Outline of an Auditory Theory of Speaker Normalization

R.A.W. Bladon, C.G. Henton and J.B. Pickering
Oxford, United Kingdom

Research towards systems of speaker-independent speech recognition and theoretical research on speech perception are both confronted by the problem of speaker normalization, perhaps most forcibly illustrated by the speaker sex problem in acoustic phonetics. This problem can be summed up as follows: two vowels (cf. the data of Peterson and Barney, 1952) or two fricatives (cf. the data of Schwartz, 1968), one of which is spoken by a male and one by a female can be judged 'the same' even by a trained phonetician, yet our analysis equipment reveals great differences. The differences are exemplified by - but not exhausted by - the well known ones of formant frequency.

According to Fant (1975) the problem is compounded, because the measured differences (his 'k-factors') are not consistent from vowel to vowel, nor from formant to formant. One is reminded of his dictum that (ibid.), 'In terms of the acoustic code, female speech remains an obscure dialect.'

Hitherto, approaches to the speaker sex problem (see Disner, 1980 for a review) have mostly concentrated on scaling the acoustic data according to inferred or observed differences in speaker physiology such as vocal tract length.

Our approach is different: it is listener-orientated, and it draws on current knowledge about human auditory analysis, in the belief that this forms an important building-block in the modelling of speech perception. As our point of departure we take Potter and Steinberg's old (1950) idea that, 'a certain pattern of stimulation along the basilar membrane may be identified as a given sound, regardless of position along the membrane'.

That their idea was attractive can be deduced from Figure 1. Now that a good estimate of basilar membrane frequency analysis is available to us (in the psychophysical form of the critical-band scale of Bark units) it is instructive to plot on a Bark scale the same heterogeneous cross-language data from Fant (1975) whose apparent non-uniformity disturbed us at first sight. The plot in Figure 1 is of inter-formant distance F_1-F_2 (in Bark), female against male. The emergent signs of correlation are encouraging.

Our auditory theory of normalization goes somewhat further. As reported at length elsewhere (Bladon and Lindblom, 1981), we postulate a series of acoustic-to-auditory spectral transforms which contain not only a conversion to the Bark scale, but also an auditory filter designed to reflect aspects of

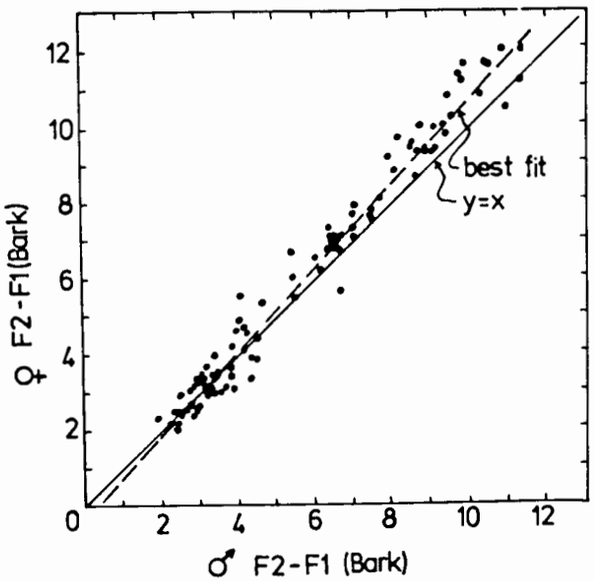

Figure 1. Inter-formant distance F2-F1 (in Bark), females against males. Data from six languages taken from Fant (1975).

masking, together with transforms of intensity level in terms of equal loudness curves and of total loudness calibrated in terms of loudness density per Bark. Further details of this 'auditory model' need not be rehearsed here: suffice it to say that the output is a quasi-auditory spectrum (of e.g. a vowel), which is meant to correspond to that vowel's excitation pattern on the auditory nerve.

Next, take two vowels represented as auditory spectra in this sense: a male vowel and a female equivalent. Suppose that we follow the Potter and Steinberg idea and, analogically speaking, preserve the two excitation patterns in the auditory system but displace the position of one of them. In our terms, we effect a simple linear Bark scale shift of the female pattern. Figure 2 illustrates this procedure applied to several vowels in our data, using a shift of 1 Bark. The coincidence of the resultant spectral shapes is not complete, but it is encouraging as a first approximation. Some progress has in fact been made upon the modifications which are apparently needed, and these are being reported elsewhere. Foremost among these modifications is a warping of the spectrum, especially in the F_1 region, owing to interference from F_0 (see Bladon, 1982).

This auditory theory of normalization has to date been tested on seven sets of male/female vowel data from four languages. Preliminary indications from this (far from satisfactory) database are that the optimal male/female normalization (expressed on average for the vowels of a particular language or dialect) may not necessarily be by 1 Bark. The optimal normalization

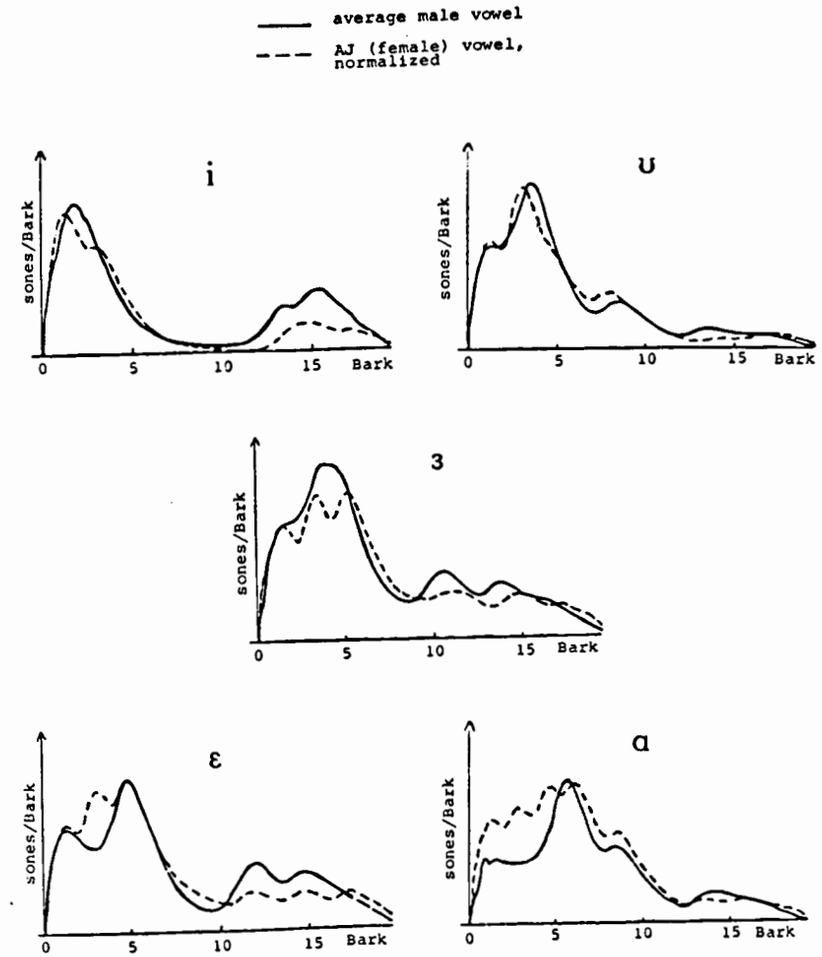

Figure 2. Vowel pairs, male and female, after normalization of the female (dotted vowel) by a downward shift of 1 Bark. The male vowel spectra are averaged over 5 speakers; the female spectra are from one speaker AJ; all are speakers of Middle Northern British English.

displacement varies considerably across speech communities. Figure 3 demonstrates this finding quantitatively, insofar as present data permit. The suggestion is, rather as observed by Labov (1978) in Martha's Vineyard speakers, or as concluded by Goldstein (1980) from her vocal tract modelling, that males and females may in some speech communities speak more unlike (or, more like) each other than their vocal tract physiology would predict. In other words, the data of Figure 3 appear to implicate a learned, socially motivated factor for part of a model of speaker normalization.

However, in order to establish with any certainty these tentative suggestions of phonetic role-stereotyping in vowels, the investigator must be conscious of the need to control a large number of variables. These include

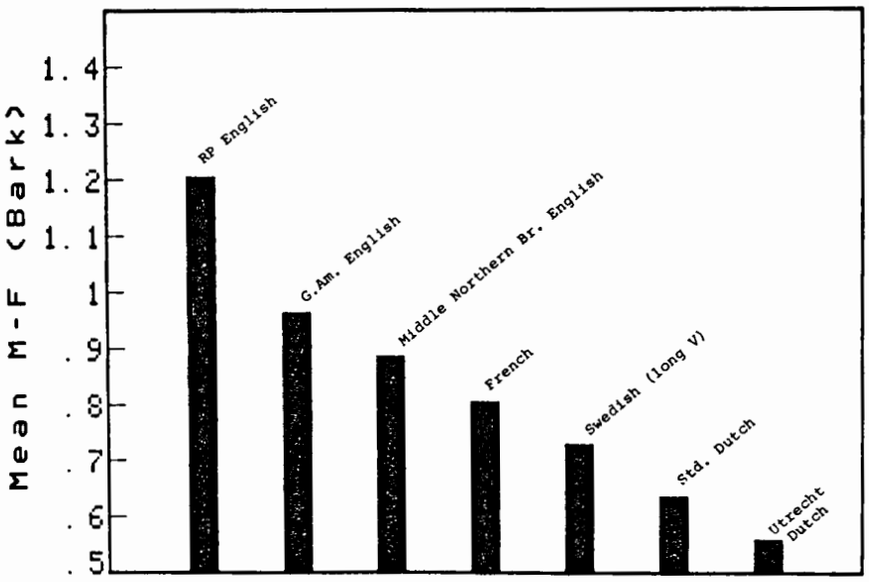

Figure 3. Optimum mean male/female vowel normalization (in Bark), for seven languages/dialects. The data for RP and Middle Northern British English are our own. The remaining sources are: General American (Peterson and Barney 1952), French (Carton 1974 and Mettas 1979), Swedish (Fant 1979), Standard Dutch (Pols et al. 1973 and van Nierop et al. 1973), Utrecht Dutch (Koopmans-van Beinum 1973).

speaker variables such as physique, age, socio-economic background and state of health, as well as experimental variables such as linguistic context, F_0 used and recording conditions. For a discussion of these problems see Henton (1983). It is unfortunate that a good many important controls are missing from most of the data examined in this paper; consequently the conclusions should be treated as no more than suggestive at this stage.

Acknowledgement

This work was supported in part by a Scientific Investigations Grant from the Royal Society.

References

Bladon, R.A.W. (1982). Problems of normalizing the spectral effects of variations in the fundamental. *Proc. Inst. Acoust. 1982*, A51-A55.

Bladon, R.A.W. and Lindblom, B. (1981). Modelling the judgement of vowel quality differences. *J. Acoust. Soc. Am.* **69**, 1414-1422.

Carton, F. (1974). *Introduction à la Phonétique du Français.* Paris: Bordas.

Disner, S.F. (1980). Evaluation of normalizations. *J. Acoust. Soc. Am.* **67**, 253-261.

Fant, G. (1975). Non-uniform vowel normalization. *RIT Stockholm Qu. Prog. Stat. Rep.* 2-3/1975, 28-52.

Fant, G. (1979). *Speech Sounds and Features.* MIT Press.

Goldstein, U. (1980). An Articulatory Model for the Vocal Tracts of Growing Children. Doctor of Science Thesis, MIT.

Henton, C.G. (1983). Changes in the vowels of Received Pronunciation. *J. Phon.* **11** (forthcoming).

Koopmans-van Beinum, F.J. (1973). Comparative Phonetic vowel analysis. *J. Phon.* **1**, 249-261.

Labov, W. (1978). *Sociolinguistic Patterns.* Oxford: Blackwell.

Mettas, O. (1979). French oral vowels analysed from recording spontaneous conversations. In: B. Lindblom and S. Öhman (eds.), *Frontiers of Speech Communication Research.* London: Academic Press.

Peterson, G.E. and Barney, H. (1952). Control methods used in a study of the vowels. *J. Acoust Soc. Am.* **24**, 175-184.

Pols, L., Tromp, R. and Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *J. Acoust. Soc. Am.* **53**, 1093-1101.

Potter, R.K. and Steinberg, J.C. (1950). Towards the specification of speech. *J. Acoust. Soc. Am.* **22**, 807-820.

Schwartz, M.F. (1968). Identification of speaker sex from isolated voiceless fricatives. *J. Acoust. Soc. Am.* **43**, 1171-1179.

Van Nierop, Pols, L. and Plomp, R. (1973). Frequency analysis of Dutch vowels from 25 female speakers. *Acustica* **29**, 110-118.