frequency contour was scaled down by half for all forty five vowels using an editing facility. This therefore produced forty five synthetic stimuli with fundamental frequencies representing a typical adult male while preserving the adult female formant frequencies. These synthetic data represented stimulus set A.

Editing the female stimuli to manipulate f0 and formant frequencies
One set of the steady states of nine vowels produced by the female speaker was randomly chosen. These were then edited and resynthesized to generate three groups of stimuli with: 1) lowered f0 with female formant frequency values ; 2) lowered f0 values combined with the female fundamental frequency values and 3) lowered f0 combined with lowered formant frequencies. All lowered values were modelled on the reference male speaker using the onset, midpoint and offset values resulting from the analysis (midpoint values (Hz) are given in Tables 1 & 2 for the f0 and formant frequency data respectively). These twenty seven synthetic data made up stimulus set B.

2.6 Perception tests

Stimulus set A
Stimulus set A was randomised and recorded onto cassette using a Sony Professional Walkman (WM D6C). A ten second gap was left between each stimulus so that listeners had ample time to record their responses. The stimuli were presented to a group of ten listeners who reported having no history of speech, language or hearing problems. The listeners were asked to write down whether they perceived the speaker to be a male or a female. They were also asked to assess how confident they were by using a seven point scale where 1 indicated a guess and 7 represented complete confidence.

Stimulus set B
The method here was identical to that used for simulus set A with the exception that a different group of eleven listeners took part in the perception test.

3. RESULTS AND DISCUSSION

3.1 Analysis
Tables 1 and 2 show the mean, standard deviation and standard error for the fundamental frequency and formant frequencies (F1 to F4) for each of the nine vowels at the midpoint for both the male and female speakers.

Vowel	i	ɪ	ɛ	æ	ɔ	ɒ	ʊ	u	ɜ
Male Mean	119	118	117	113	118	116	117	119	117
Standard dev.	3.4	5.7	2.9	6.0	4.8	5.5	2.5	4.7	6.3
Standard error	1.5	2.6	1.3	2.7	2.2	2.5	1.3	2.1	2.8
Female Mean	214	216	210	212	210	212	220	218	211
Standard dev.	3.6	5.7	3.8	4.9	7.9	8.8	8.0	6.3	11.3
Standard error	1.6	2.6	1.7	2.2	3.5	3.9	3.6	3.2	5.1

Table 1. Mean, standard deviation and standard error values (Hz) for male and female speakers: fundamental frequency.

An overall view of the fundamental frequency results (see Table 1) revealed that the female speaker had a fundamental frequency approximately 1.8 times that of the male speaker. This compares to a scale factor of 1.6 suggested by Titze [13]. In addition there appeared to be a slight declination in the f0 values across the female vowels. The values for the male speaker however, remained relatively constant (this is not represented in Table 1).

Vowel	i	ɪ	ɛ	æ	ɔ	ɒ	ʊ	u	ɜ
F1 Male Mean	324	449	604	672	522	623	417	352	497
Standard dev.	13.1	27.4	17.0	20.6	21.9	31.3	12.0	32.2	13.1
Standard error	5.86	12.3	7.60	9.21	9.79	14.0	6.0	14.4	5.86
F1 Female Mean	454	449	793	884	498	713	514	411	629
Standard dev.	20.1	21.5	57.8	36.9	65.2	89.7	35.7	20.0	32.8
Standard error	8.99	9.62	25.8	16.5	29.2	40.1	16.0	10.0	14.7
F2 Male Mean	2137	1876	1653	1424	1010	1078	1039	1194	1590
Standard dev.	27.7	40.5	36.9	100.9	10.7	21.9	19.6	47.1	20.1
Standard error	12.4	18.1	16.5	45.1	4.79	9.79	9.8	21.1	8.99
F2 Female Mean	2920	2499	2089	1757	1020	1208	1387	1007	1889
Standard dev.	20.1	55.6	69.5	110.9	67.2	101.2	43.9	64.1	19.4
Standard error	8.99	24.9	31.1	49.6	30.1	145.3	19.6	32.1	8.68
F3 Male Mean	2737	2616	2547	2427	2422	2442	2405	2243	2437
Standard dev.	64.7	31.2	27.9	55.7	131	17.3	12.0	91.4	26.8
Standard error	28.9	13.9	12.5	24.9	58.6	7.74	6.0	40.9	12.0
F3 Female Mean	3467	3133	2998	3009	2858	2703	2737	2721	2896
Standard dev.	43.6	36.8	38.4	90.2	79.9	139.4	78.8	132.0	46.4
Standard error	19.5	16.5	17.2	40.3	40.0	62.3	35.2	66.0	20.8
F4 Male Mean	3360	3578	*	3566	3264	3307	3058	3152	3452
Standard dev.	85.3	90.6	*	*	64.7	80.3	41.6	50.2	57.7
Standard error	38.1	40.5	*	*	28.9	35.9	20.8	22.5	25.8
F4 Female Mean	*	*	*	3772	3578	3702	3771	*	3874
Standard dev.	*	*	*	*	56.7	*	*	*	12.0
Standard error	*	*	*	*	25.4	*	*	*	6.0

* missing data

Table 2. Mean, standard deviation and standard error values for male and female speakers: Formant frequencies

The formant frequency values in Table 2 suggest that overall the female speaker's formants are approximately 1.3 times higher in frequency compared to her male counterpart. This pattern is a well established finding [12]. There are however instances where the male had either higher or equal formant frequencies. These were in the first formant frequency of the vowels /ɪ/ and /ɔ/ and the second formant of the vowel /u/. The first and second formant frequencies are plotted in Figure 1 to highlight the larger vowel space of the female speaker.

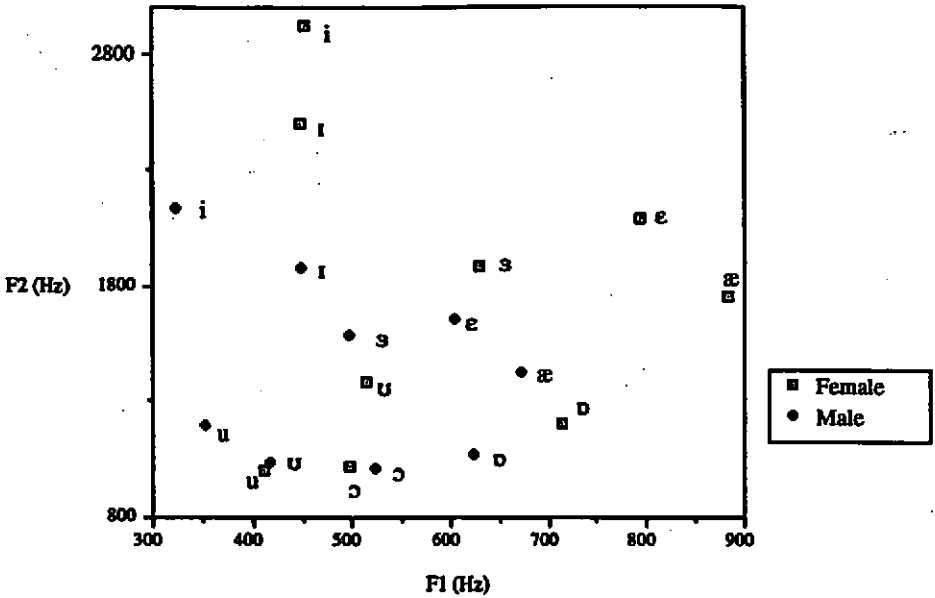

Figure 1. Vowel spaces of male and female reference speakers.

3.2 Perception tests

Stimulus set A
The majority of synthetic tokens in stimulus set A (with a lowered f0) were identified as male. The results are represented graphically in Figure 2 with total confidence levels. The results indicate that those synthetic stimuli that were identified as female were the front vowels /i, ɪ, ɛ/. This suggests that the much higher F2 and F3 values for these vowels were salient cues even in the presence of a lowered fundamental frequency. On the other hand the back vowels were mainly identified as being male. Furthermore, listeners were more confident when making male judgements.

Stimulus set B
Figure 3 graphs the results of identification scores and confidence levels for the stimuli generated using female formant frequency values and male fundamental frequency values. Once again the findings highlight the robustness of fundamental frequency as a cue to a speaker gender where a low fundamental frequency is cueing maleness. In addition listeners appear to be very confident about their decisions. There are a few instances where femaleness judgements are being made. These appear once again to occur with vowels (/i ɛ æ/) where female second and third formant values are substantially higher than the male speaker.

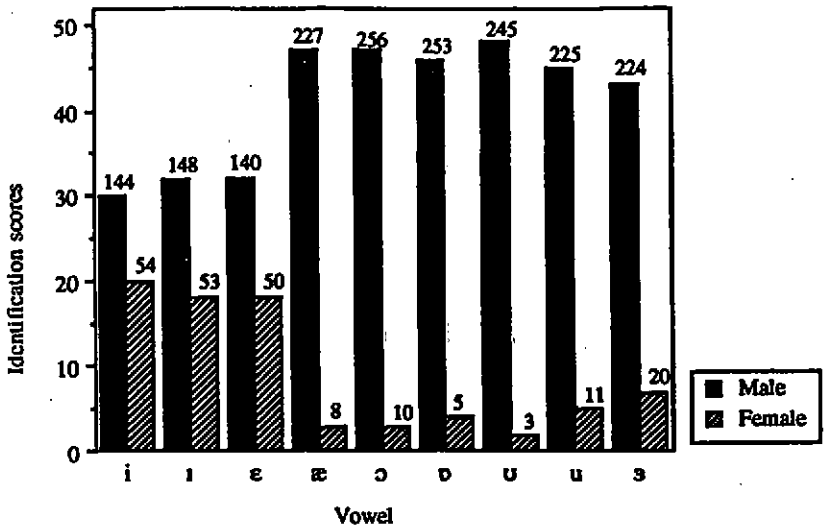

Figure 2. Identification scores for stimulus set A.

Figure 4 illustrates the results of identification scores and confidence levels for the stimuli generated using male formant frequency values and female fundamental frequency values. On the whole listeners were far more confident in judging the stimuli as female. There are a few exceptions to this, the most notable being /ɔ/. Figure 5 gives the results for the stimuli generated using male formant frequency values and male fundamental frequency values. Listeners were highly confident in judging the stimuli as male which is not a surprising finding. The responses of two individuals however suggest that there may be remnants of femaleness in this group of stimuli.

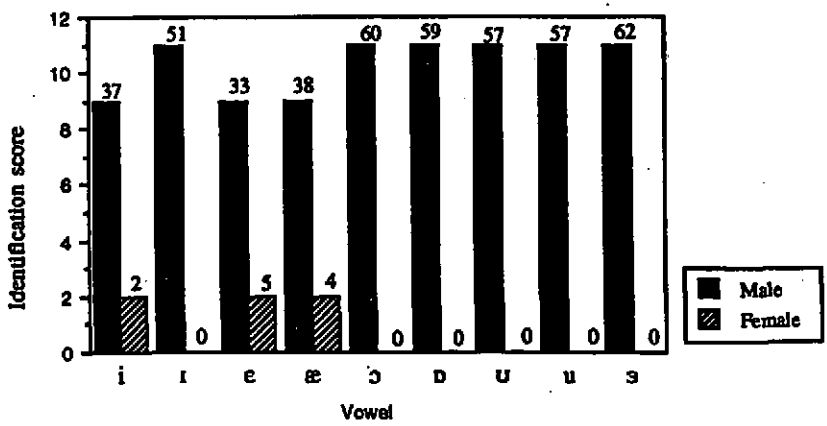

Figure 3. Simulus B identification scores - female formant frequencies and male f0.

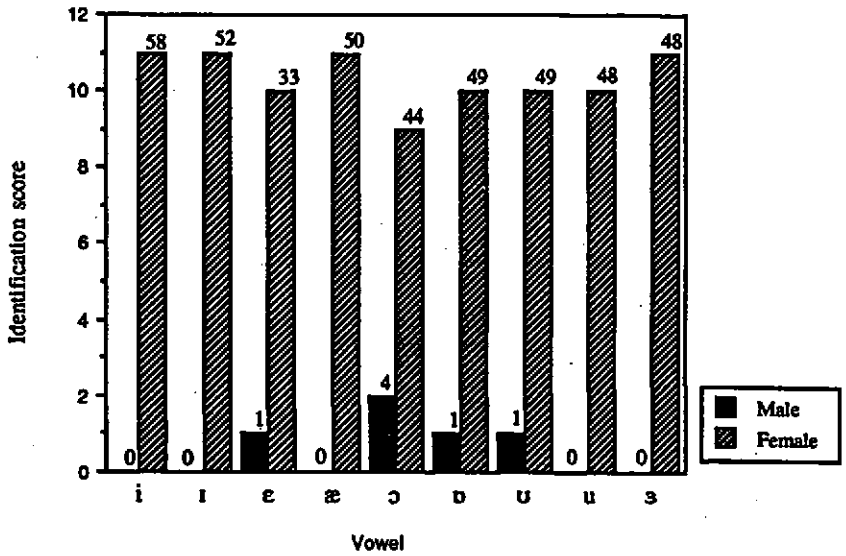

Figure 4. Simulus B identification scores - male formant frequencies and female f0.

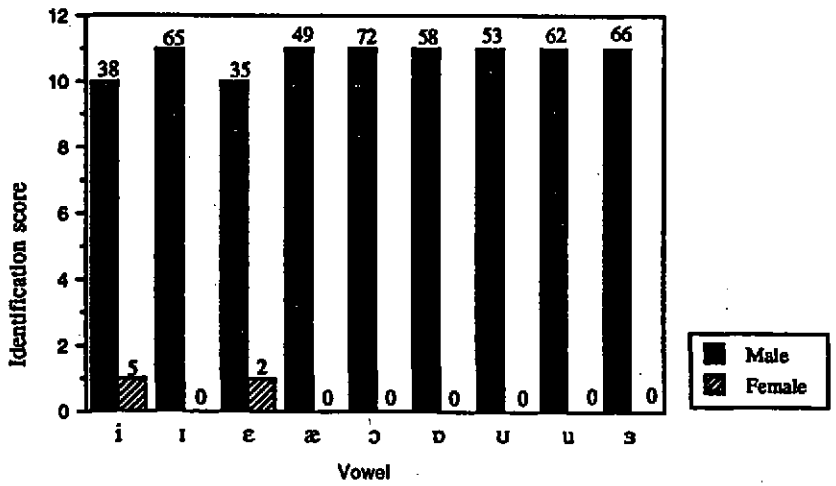

Figure 5. Simulus B identification scores - male formant frequencies and male f0.

4. GENERAL CONCLUSIONS

The results of this study suggest that fundamental frequency is the most salient cue that listeners use when judging the gender of synthetic stimuli. However, the importance of this cue seems to vary across vowels. The data here suggests that in vowels where F2 and F3 are particularly high as in the case of the close front vowels /i ɪ ɛ æ/, listeners perceive femaleness even when the fundamental frequency has been lowered to model the fundamental frequency of an adult male. Our findings provide some support for previous research which has highlighted the perceptual salience of vocal tract resonances ([6, 7, 8, 9, 10]).

5. REFERENCES

[1] Karlsson, I. (1988). Glottal waveform parameters for different speaker types. *Speech Transmission Laboratory- Quarterly Progress and Status Report, KTH, Stockholm*, 2-3, 61-67.

[2] Aronovitch, C. D. (1976). The voice of personality: stereotyped judgments and their relation to voice quality and sex of speaker. *Journal of Social Psychology*, 99, 207-220.

[3] Murry, T. & Singh, S. Multidimensional analysis of male and female voices, *Journal of the Acoustical Society of America*, 68: 1294-1300 (1980).

[4] Lass, N. J., Hughes K. R., Bowyer, M. D., Waters, L. T. & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered and isolated vowels. *Journal of the Acoustical Society of America*, 59: 675-678.

[5] Singh, S. & Murry, T. (1978) Multidimensional classification of normal voice qualities, *Journal of the Acoustical Society of America*, Suppl. 1, 64, p. 87.

[6] Bennett, S & Montero-Diaz, L. (1982) Children's perception of speaker sex, *Journal of Phonetics*, 10: 113-121.

[7] Schwartz, M. F. Identification of speaker sex from isolated, voiceless fricatives, *Journal of the Acoustical Society of America, 43*, 1178-1179 (1968).

[8] Schwartz, M. F. & Rine, H. E. Identification of speaker sex from isolated whispered vowels, *Journal of the Acoustical Society of America, 44*, 1736-1737 (1968).

[9] Childers, D. G. and Wu, K. (1991). Gender recognition from speech. Part II: Fine analysis. *Journal of the Acoustical Society of America*, 90(4), 1841-1856.

[10] Wu, K. and Childers, D. G. (1991). Gender recognition from speech. Part I: Coarse analysis, Journal of the Acoustical Society of America, 90(4), 1828-1840.

[11] Thomas, I. B. Perceived pitch of whispered vowels, *Journal of the Acoustical Society of America, 46*, 468-470 (1969).

[12] Peterson, G. E. & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2), 175-184.

[13] Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices, *Journal of the Acoustical Society of America*, 85, 1699-1707.

6. ACKNOWLEDGEMENTS

Our thanks to the speakers and listeners who took part in this study.