EFFICIENT DERIVATION OF FORMANT-LIKE INFORMATION FROM SPEECH WAVEFORMS

M. O'Kane and J. Gillis

School of Information Sciences and Engineering
Canberra College of Advanced Education

ABSTRACT - A computationally fast technique for deriving spectral information including formant information from speech waveforms is presented. Its use in automatic recognition of vowels and some consonants in continuous speech is briefly described and the relationship of the spectral display obtained using the new technique to the Bark scale is discussed.

INTRODUCTION

In interviewing one of the expert phoneticians working on the FOPHO project (O'Kane, 1983) we discovered that one of his specialisations was an ability to decipher printed speech waveforms in that he was able to perform rough phoneme-level segmentation of the plotted waveform and could label the phoneme segments with various perceptually meaningful features. These features include the 'obvious' ones of fundamental frequency (and hence likely sex of the speaker), rate of speaking (time-scale is provided), and locations of stressed vowels, fricatives and silence. More interestingly however, our phonetician can also give the approximate identity (place of articulation) of vowels and fricatives and can often label laterals, nasals and plosives as such. Where he is uncertain of the manner and place classifications of a sound he can produce a likely set of alternatives. Where he is uncertain of the number of phonemes in a particular segment of speech he can, by reasoning about the length of time involved and the speaker's typical speaking rate, put upper and lower bounds on the number of phonemes present. Furthermore, he can often spot peculiarities in phonation such as a creaky voice.

We decided that an attempt to capture this waveform reading expetise would provide a useful alternative recognition knowledge source to the LPC-based FOPHO algorithms. This paper outlines results obtained when investigating extraction of information for vowel recognition.

SIGNAL PROCESSING TECHNIQUES USEFUL FOR WAVEFORM READING

To capture the expert's knowledge we asked the phonetician to hand-segment and label a large number of passages of continuous speech produced by a variety of speakers from the Australian English database (O'Kane, Millar and Bryant, 1983) and furthermore we asked him to provide a rationale for his segmentation and labelling decisions. This rationale seemed to often be given in terms of features relating to both zero-crossings per unit time and waveform amplitude information.

To extract these features automatically we developed several related waveform analysis techniques the outputs from which were used for specialised feature extraction. One technique which led to a computationally fast segmentation algorithm has been described (O'Kane, Gillis, Rose and Wagner, 1986). Another useful function is the function we call M1. This function is defined as the inverse of the time between adjacent points (called W1 points) which are themselves mid-way in time between valleys and peaks in the positive-gradient sections of the speech waveform. For speed of computation M1 is actually computed by noting the number of waveform samples between W1 points and assigning as the frequency value associated with the M1 point a pre-computed value equal to the sampling frequency divided by this number. The time between W1 points will always be equal to integral multiples (by an integer ≥ 4) of half the sampling period - that is why the frequency values can be pre-computed and that is why in Figures 1(a) and (b), both of which are examples of M1 for speech sampled at 20kHz, one notices hyperbolically quantised striations (particularly noticeable in the top of the diagrams where the striations are most separated). The time value associated with an M1 point is the time mid-way between the time values of the W1 points used to obtain the M1 frequency.

In Figure 1(a) the M1 graph for the front diphthong /eɩ/ is shown while Figure 1(b) displays the M1 graph for the low back vowel /ɔ/. As can be seen from the figures, for the front vowel region M1 has no values below about 950 Hz and a scatter of values above 950 Hz. The back vowel, on the other hand, is characterised by having almost all its M1 values in the range 500-1400 Hz. Furthermore, the M1 graph for the back vowel displays relatively stable 'tracks', the first two of which correspond to the formant tracks for this vowel. Other interesting features of the M1 graphs are that examples of /s/ are characterised by having no values below about 4400 Hz while examples of nasals, apart from the devoiced section, have no values above 950 Hz.

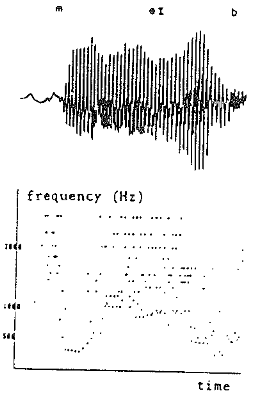


Figure 1(a): The function M1 for speaker 18 producing /meib/ from "may be".

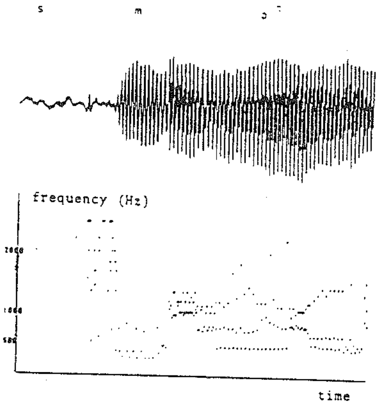


Figure 1(b): The function M1 for speaker 18 producing /smɔ/ from "small".

RELATIONSHIPS TO PERIPHERAL AUDITORY PROCESSING

The hyperbolic quantisation of the frequency scale in Figure 1 is somewhat reminiscent of the Bark scale of frequency which has been used for depicting spectral information in disucssions (cf. Bladon and Lindblom, 1981) of theories of peripheral auditory processing based on the concept of critical bands (Zwicker, 1961). Noting this we described the information displayed in Figure 1 in another way. In the diagrams in Figure 2 the number of dots in each striation in M1 style displays taken across the steady-state region of a vowel have been counted and the overall number of points is normalised out of 100. For ease of display every second striation band is represented; points from non-displayed bands have been assigned to neighbouring displayed bands and the result has been re-normalised. Examples of such 'averaged M1' displays for a front, a central and a back vowel produced by a male speaker of Australian English are given in the figure. Such averaged M1 displays are typical of the type of patterns seen across all male speakers studied so far (15 male speakers each reading 11 different /h-d/ words). From a strictly automatic speech recognition approach a method of distinguishing between these different · patterns would be sufficient (cf. Kuhlwetter's (1978) vowel recognition algorithm also using a waveform-based technique). However as many of the other algorithms to be incorporated into FOPHO are formant or formant track based and because of the number of formant based studies in the speech literature, we decided that formant extraction was desirable even from a speech recognition point of view. This is discussed in the next section.

As in displays of spectral information using a Bark scale the information in the region of low frequency in the diagrams in Figure 2 is much more finely quantised than it is in the high frequency region. However, the axis scale in these diagrams is not the Bark scale as can be seen by considering Figures 3(a) and (b). Figure 3(a) is a graph of critical band number or Bark versus frequency. Figure 3(b) is a graph of striation band number versus frequency for 10 kHz sampling where the striation band

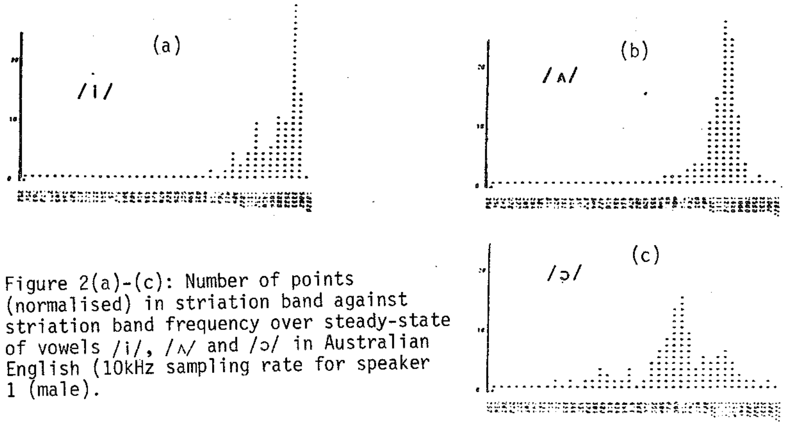
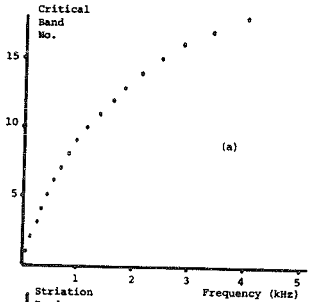
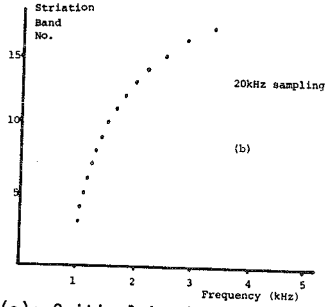


Figure 2(a)-(c): Number of points (normalised) in striation band against striation band frequency over steady-state of vowels /i/, /ʌ/ and /ɔ/ in Australian English (10kHz sampling rate for speaker 1 (male).





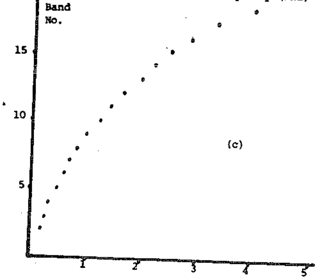


Figure 3(a): Critical band number versus associated frequency - data from Zwicker(1986).

Figure 3(b): Striation band number versus associated frequency for 10kHz sampling.

Figure 3(c): Striation band number versus associated frequency for mixed sampling rates.

(5-2 kHz - 10 kHz sampling
2-1 " - 5 " "
1-.67 " - 2.5 " "
.67-.25 " - 1.25 " ")

numbering is assigned with a view to making the fit to the Bark Scale as good as possible. In an attempt to obtain a better fit to the graph in Figure 3(a), we considered the graphs of striation band versus frequency associated with sampling rates lower than 10 kHz. We found that the best fit to the graph in Figure 3(a) was the one displayed as Figure 3(c) which while still being striation number versus frequency is a composite of the striation number versus frequency for 10 kHz sampling from 5 kHz down to 2 kHz, for 5 kHz sampling from 2 kHz to 1 kHz, for 2.5 kHz sampling from 1 kHz down to .67 kHz and 1.25 kHz sampling from .67 kHz to .25 kHz. It will be noted that the graph in Figure 3(c) is a very good approximation to Figure 3(a) apart from the sampling rate changeover points.

RECOVERING FORMANT INFORMATION

Studies of composite M1 displays obtained using the composite striation band scale of Figure 3(c) were rather inconclusive in the sense that such composite M1 displays did not look particularly close to the Bladon and Lindblom (1981) spectra. However in deriving these composite averaged M1 spectra we studied the standard averaged M1 spectra (diagrams of the type displayed in Figure 2) for the same piece of speech sampled at several different frequencies (obtained by downsampling) and noted that the vowel formants seemed to 'appear' in the displays for various sampling rates and not to be so prominent in displays for other sampling rates. This led us to an attempt to recover formant information from averaged M1 displays. In this context it should be noted that studies by Young and Sachs (1979) have shown that the auditory system is capable of locating vowel formants.

Our algorithms for calculation of the first and second formants are as follows. The first formant is calculated from the normalised M1 display associated with 1,666 Hz sampling as follows:

First formant =

$$\frac{\sum\limits_{\substack{\text{all}\\ \text{striation bands}}} (\text{number of points in striation band x striation band frequency})}{100}$$

The algorithm for the second formant is the same except for the fact that 10 kHz sampling is used.

Using these algorithms the results for the average vowel formants for 15 males and 18 females obtained by this method are shown in Figure 4. For comparison the formant values got by applying the McCandless algorithm (McCandless, 1974) to FFT and Linear Prediction analyses are given in Figure 5. The match in the front vowel region is not as good as the back vowel region - we have some evidence that 20 kHz sampling might be more appropriate in the second formant algorithm. Nevertheless the same overall trends are to be seen between the two types of formant extraction. Again it must be emphasised that our technique for formant extraction is computationally much faster than the FFT/LPC/McCandless method.

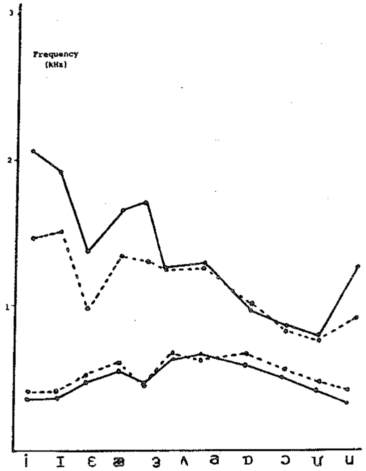


Figure 4: Average for 15 male speakers (dark lines) and 18 female speakers (dotted lines) of the first and second formant values for vowels in Australian English (high front to low back). Formant extraction method used is the technique described here.

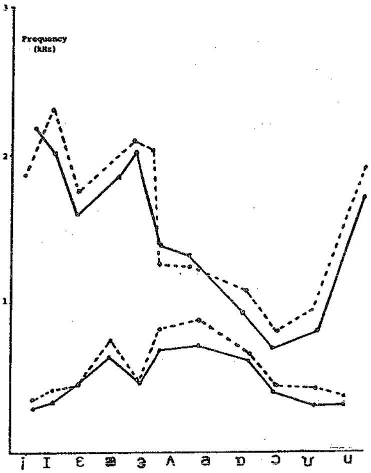


Figure 5: Average for 15 male speakers (dark lines) and 18 female speakers (dotted lines) of the first and second formant values for vowels in Australian English (high front to low back). Formant extraction used is FFT/LCP/McCandless.

Formants extracted using our algorithm and vowel durations obtained from the segmentation algorithm (O'Kane et al. 1986) provide computationally inexpensive input to a previously developed vowel recognition algorithm for Australian English (O'Kane, 1981). Tested on the same material as was used for testing the segmentation algorithm, this vowel recognition algorithm correctly classified all vowel segments according to broad tongue height descriptions (high, mid, low), broad front-back criteria (front-central and central-back) and dual length criteria (short and long).

CONSONANT RECOGNITION

Recognition algorithms for three classes of consonants - nasals, fricatives and plosives - have also been developed using the averaged M1 function. However we have not as yet perfected a recognition algorithm that is not speaker dependent for identifying the nasal consonants. For any given speaker the distinction between different nasal consonants is easy to define but generating an algorithm to work across speakers is difficult. But then the speaker idiosyncrasy of nasals is what makes them so useful in speaker recognition algorithms so maybe this result is not too surprising.

However it was relatively simple to devise a speaker-independent algorithm to distinguish between fricatives associated with each place of articulation. Overall the fricatives had an 82% correct classification rate when tested on a one minute reading passage read by 5 speakers.

REFERENCES

BLADON, R.A.W. & LINDBLOM, B. (1981) "Modeling the judgement of vowel quality differences", J. Acoust. Soc. Am., 69, 1414-1422.

KUHLWETTER, J. (1978) "Vowel recognition in the time domain", Proceedings of the 4th International Joint Conference on Pattern Recognition, Kyoto.

McCANDLESS, S.S. (1974) "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Transactions Acoustics Speech and Signal Processing, AASP-22, 135-141.

O'KANE, M. (1981) "Acoustic-phonetic processing for continuous speech recognition", Ph.D. thesis, Australian National University.

O'KANE, M. (1983) "The FOPHO Speech Recognition Project", Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, 630-632.

O'KANE, M., MILLAR, J.B. & BRYANT, P. (1983) "A Database of Spoken Australian English : Design and Collection", Technical Note No.6, School of Information Sciences, CCAE.

O'KANE., GILLIS, J., ROSE, P. & WAGNER, M. (1986) "Deciphering speech waveforms". Proceedings of IEEE-IECEJ-ASJ International Conference on Acoustics, Speech, and Signal Processing, Tokyo, 2227-2230.

YOUNG, E.D. & SACHS., M.B. (1979) "Representation of Steady-State Vowels in the Temporal Aspect of the Discharge Patterns of Populations of Auditory-Nerve Fibers", J. Acoust. Soc. Am., 65, 1381-1403.

ZWICKER, E. (1961) "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)", J. Acoust. Soc. Am., 33, 248.