USING AUDITORY MODELS FOR SPEAKER NORMALIZATION IN SPEECH RECOGNITION

Anthony Bladon

Phonetics Laboratory, University of Oxford,
41 Wellington Square, Oxford OX1 2JF, U.K.

Auditorily-transformed versions of the speech spectrum may well be a useful way of reducing the apparently nonuniform physical differences between speakers. A speaker normalization technique of this kind is however justified to different degrees by different kinds of speech event. Does this presuppose a need for higher-level (phonetic class) information at the acoustic level in speaker-independent ASR?

> "It is obvious from our experiment that the unqualified assumption does not hold - auditory models used as speech recognition front ends will not consistently improve performance."

Blomberg et al.'s (1984) ominous words are ones which this symposium ought to take seriously to heart. They conflict with our initial theoretical expectations. This paper will not attempt to investigate what reasons lie behind the inconsistent results which some authors have found. Rather, we will focus on an aspect of the speech recognition task where the prognosis for auditory modelling promises to bear some fruit, namely, speaker differences (in speaker-independent speech recognition).

Speaker normalization for vowels

Normalizing the acoustic differences found between speakers - to take the best known example, differences of formant frequency in male and female vowels - used to be a formidable prospect. Fant showed how formant frequency differences were not just sex-specific, but also formant-specific and vowel-specific too. Methods of normalizing these data based upon reconstructing vocal tract shape fell foul of the problem that the solution to this exercise is nonunique. But if we apply auditory insights to the question, and compare not measured acoustic formants but auditorily transformed spectra, it can be shown that the nonlinearities which plagued Fant's data largely disappear. We argued (in Bladon et al., 1984) that the application of an auditory model which includes an auditory filter and a Bark scale, together with a displacement notion which has a simple physiological analogue, combine to generate a high degree of spectral match between male and female vowels. A large quantity of data, assembled by us and by others across a range of dialects and languages, has broadly supported this contention. Examples of vowels normalized in this way can be found in the above reference, and will be shown in the symposium.

How far is it worthwhile to extend an auditory model of speaker normalization beyond the vowel sounds? The theoretical answer seems to be: in part. At the present stage of research this answer has to be arrived at largely by inference from scattered pieces of the work of others, supported by some sporadic experimental confirmation of our own.

Voiceless vowels

Schwartz and Rine (1968) demonstrated that listeners could confidently identify a speaker's sex from individual steady-state vowels which were whispered. This is a finding of interest because it demonstrates that the role of voice pitch in speaker normalization is not a necessary one (though this does not exclude the possibility that pitch may have an ancillary role). As a result, the spectral characteristics of the whispered vowels are firmly implicated as a source for the listener's ability to identify sex.

Transforming the Schwartz and Rine whispered vowel spectra into auditory representations enables us to judge the effect of normalizing them by our method. It turns out that this procedure neutralizes much of the male/female difference. Whispered vowels, then, should be encompassed straightfowardly in an auditory model for speaker-independent ASR. It is not just voiced sounds which differ across speakers.

Plosives

This being so, what of plosives (voiced and voiceless)? The burst spectrum, widely believed to be of service as a differentiator of place in plosives, appears not to be a candidate for normalization. This statement derives from work in progress at Amsterdam by Weenink and remains to be fully confirmed. Weenink is finding that, while the plosive burst spectrum is sufficient to identify the plosive place in 85% of cases (thus corroborating the position long held by Stevens and others), listeners cannot identify the speaker's sex from the burst spectrum. When we recall the well-known templates for burst spectra, it is not difficult to guess why plosive bursts carry so little speaker information. The burst spectra are very variable, partly due to phonetic context; consequently the templates which fit each plosive are large, extensive both in frequency and in amplitude.

Even so, there is some evidence that normalization is appropriate for plosives, in respect not of their bursts but of their transitional spectra. This evidence comes from both production (O'Kane, 1984) and perception (Rand, 1971). Rand showed how, in a synthetic plosive-vowel sequence, the onset of formant transitions was at a frequency position which varied with speaker type. (His speaker types were "a large vocal tract" and "a small vocal tract".) He deduced that the same applies to the plosive locus frequency. In fact, unnoticed by Rand, the average [d] onsets needed to be 1.1 Bark different. It is striking, and unlikely to be coincidental, that this difference is reminiscent of an auditory displacement of the same magnitude which we have been discovering in vowel sounds.

The second piece of evidence is the measurements by O'Kane (1984) of locus frequencies, from the Australian English plosives spoken by 5 males and 5 females. She reported the overall locus ranges only in fairly gross terms: and, of course, ranges can give a misleading picture of the typical behaviour. Nevertheless, once again, when converted to a Bark scale, the female measured plosive loci can be seen to exceed the male values by a generally constant amount. One Bark would be a representative value. And so, while noting that plosive transitions have so far been only superficially investigated, it may be concluded that plosive transitions look like conforming to the normalization pattern.

Liquids and nasals

For many other classes of speech event there is at present no known evidence which would indicate how far, if at all, they are susceptible to variation with speaker-type, and hence, how far normalization is called for. This applies to laterals, nasals and trills, for instance. *Prima facie*, since these sounds have a prominent spectral content, they may possibly also carry the speaker-type information in a similar way to vowels. Alternatively, it may be that the spectral content in a nasal, with its large number of heavily damped formants, may be too elusive to have a clear auditory image which could be used in a normalization role. Pending further work, these matters have to be left open.

Fricatives

For fricatives, on the other hand, there is some well-documented evidence. Initially we will consider just the sibilant fricatives such as [s, ʃ, ç]. Schwartz (1968) published illustrations of speaker sex difference among voiceless English [s] and [ʃ]. Once again, we find that a conversion to auditory spectra leads to a greatly improved congruence of spectral shape.

Male and female [s] spectra were also investigated by us in British English. From a tightly controlled database and in an identical linguistic context, 55 male tokens (from five speakers) were compared with the same number of female tokens. Auditory spectra of these fricatives confirmed the tendency to congruence noted in the Schwartz data and further revealed that an especially constant feature of [s] was the (15 phons/Bark) low-frequency edge of the [s] peak. As with vowels and other sounds, this edge is so located as to suggest a constant male/female normalizing factor in auditory space.

Whether this behaviour extends to fricatives other than the sibilants mentioned is currently a matter of uncertainty: the basic work remains to be done. A fairly confident summary would be as follows. It is known from the study by Ingemann (1968) that speaker sex is identifiable from steady-state productions of the glottal fricative [h], with an accuracy comparable to that of the sibilants. Also identifiable at better than chance accuracy, according to the same study, are uvular [χ] and velar [x]. Spectra of these back fricatives show a somewhat vowel-like superimposition of vocal tract cavity resonances, and hence will be expected to behave in speaker normalization very much as vowels do. This is especially likely of [h] since the resonance patterns will not differ markedly from those of a whispered vowel. On the other hand the front fricatives [ɸ, f, θ] are not identifiable for sex. This is understandable, given that the front fricatives with little or no resonance cavity ahead of their friction source, do not have a very distinctive spectral shape. Intensity level is their prime cue. Speaker sex differences do not seem to exploit this.

Conclusion

Extrapolating somewhat beyond the rather superficial review above, it seems reasonable to say that, as a useful basis for speaker-independent ASR, an auditory model can in general be used to normalize the running-speech spectral shape. Fairly clear exceptions to this are the front fricatives (those which are more advanced than alveolar) and the plosive bursts, whose spectra appear not to be capable of signalling information on speaker type.

If this is so, then in an actual speech recognition system two empirically testable alternatives can be explored. One is the possibility that a decision on whether or not to normalize the currently incoming spectrum for speaker differences must be made, depending on a decision about its phonetic class. This alternative clearly implies the intervention of some higher-level expert. The other possibility is that no such decision needs to be made at all: the recognizer can safely normalize the whole signal, because those phonetic classes of event which do not show evidence of sex-based physical difference are anyway spectrally rather flat or heavily smeared.

In order to choose between these alternatives we propose to examine recognition test results to see whether (or how far) deterioration ensues, when the whole set of phonetic events in speech (as opposed to a partial set excluding front fricatives and plosive bursts) is first subjected to an auditorily-based normalization for speaker sex.

References

BLADON R.A.W., HENTON C.G. and PICKERING J.B. (1984a). Outline of an auditory theory of speaker normalization. In Van den Broecke M.P.R. and Cohen A. (eds.), Proceedings of the Tenth International Congress of Phonetic Sciences (Dordrecht, Foris), 313-317.

BLOMBERG M., CARLSON R., ELENIUS K. and GRANSTROM B. (1984). Auditory models in isolated word recognition. IEEE ICASSP 1984, 17.9.1-17.9.4.

INGEMANN F. (1968). Identification of the speaker's sex from voiceless fricatives. J. Acoust. Soc. Am. 44, 1142-1144.

O'KANE M. (1984). Extensions to the locus theory. In Van den Broecke M.P.R. and Cohen A. [see Bladon above], 331-337.

RAND T.C. (1971). Vocal tract size normalization in the perception of stop consonants. Haskins Labs. Stat. Rep. Sp. Res. 25/26, 141-146.

SCHWARTZ M.F. (1968). Identification of speaker sex from isolated voiceless fricatives. J. Acoust. Soc. Am. 43, 1178-1179.

SCHWARTZ M.F. and RINE H.E. (1968). Identification of speaker sex from isolated whispered vowels. J. Acoust. Soc. Am. 44, 1736-1737.