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 naturalness, 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 (recently DICTalk, Calltext) are producing readable male voices, distinctly female voices remain elusive. 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 dollar market?

Outlined are phonetic and social reasons for the paucity of synthetic female voices. There follows discussion of the acoustic specifications of female voices which are relevant to synthesis. Recent research pertaining to female voice quality and a ranking of these variables is proposed.

BACKGROUND

Phonetically, the female voice has been largely ignored for two reasons. The first is a result of minimal female data. A cross-section of survey of phonetic studies conducted 1952-1985 [14] which categorically provided 'representative' adult acoustic data, showed that 82 studies, 40% assembled solely male speakers. 23% incorporated more males than females. Only one study (24%) incorporated more females than males. Studies of females alone are a mere 4%. The first point is then, that in acoustic phonetic research, the female voice has been either excluded or minimized.

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 problematic nature of formant frequency determination for female speakers, using spectrography are made by Ladeeford [25] and Ladeeford and Bladon [26]. The apparent source of the problem of female speech appears in an article by Johnson 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 analysis uncertain, and hence, information on the voice source unsafe." The logic may be chopp'd, and the association of formants with the voice is realized, but the message is clear: the female voice is puzzling because it is not the same as a male one. Base-line layering of the work of Klatz [23]. Reviewing the efficacy of the formant, from both linguistic and acoustic conclusions, he states, "As far as speech research concerned, it is inconceivable that the sound spectrograph has had an overall detrimental influence for male speech. Although it presents spectrograms of the same utterance produced by both a woman and a child: the woman and child speak with a much higher fundamental frequency, have a more defined voice quality, and also have shorter vowel tracts, implying higher formant frequencies." In addition, formant frequencies have been discussed further, below; meanwhile Klatz asserts that (1983) "...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...essentially has not changed." Such a situation surely reflects inadequacy in consideration of the production of female voices. Here, though, the idea is not even for the spectrographic immunsorbability of male and female voices, but simply for the mass-rejection of fundamental inaudibility of female voices.

The impetus from many such comments about female speech is that there is something intrinsically '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 we even use more so, although evidence seems somewhat variable [6], [12], [23], and [14; 312 ff.].

It is possible to infer that a great deal more could be known about female voices, if the technology were improved for processing speakers with higher fundamental frequency, namely the 'quantifiable' females. Unfortunately not entirely in keeping with some of the little evidence (in scientific positions generally), few phoneticians and electronic engineers are female. So there is little grassroots motivation for improving females' analysable lot.

WHAT NEXT

Notwithstanding such a negative background, the demand for incorporation of synthetic female speech. Naval pilots, for example, apparently react more favorably when women speak through an opposing system or potential problem in female voice. It is clear that female voices are going to be needed in the near and far future. It is evident that... researchers concerned. In what do we know about female voices to contribute to their better kind? Isolating the changes of the voice described in [14], [15] and [16] indicates a potential for perceptual and acoustic social factors as being influential in the perception of female voices. We will briefly outline some of those categories.

Running with the generation of formant frequency values, we then appropriate the different values, apply them through matrices representing the formant frequency values for vocal and possible semi-vocal social factors in constant production, and lastly, provide an experimental consideration of voice quality associated with female speech. Glottal source characteristics. It would seem that the art of synthesizing a convincing female voice might be considered as adding pitch, it would be sufficient to simply double that of a male voice, increasing from say 120Hz. to 240Hz. Several studies have shown, however, that there are marked differences in the glottal sound source of the male and female adult voice can vary within a wide range, in respect to F0 and resonance intensity, the appearance and shape of the waveform, and the phase of the source. The important argument here is, however, is that the fact that all these variables can be influenced by the variables speaker 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-clos phase 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 [41],[14]. Details do not bear reiteration here. In all seven languages/dialects [14] may be summarised: for example, F1 and F2 normalisation of approximately one Bark appears appropriate for Fi and F2, there accordingly variations that different speech communities need different normalisation. That is to say, in some communities females and males appear to speak more linguistically but their voice frequencies other than their vocal anatomy would predict. A socially-conditioned element in speech production is thus posited. Hence the production of a more linguistically in keeping with the socially-learned component must be accounted for. The amount of physiological input compared to the socially-learned component may be large.

In addition, the spectral tilt of female formant frequency is lower than that of males. The evidence of sex-differentiation, we might predict that the difference in frequency decreases, but would be somewhat still lower. What is that the difference? If the female voice a decrease is approximately 1500Hz in female voices, the common difference across sex, males being wider than males [11]. It is not clear whether this it is not clear whether this is perceptible. Whereas.
have been noticed, it has been for ethological or topographical reasons. The relative linguistic function of vowels and consonants indicates that vowels exhibit the greatest perceptual variation in speaker-sex (along with tone, affective tone, and intonation) as shown by their frequency of use. Paralinguistic information such as vocal tracts, voice, and intonation properties which would show acoustic spectral differences of a sex-linked kind. Specific behavior 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. Wennergren [38] suggests that male-female differences in plosive bursts are not always 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 individuates' 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' nasal vowels are more likely to bear the lower frequency, more nasalized, more stressed nasal vowels than male's according to place of articulation, manner of articulation (the airstream, voicing, or any combination of these four parameters). It is probable, then, that, sibilants apart, manipulating consonants to interpret synthetic female speech will be unprofitable (unless a language uses consonantal variation as a significant sex-marking device). The pitch, tone, vowel length, intensity, hesitancy, stress (word and sentence) and intonation are rich areas for the ethnolinguistic investigator. Many argue that suprasegmental features are the primary cues of speaker-sex. Surprisingly little empirical objective research into these areas has been conducted. Piatt has received some attention. Hollin and colleagues [18] indicate that the speaking fundamental frequency (SFF) of men is considerably higher than that of women. For word and sentence stress, there is an additional component of intuitions. (20) The linguistic function of stress is a conspicuous characteristic of female speech; women have used to set off the higher falls of the intonation. Akin to the male's higher speaker's voice, the higher falls are more common in their utterances. The stimulus that results in a higher pitch may be seen as a robust marker of female speech. An alternative explanation is that the speaker's voice pitch may be a marker of social status.