Proceedings of The Institute of Acoustics

NORMALIZATION: FUNDAMENTAL PROBLEMS

C.G. Henton

Phonetics Laboratory, University of Oxford

INTRODUCTION

If asked what basically distinguishes female and male voices, average listeners would answer 'pitch'. Approaches to speaker sex normalization have reflected this attitude and assumed that if differences between female and male F0 can be eliminated, then female and male speech will be more comparable. This assumption is here shown to be simplistic and insufficient.

The voice fundamental frequency of females and males is not the only differentiating factor: formant frequencies and a learned, socially-defined behaviour are also influential. Therefore more is needed in normalization procedures than a straightforward transform of F0. Further, when the amount by which F0 is to be normalized is deduced, we have to remember that F0 is under speakers' conscious control and that speakers chose to exercise that control in different ways in different societies. So, while F0 is amenable to normalization, it is not necessarily reliable.

A. PREVIOUS STUDIES: PROBLEMS AND MISCONCEPTIONS

Much work has been devoted to establishing just how important is F0, compared with formants, in speakeer identification tasks. Schwartz [1968] showed that speaker-sex could be established by two of four voiceless fricatives (viz. /s/ and /ʃ/). Ingemann [1968] extended the fricative range examined, provided slightly less convincing results than Schwartz, but still concluded that speaker sex is recognizable from voiceless sounds. These steps were taken further by Schwartz and Rine [1968], who suggested the dispensability of F0 by showing that isolated whispered vowels were identifiable for sex of speaker.

The normalization theory developed in Oxford has an auditory basis. It uses the Bark scale, for which justification is given elsewhere (see Bladon, Henton and Pickering, 1984). It is interesting here to see how well the Bark normalization applies to the data of Schwartz and Rine. Figure 1 reveals how effectively the female /i/ vowel would 'map' onto their male /i/ if it were transposed downwards in frequency by one Bark (the amount previously calculated to normalize between speakers' formants). The figure also shows the logarithmic nature of the Bark scale, with higher frequencies correlating with more than one Bark.

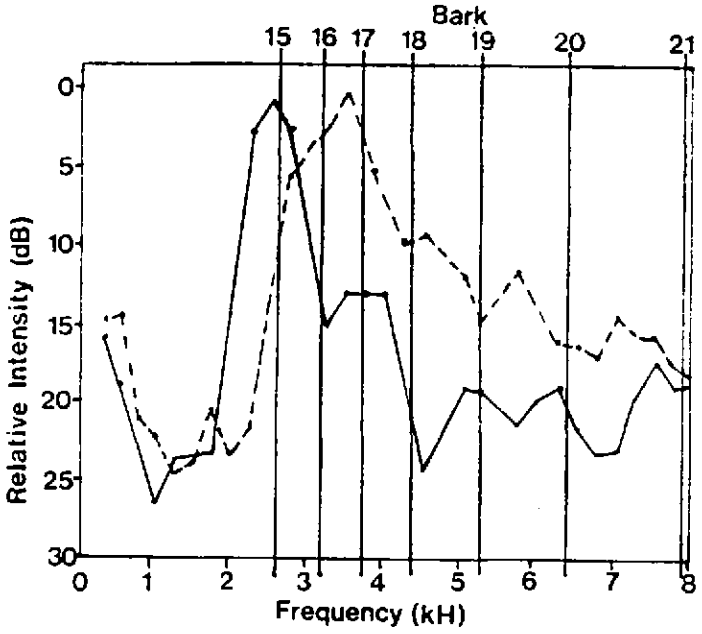

FIGURE 1 Average spectra for the whispered vowel /i/ from five females (dashed line) and five males (solid line), from Schwartz and Rine [1968]. The vertical lines show how a 1 Bark transformation downwards in frequency generally map the female peaks on to the males'.

Previous studies, then, imply that the presence of a fundamental frequency is not vital for classification by sex. The next question to be broached is whether F0 is disposable in general, or whether it combines with formant frequencies in vowels to provide stronger cues to speaker identification.

Coleman examined the comparative contributions of F0 and formants to female and male voice quality apparently exhaustively. His first study [1971] leads one to believe that perception of sex as engendered solely by F0 is not supportable. Reasonable accuracy appeared in sex-assignment for /i/ and /u/ from their formant frequencies alone for male speakers; female speakers seemed more troublesome, with two females being judged male. The reason is obvious and unfortunately undermines the validity of Coleman's findings, both here and in the later studies [1973a and 1973b]. The F0 generated by the electrolarynx had a frequency of 85 Hz! This androcentric bias apart, he also concedes that, "It is... possible that males and females differ

in some learned speech characteristics..." Quite so: this aspect will be highlighted in Section C.

Coleman's results should not be extended to females. Any conclusions applying to female F0 in relation to formants should only rest on the artificially-produced F0 being a more representative frequency, say 210 Hz. When applied to males, though, there does seem to be evidence from Coleman's work for speakers being identifiable from non-glottal characteristics. His [1973b] paper reports findings which were much better controlled. The conclusions are that, in natural speech, the sex of the speaker is determined primarily by F0, with vocal tract resonances contributing negligibly to sex perception. Conversely, when an artificial male F0 was combined with the vocal tract resonance values of a female and vice versa, the male characteristics preserved their perceptual prominence, while the female F0 was much weaker. Thus the findings of these two experiments are inconsistent, but a loophole remains with Coleman's admission that the particular electro-larynx used may generate a more natural-sounding F0 for males than for females. More interesting for the purposes here is his further comment that, "It is also possible that the glottal source in females differs from males in some basic way besides simply that of pitch", [1973b:21]. That possibility is precisely what is explored in Section C.

Lass et al. [1976], spotting a set of variables which had not previously been juxtaposed in his and other colleagues many speaker-identification experiments, used six vowels in three conditions to conclude that other experiments were correct: the laryngeal fundamental provides a more salient cue to speaker sex than do formants. They thus agree largely with Coleman, but do not necessarily find support from the Schwartz studies or from Ingemann, as Lass et al. infer they do.

Conflicting arguments exist, then, for the relative roles of F0 and formants in the assignment of speaker sex. On the one hand, there is evidence that sex can be established with no F0 information at all; on the other, there are the studies which indicate the supremacy of F0 in perceptual terms.

Sekimoto [1983], following on from and concurring with Fujisaki and Kawashima [1968], finally concludes that for Japanese at least, it is the concurrent variation between the fundamental and the spectrum envelope, together with the ratios between the formant frequencies (including the higher frequencies) which are the important factors to be taken into account in a normalization algorithm. The auditory normalization theory takes a similar view. How it fares with F0 normalization in RP is examined in the next section.

NORMALIZATION: FUNDAMENTAL QUESTIONS

B. F0 NORMALIZATION IN RP VOWELS

Twenty female and twenty male speakers of Received Pronunciation recorded eleven vowels in an hVd context in citation form. The vowels were /i:,ɪ,e,æ,ɑ:,ɒ,ɔ:,ʊ,u:,ɜ:,ʌ /. Recording equipment and procedures were as reported in Henton [1983]. Unusual care was taken to ensure speaker homogeneity in terms of age, stature, socio-economic background, state of health etc.

Narrow band spectrum analyses were made of the vowels, using a Brüel and Kjaer Narrow Band Spectrum Analyzer, Type 2031, selecting a frequency range of 10-5000 Hz, in combination with an oscillographic chart recorder (Brüel and Kjaer X-Y Recorder, Type 2308). Spectral sections were taken at steady states and the F0 for each vowel calculated. This was done by locating the tenth harmonic, or above, and dividing by that number. This method is considered to provide a more accurate representation of F0 than merely taking the frequency of the lowest harmonic.

The mean F0 value for each vowel was calculated for the twenty speakers of each sex. These values are plotted, in Hertz, in Figure 2.

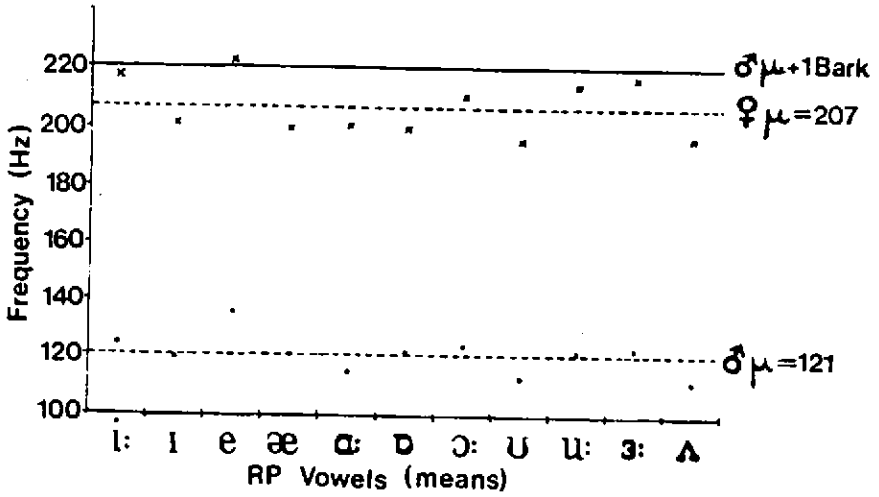

FIGURE 2 Mean values for F0 of eleven RP vowels, spoken by twenty males (dots) and twenty females (crosses). Dotted lines show the mean F0 for all the vowels. The solid horizontal line indicates an increase of 1 Bark (the expected normalization value) above the mean male value for all the vowels.

The overall uniformity of pattern between the productions of the vowels by each sex is quite pleasing. Surprise deviations emerging will be discussed in greater detail when space and time permit. What is of more immediate concern is the average difference between the mean male and female F0 across the vowels. This difference is 86.3 Hz, or approximately 0.8 Bark.

For the normalization of the first two formants of the RP vowels, though, a figure of 1.1 Bark has been derived, and it has seemed reasonable to expect this to be appropriate for F0, too. The reasons for the apparently different amount of normalization required by F0 therefore need to be examined. Possible explanations include: (i) the Bark scale is inaccurate, or not appropriate at frequencies below 100 Hz, and suffers from an "end of scale effect" (see Traunmüller, 1981: Fig.6); (ii) an element of social conditioning has to be admitted into calculations for the normalization of F0.

C. LEARNED ELEMENTS IN PITCH

Lieberman [1967] and Meditch [1975] have indicated that babies and pre-pubertal children manipulate their pitch according to the sex of their interlocutor and social role expectations. Brend [1972] implies that female speakers exploit a greater pitch range than do males. With typical unattested anecdote, Lakoff [1975] claims that American women exclusively use high rise intonation tunes for declarative answers to questions; counter-arguments to her explanation are provided by Dubois and Crouch [1975]. The connection of these findings to the RP data is discussed. Female RP speakers do not appear to be using a greater pitch range than males, but they do seem to be conforming with an idea of 'minimal separation'.

RP spoken by females has been called the voice of 'perceived androgyy' among the dialects of British English [Elyan, Smith, Giles and Bourhis, 1978]: that is, they are deemed to possess some more 'male-like' qualities in their voices. Figure 2 supports this notion: the fundamental frequencies of the vowels of female and male RP speakers are separated by a considerably smaller amount than are their formant frequencies. This may imply that either RP males are speaking 'higher' or RP females 'lower' than one would predict: the latter is more probable. Preliminary results from a comparison of RP with data for Modified Northern speakers certainly supports the notion that RP-speaking women do lower their pitch in the production of isolated vowels. Since physical and contextual variables were controlled for, the women are speaking 'lower' for social, and not necessarily physical, reasons.

In accord with this notion that voice pitch is a learned, socially-motivated factor in voice quality, further data is also being assembled from television broadcasts by female newsreaders.

From the results it is hoped that support will be gained for the (previously unempirical) statements that (a) female speakers of RP use a lower than average long-term pitch; (b) that female newsreaders in particular employ lower than average pitches, but do not use a wider than average pitch range, and (c) that voice pitch is exploited consciously for the conveyance of social prestige, power and as the 'voice of authority'. Links can be made with Ohala's [1983] 'frequency code' where pitch is used as a signal of dominance or submissiveness.

D. CONCLUSION

If fundamental frequency can be established as an important, inter-active indicator of speaker sex, then the fundamental frequency has to be normalized accurately and, it appears in the case of RP at least, by a different amount than the formants. The convincing synthesis of RP speech will have to take this finding into account.

However, if fundamental frequency is very much under individual speakers' control, then it has to be regarded as both a physiologically and socially-determined quality. Uniform normalization, across sociolects and perhaps even across the speech of the same speaker in different interactive contexts, will be insufficient and misrepresentative of the complexity of social usage of pitch.

Acknowledgement The preparation of this paper was supported by the Economic and Social Research Council, Grant No. C00232033.

REFERENCES

Bladon, A., Henton, C.G. and Pickering, J.B. [1984] Towards an auditory theory of speaker normalization. Language and Communication, 4:59-69.

Brend, R.M. [1972] Male-female intonation patterns in American English. Proc. Seventh International Congress of Phonetic Sciences, 1971. The Hague, Mouton.

Coleman, R.O. [1971] Male and female voice quality and its relationship to vowel formant frequencies. Jour. of Speech and Hearing Research, 14:565-77.

Coleman, R.O. [1973a] Speaker identification in the absence of inter-subject differences in glottal source characteristics. JASA, 53:1741-1743.

Coleman R.O. [1973b] A comparison of the contributions of two vocal characteristics to the perception of maleness and femaleness in the voice. STL-QPSR, 2-3:13-22.

NORMALIZATION: FUNDAMENTAL QUESTIONS

Dubois, B.L. and Crouch, I.M. [1975] The question of tag questions in women's speech: they don't really use more of them, do they? Language in Society, Dec. 1975:289-294.

Elyan, O., Smith, P., Giles, H. and Bourhis, R. [1978] RP-accented female speech: the voice of perceived androgyny? In Trudgill, P. (ed). Sociolinguistic Patterns in British English. London, Arnold.

Fujisaki, H. and Kawashima, T. [1968] The roles of pitch and higher formants in the perception of vowels. IEEE Trans. Audio. Electroacoust., AU16 (1):73-77.

Henton, C.G. [1983] Changes in the vowels of received pronunciation. Jour. of Phonetics, vol.11:353-371.

Ingemann, F. [1968] Identification of the speaker's sex from voiceless fricatives. JASA, 44:1142-1144.

Lakoff, R. [1975] Language and Woman's Place. New York, Harper and Row.

Lass, N.J., Hughes, K.R., Bowyer, M.D., Waters, L.T. and Bourne, V.T. [1976] Speaker sex identification from voiced, whispered, and filtered isolated vowels. JASA, 59:675-678.

Lieberman, P. [1967] Intonation, Perception and Language. Cambs.Mass. MIT Press.

Meditch, A. [1975] The development of sex-specific patterns in young children. Anthropological Linguistics, 17,9:421-433.

Ohala, J.J. [1983] Cross-language use of pitch: an ethological view. Phonetica, 40:1-18.

Schwartz, M.F. [1968] Identification of speaker sex from isolated, voiceless fricatives. JASA, 43:1178-1179.

Schwartz, M.F. and Rine, H.E. [1968] Identification of speaker sex from isolated, whispered vowels. JASA,44:1736-1737.

Sekimoto, S. [1983] Normalization of the speaker difference in vowel perception. Annual Bulletin RILP, 17:83-96.

Traunmüller, H. [1981] Perceptual dimension of openness in vowels. JASA, 69: 1465-1475.