PHONETIC CONSIDERATIONS FOR THE SYNTHESIS OF FEMALE VOICES

CAROLINE G. HENTON

Linguistics Program, University of California, Davis,
Davis, CA 95616, U.S.A.

ABSTRACT

Synthesized female voices are scarce and lack naturality, but they are growing in demand. Acoustic and sociophonetic criteria are supplied for the improvement of female voices, and a ranking of importance suggested.

"Synthesis is going to be the next barn-burning technology," was an informed forecast three years ago [9]. It is further predicted that by 1992, the combined American and European markets for electronic speech synthesis and recognition devices will approach $5 billion [8]. While speech synthesis is a priority in speech technology research, and several commercial packages (e.g. DECtalk, Calltext) are producing successful male voices, distinctly fewer felicitous female synthetic voices are available. The widespread appearance of synthetic female speech is slow. Why is this? Are female voices not to be included in the barn-burning, or contribute to the multi-billion sales?
Outlined here are phonetic and social reasons for the paucity of synthetic female voices. There follows discussion of some acoustic specifications of female voices which are relevant to synthesis. Recent research pertaining to female voice quality is reported and a ranking of these various factors proposed.

BACKGROUND

Phonetically, the female voice has been largely ignored for two reasons. The first is as a result of minimal productional data. A cross-language survey of phonetic studies conducted 1952-1985 [14], which ostensibly provided 'representative' adult acoustic data, shows that among 42 studies, 40.5% assembled solely male speakers. 21.4% incorporated more males than females. Only one study (2.4%) incorporated more females than males. Studies of females alone are a meagre 4.8%. The first point, then, is that in acoustic phonetic research, the female voice has been either excluded or minimalized.
Secondly, female voices have been rejected acoustically (and hence, disregarded in phonetic theory) owing to inadequacies in analytic hardware. That should be obvious to anyone who has wrestled with interpreting spectrograms of female voices. Until recently, the sound spectrograph has been the most frequently-used tool in acoustic speech analysis, and other instruments (such as narrow-band spectrum analyzers) are still imperfect in analyzing females' speech. Criticisms of the problematicity of formant frequency determination for female speakers, using spectrography, are made by Ladefoged [25] and Ladefoged and Bladon [26].
The apparent source of the 'problem' of female speech appears in an article by Johansson et al. [21]: "Comparatively little is known about the characteristics of the female voice as compared with the male voice. The background is the high fundamental frequency range of the female voice which makes formant frequency estimates uncertain, and hence, information on the voice source unsafe." The logic may be chopped, and the association of formants with the voice source misleading, but the message is clear: the female voice is puzzling because it is not the same as a male's. This issue has also caught the attention of Klatt [23]. Reviewing the efficacy of spectrograms from which to draw acoustic conclusions, he states, "As far as speech ...research is concerned, it is not inconceivable that the sound spectrograph has had an overall detrimental influence over the last forty years by emphasizing aspects of speech spectra that are probably not direct perceptual cues." He presents spectrograms of the same utterance produced by a man, woman and child: "The woman and child speak with a much higher fundamental frequency, have a more breathy voice quality, and also have shorter vocal tracts, implying higher formant frequencies...". These traits are discussed further, below; meanwhile Klatt asserts that (p.83), "...it seems to be generally believed that the speech patterns of men and women could be made to look more similar if minor modifications were made to the sound spectrograph...Yet, here we are, nearly forty years later, and the sound spectrograph machine essentially has not changed." Such a situation surely reflects androcentric reality in the obscuration of females' voices. Here, though, the plea is not even for the spectrographic commensurability of male and female voices, but simply for the *measurability* (and hence reproducability) of female voices.
The implication from many such comments about female speech is that there is something intrinsically more difficult analytically, or just deviant, about female voices. The assumption is incorrect, but too few authors have thought to blame the design of the technology rather than females for producing analytical problems. Female voices only appear more 'difficult' because of the limitations of some present instrumentation. They are not more 'difficult' to the human ear: females are not any less intelligible than males, and may even be more so, although evidence seems somewhat variable [6], [12], [23], and [14: 312 ff.].
It is possible to infer then that a great deal more could be known about female voices, if the technology were improved for processing speakers with higher fundamental frequencies, namely the 'unquantifiable' females. Unfortunately (but entirely in keeping with the distribution of women in scientific positions generally), few phoneticians and even fewer technologists are female. So there is little grassroot motivation for improving females' analyzable lot.

WHAT NOW?

Notwithstanding such a negative background, demand is increasing for synthetic female speech. Naval pilots, for example, apparently react best to the voice of a young woman when warning them of upcoming obstacles or potential problems in the cockpit [8] It is clear that female voices are going to be needed more, for 'smoothing' and other messages.
So what *do* we know about female voices to contribute to their better synthesis? Sociophonetic evidence of the kind described in [14], [15] and [16] indicates a broad range of acoustic, perceptual and social factors as being influential in the synthesis of female voices. We will now briefly outline some of those categories. Starting with the generation of an appropriate glottal waveform, we then address formant frequency values for vowels, possible sex-specific factors in consonant production, and lastly, but perhaps most importantly, suprasegmental considerations and types of voice quality associated with female speech.

Glottal source characteristics. Naively it might be thought that to synthesize a convincing female voice pitch, it would be sufficient to simply double that of a male, increasing from, say 120Hz. to 240Hz. Several studies have shown, however, that there are marked differences in the glottal gestures of females and males [24], [31]. The latter show the glottal sound source of a normal adult can vary within a wide range, in respect to F0 and rms intensity, the appearance and shape of the waveform, and the phase and intensity spectra. Most important for the argument here, is the fact that all these variants can be influenced by the variables speaker sex, voice register and linguistic context. In addition, the female glottal waveform tends to have a less steep closing phase and a more rounded 'shoulder' at the end of that phase, and consequently, a higher ratio of open-to-close time which could result in more glottal leakage or weaker excitation of higher harmonics. Generating an appropriately varying female glottal waveform is thus vital for natural-sounding synthesis.

Formant frequencies. Male/female differences in formant frequencies of vowels have been reported widely elsewhere [4],[14]. Details do not bear reiteration here. Cross-linguistic findings from seven languages/dialects [14] may be summarized: while a male-female auditory normalization of approximately one Bark appears appropriate for F1 and F2, there are also indications that different speech communities need different amounts of normalization. That is to say, in some communities females and males appear to speak more like/unlike each other than their vocal anatomies would predict. A socially-conditioned element in speech production is thus posited. Hence the amount of physically-based input to the voice signal as compared to the socially-learned component must be weighed carefully when synthesizing speech.
In addition, the spectral tilt of female and male vowels might yield further evidence of sex-differentiation. we might predict that the angle of tilt of the spectrum, as it increases in frequency, would be somewhat steeper for females than for males. Indications are [31] that the decrease is approximately -12dB per octave in male vowels, but -15dB in female vowels. Inter-sex differences of formant bandwidth may also appear, with females' bandwidths being wider than males [11]. It is not known yet whether this difference is perceptible.
Sex-appropriate consonants. Whereas vowels exhibit important sex-specific cues, the evidence for specifically female consonants is less convincing. Generally, consonants have been explored less thoroughly: where male-female differences

have been noticed, it has been for ethnographic or tangential reasons. The relative linguistic function of vowels and consonants indicates that vowels exhibit the most individual-speaker traits. So speaker-sex (along with tone, affective state etc.) is more likely to be shown by vowels. Paralinguistic information such as exclamations, expression of pain, anger and so forth is also most often conveyed by vowel-like sounds. Consonants, conversely, are more language-specific, conveying linguistic information. Certain consonants, e.g. front fricatives and many stops, are unlikely to have vocal tract resonance properties which would show acoustic spectral differences of a sex-linked kind. Sex-specific behaviour in fricatives has been examined to a limited extent [3], [19], [35]. They indicate that different fricatives seem to carry differing amounts of sex-specific information. Weeninck [39] suggests that male-female differences in plosive bursts are also not anticipated. Other consonants are unlikely to show a sex-linked difference because there is already much individual speaker variation in their production. This argument applies to nasals, where individuals' nasal structures and degree of constriction may vary greatly. When speakers do appear to use consonants sex-specifically, they do so in an apparently unpatterned way: thus, females' realizations may differ from males' according to place of articulation, manner of articulation, direction of the airstream, voicing, or any combination of these four parameters. It is probable, then, that, sibilants apart, manipulating consonants to improve synthetic female speech will be unrewarding (unless a language uses consonantal variation as a specific sex-marker - which European languages, on the whole, do not).

Suprasegmental markers. Pitch, tone, vowel length, intensity, hesitancy, stress (word and sentence), rhythm and intonational tunes are rich areas for the exhibition of femininity/masculinity. Many argue that suprasegmental features are the primary cues of speaker-sex. Surprisingly little empirical, objective research into these areas has been initiated, however. Pitch has received some attention. Hollien and colleagues [17], [18] have indicated that the speaking fundamental frequency (SFF) of males may be lower than differences in stature would predict. This tendency is further borne out by Henton [13] for British English, and the conclusion must be that voice pitch is to some extent learned, and subject to sociocultural expectations. Such behaviour should not be overlooked in synthetic voices. Still, no researchers seem to be asking whether, for example, females habitually use a higher/lower pitch than vocal anatomy would predict; or are certain long-term SFFs associated with one sex, as interacting with socio-economic status; or do speakers of either sex alter their SFF by differing amounts according to register, or speech style? Preliminary answers to these sorts of questions, which are after all vital to successful synthesis, are beginning to appear incidentally [33], but this is hardly a broadscale line of enquiry.

Crystal [7] cites anecdotal observation of male-female suprasegmental differences such as glissando effects, complex-tone usage, breathiness and moving to falsetto involved in the production of 'simpering' voice in English. Aronovitch [1] found that his parameters of 'Intensity average; intensity variance; rate; F0 average; F0 variance and sound-silence ratio' were used differentially by raters for the two sexes. His goal, however, was to investigate if there are any connections between voice quality and stereotyped judgements, and not to examine diverse suprasegmental behaviour of the sexes. Unfortunately, his findings have yet to tantalize other researchers into further work along these lines.

Intensity has been examined to a limited extent. One study [29] revealed that males speak with greater average intensity in interpersonal communication than did females, although both sexes address a member of the opposite sex with greater intensity than a same-sex member. Another investigation [34] showed that women raise their intensity more than men when compensating for external noise. There is a brief summary in [38] of the few studies of sex-linked verbal fluency in adults. A comment by Smith [37:125] is still valid however: there are "only two paralinguistic features (viz. loudness and speech fluency, my parentheses) for which there is even a modicum of data."

For word and sentence stress, there is an array of intuitive remarks. Jespersen [20] claimed that "exaggeration of stress" is a conspicuous characteristic of female speech; and his intuitive successor, Lakoff [27:56] states that "Women speak in italics." As far as we are aware, no empirical investigation of stress in female speech is known as yet. Nor is there any of vowel length.

Sex-based intonation tunes *have* attracted attention, but too often in an unsystematic way, with little use of experimental measurement and too great a reliance on subjective judgement: [5], [22], [30]. Better results are provided by Pellowe and Jones [32], who showed that females use more rising tones than men. Women also exploited a greater variety of intonational tunes. Elyan's results [10] lend more support to this observation. More extensive exploration of these tendencies, across accents, is still

required. Male speech is characterized by monotonicity [2], while on the other hand female speech has greater *intonational dynamism* [30], varying more in respect of width of pitch-range; more frequent and rapid movements within the range; greater amplitude changes, and selection of differing tunes. McConnell-Ginet [30] hypothesizes that the female habit of changing pitch and loudness frequently may have great communicative importance, both attracting and keeping the listener's attention. As that is a common obstacle in speech synthesis, paramount attention should be given to such suprasegmental behaviour in females.

Voice quality. A major contribution to the paralinguistic component of speech is made by habitual vocal settings, or voice qualities, such as harshness, breathiness or creak. Perceptual correlates of voice qualities have been studied incidentally [1], but essential baseline information about the sociophonetic production and sex-specific use of different voice qualities is slow to emerge. Recently, though, Henton and Bladon [15], [16] have explored sex-related incidence of two common voice qualities: breathiness and creak. Both studies were conducted using a large corpus (80 speakers) of two accents of British English.

For breathiness, it was found, at least in open vowels, females employ significantly greater amounts of this voice quality, thus supporting previous tangential observations, [7], [23]. Implications for the diagnosis of normal versus pathological speech in females, as well as for speech synthesis are discussed further in that paper [15].

Creak, on the other hand, appears to be a marker of male speech in British English (and, by informal observation, in American English, too). Summarizing the findings of the creak study [16], we found that (a) creak was an utterance-final phenomenon, with a linguistic function of pre-pausal demarcation; (b) speaker-sex predicts drastically rate of creak: creak may be seen as a robust marker of male speech; (c) creak may be employed to different degrees in different accents to indicate hypo/hyper-masculinity; (d) creak is used habitually by overwhelming numbers of non-pathological male speakers. The incidence of creak thus varies with sex and accent, as well as with utterance-position. Male voice synthesis will obviously benefit from a sprinkling of creak, with female creak limited to utterance-final position (unless the desired voice is purposely androgynous).

These explorations make us ask more widely whether other voice qualities are sex-indicative. Are, for example, nasality, rhoticity, or pharyngealization typical of one sex or the other? Since voice quality appears to carry strong associations in the assessment of personalities [1], it seems essential to exploit this fact in speech synthesis. A female voice should therefore include greater breathiness, creak before relinquishing a turn and, speculatively, be more nasalized for appropriate accents.

SUMMARY AND PERCEPTUAL RANKING

After reviewing these various physical, segmental and suprasegmental aspects necessary for synthesizing female voices, it seems incumbent to provide a ranking of these features, according to perceptual salience. The ranking, in order of importance, is: voice quality, pitch, suprasegmentals, vowels, consonants. The synthesis of current synthetic female voices may not have adopted these criteria and so continue to sound unnatural. The increased production of convincing female voices, hand-in-hand with eliminating the female voice as "one of the mysteries of the universe" [36] is undoubtedly a profitable goal for speech synthesis.

[1] C.D. Aronovitch, The voice of personality: stereotyped judgements and their relation to voice quality and sex of speaker. *Jour. Soc. Psych.*, 99: 207-20, 1976.
[2] S. Bennett & B. Weinberg, Sexual characteristics of pre-adolescent children's voices. *Jour. Acoust. Soc. Amer.*, 65: 179-89, 1979.
[3] A. Bladon, The use of auditory modelling for speaker normalization in speech recognition. In P. Mermelstein (ed.) *Montreal Symposium on Speech Recognition*. 1986.
[4] R.A.W. Bladon, C.G. Henton & J.B. Pickering, Towards an auditory theory of speaker normalisation. *Lang. & Speech*, 4: 59-69, 1984.
[5] R. Brend, Male-female intonation patterns in American English. In B. Thorne & N. Henley (eds.) *Language and Sex: Difference and Dominance*. Newbury House: 84-87, 1975.
[6] F.R. Chen, *Acoustic Characteristics and Intelligibility of Clear and Conversational Speech at the Segmental Level*. S.M. thesis, Dept. Electrical Engineering, MIT, 1980.
[7] D. Crystal, *The English Tone of Voice*. Arnold, 1975.
[8] B. Darrow, Research spurs development of talking machines. *Design News*, 12 March, 1984, p.110.
[9] L. Drinkwater, of Wang, quoted in *Industry Week*. October 15, 1984, p.82.
[10] O. Elyan, Sex differences in speech style. *Women Speaking*, 4, 1978.
[11] G. Fant, Temporal fine structure of formant damping and excitation. In J.J. Wolf & D. Klatt (eds.) *Speech Communication Papers*, Acoust. Soc. Amer: 161-66, 1979.
[12] U. Goldstein, *An Articulatory Model for the Vocal Tracts of Growing Children*. D.Sc. thesis, MIT, 1980.
[13] C.G. Henton, Normalization: fundamental problems. *Proc. Inst. Acoust.*, 6: 267-73, 1984.
[14] C.G. Henton, *A Comparative Study of Phonetic Sex-specific Differences Across Languages*. D.Phil. thesis, University of Oxford. 1986.
[15] C.G. Henton & R.A.W. Bladon, Breathiness in normal female speech: inefficiency versus desirability. *Lang. and Commun.*, 5: 221-27, 1985.
[16] C.G. Henton & R.A.W. Bladon, Creak as a sociophonetic marker. In L. Hyman & C. Li (eds.) *Language, Speech and Mind: Studies in Honor of Victoria A. Fromkin*. Croom Helm, forthcoming, 1988.
[17] H. Hollien & B. Jackson, Normative data on the speaking fundamental frequency characteristics of young adult males. *Jour. Phonet.*, 1: 117-20, 1973.
[18] H. Hollien & T. Shipp, Speaking fundamental frequency and chronologic age in males. *Jour. Sp. and Hear. Res.*, 15: 155-59, 1972.
[19] F. Ingemann, Identification of the speaker's sex from voiceless fricatives. *Jour. Acoust. Soc. Amer.*, 44: 1142-44, 1968.
[20] O. Jespersen, *Language, Its Nature, Development and Origin*. Allen & Unwin, 1922.
[21] C. Johansson, J. Sundberg & H. Wilbrand, X ray study of articulation and formant frequencies in two female singers. *Quar. Prog. Stat. Rep.*, KTH Stockholm, 4: 117-34, 1982.
[22] M.R. Key, Linguistic behavior of male and female. *Linguistics*, 88: 15-31, 1972.
[23] D. Klatt, Speech processing strategies based on auditory models. In R. Carlson and B. Granstrom (eds.) *The Representation of Speech in the Peripheral Auditory System*. Elsevier: 181-96, 1982.
[24] D. Klatt, Detailed spectral analysis of a female voice. Abstract in *Jour. Acoust. Soc. Amer.*, 80, Suppl.1: S97, 1986.
[25] P. Ladefoged, *Three Areas of Experimental Phonetics*. Oxford University Press, 1967.
[26] P. Ladefoged & R.A.W. Bladon, Attempts by human speakers to reproduce Fant's nomograms, *Sp. Comm.*, 1: 185-98.
[27] R. Lakoff, *Language and Woman's Place*. Harper Colophon, 1975.
[28] W. Majewski, H. Hollien & J. Zalewski, Speaking fundamental frequency characteristics of Polish adult males. *Phonetica*, 25: 119-25, 1972.
[29] N.N. Markel, L.D. Prebor & J.F. Brandt, Biosocial factors in dyadic communication: sex and speaking intensity. *Jour. Pers. and Soc. Psych.*, 23: 11-13, 1972.
[30] S. McConnell-Ginet, Intonation in a man's world. In B. Thorne, C. Kramarae & N. Henley (eds.) *Language, Gender and Society*. Newbury House, 1983.
[31] E.B. Monsen & A.M. Engebretson, Study of variations in the male and female glottal wave. *Jour. Acoust. Soc. Amer.*, 62: 981-93, 1977.
[32] J. Pellowe & V. Jones, On intonational variability in Tyneside speech. In P. Trudgill (ed.) *Sociolinguistic Patterns in British English*. Arnold: 101-21, 1978.
[33] O. de Pinto & H. Hollien, Speaking fundamental frequency characteristics of Australian women: then and now. *Jour. Phonet.*, 10: 367-75.
[34] W. von Raffler-Engel & J. Buckner, A difference beyond inherent pitch? In B. Dubois & I. Crouch (eds.) *The Sociology of the Languages of American Women*. Texas, 1976.
[35] M.F. Schwartz, Identification of speaker sex from isolated, voiceless fricatives. *Jour. Acoust. Soc. Amer.*, 43: 1178-79, 1968.
[36] R. Shuy, Sociolinguistic research at the Center for Applied Linguistics: the correlation of language and sex. *International Days of Sociolinguistics*. Istituto Luigi Sturzo: 849-57.
[37] P. Smith, *Language, the Sexes and Society*. Blackwell, 1985.
[38] B. Thorne, C. Kramarae & N. Henley (eds.) *Language, Gender and Society*. Newbury House, 1983.
[39] D.J.H. Weenink, Literature overview on perceptual and physical normalization of speaker variation. *Proc. Inst. Phon. Sciences, Univ. Amsterdam*, 8: 5-17, 1984.