Proceedings of The Institute of Acoustics

SECTORGRAPHS - A NEW WAY OF LOOKING AT SPEECH

R LINGGARD AND D RANKIN
DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
QUEENS UNIVERSITY OF BELFAST

INTRODUCTION

The traditional speech spectrogram is, perhaps, the most graphic way of presenting speech signals in a form easily assimilated by the human eye. It provides a means of looking at long segments of speech, and of visualising spectral energy transitions, which are so meaningful and characteristic of speech sounds. Indeed, some research workers have even taught themselves to "read" spectrograms with impressive accuracy. Unfortunately, spectrograms do not lend themselves readily to scientific analysis. Consisting, essentially, of black smudges on gray paper, they are meaningful to the human eye, but are otherwise not easily quantifiable. The problem is that, to view speech spectra in time, requires a three-dimensional display of the variables, frequency, amplitude and time. Creating the sequences of short-time spectra is easy enough the difficulty is essentially one of display.

SECTORGRAPHS

An alternative method of looking at long sequences of speech spectra is to sectionalise each spectrum and display it as a SECTORGRAPH. Fig 1 shows a typical speech spectrum which has been sectionalised into 20 sections, each of equal area. That is, the area of each section is equal to 1/20 the total area under the spectrum. Performing this operation on successive spectral frames, and plotting the sector boundaries on a two-dimensional graph, results in a "sectorgraph", an example of which is shown in Fig 2.

More formally, the short-time speech spectra are assumed to consist of a set s(n), of amplitude samples, at points n, (n=1 to N), along the frequency axis. We define S, the total area of the spectrum, as

$$S = \sum_{N=1}^{N} s(n)$$

The spectrum is divided in to P sectors, such that the sector boundaries n(i), are given by the relation

$$a(i) = \frac{1}{S} \sum_{n=1}^{n(i)} s(n)$$

Thus, a(i) is the ratio of the area under the spectrum, up to the i.th sector boundary, to S, the total area.

From the diagram of Fig 1, it will be obvious that

$$n(i-1) < n(i) < n(i+1)$$
and
$$a(i-1) < a(i) < a(i+1)$$

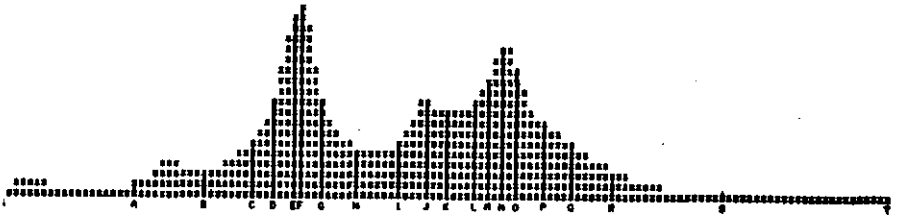
The spectral amplitudes s(n), are assumed to be on a linear scale, and the frequency sampling parameter n, may be proportional to either frequency or log of frequency.

Though the definition of the sector boundaries n(i), is somewhat complicated, in practice their computation is quite simple. The total area S, is found by summing the spectral samples s(n), and the n(i) are calculated by summing s(n) until the sum equals a(i).S. In the interests of accuracy, it is necessary that N the number of spectral samples be much greater than P, the number of sectors.

SHAPE INVARIANCE

It is obvious that the sector boundaries are invariant to changes in the overall amplitude of the spectrum. This is because the n(i) are calculated from fixed fractions of the total spectral area. However, an additional degree of shape invariance may be obtained. Linear shifts of a basic spectral shape along the frequency axis, will result only in the addition of a constant to each of the sector boundaries. This constant will disappear if sector differences are used instead of sector boundaries. Furthermore, on a log frequency scale, linear shifts of spectral shape correspond to multiplicative changes in the frequency variable, and it is well known that differences in vocal tract length manifest themselves in this manner. Thus, graphs of sector differences may be useful in the normalisation of speech spectra, since they are invariant to two major speech variables, loudness and vocal tract size. This property is illustrated in Fig 3.

FIG 1.



Typical spectrum of voiced speech. The letters A B C etc denote sector boundaries. Twenty, equal area sectors are shown.

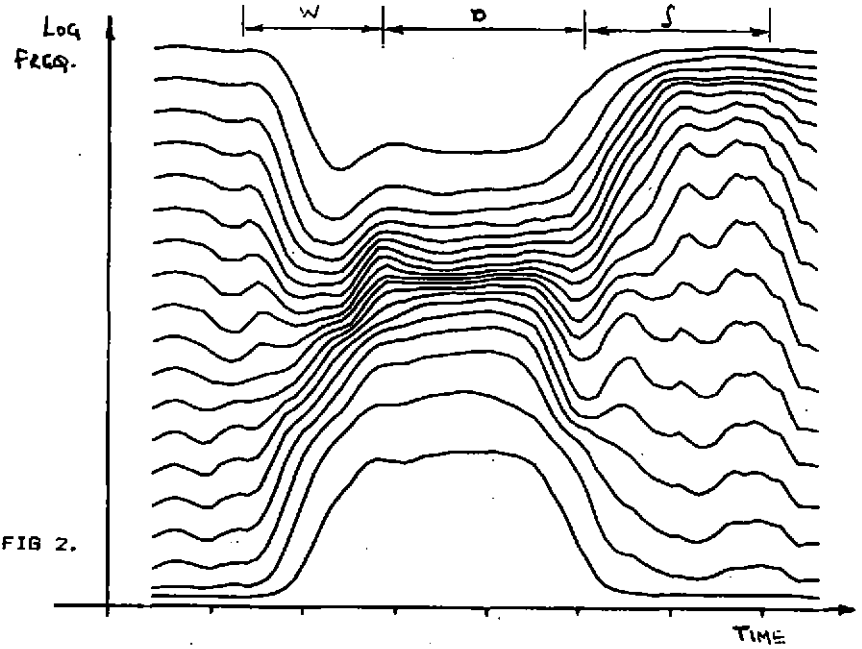


FIG 2.

Sectorgraph of the utterance "wash", female speaker. There are twenty, equal area sectors, on a log frequency scale.

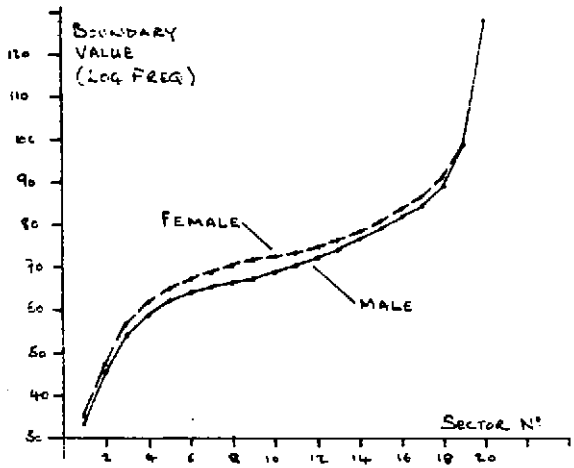


FIG 3(a). Sector boundary values for single frames of male and female speech. The frames are taken from the vowel nucleus of the utterance "wash". Amplitude differences in the two frames are already normalised out.

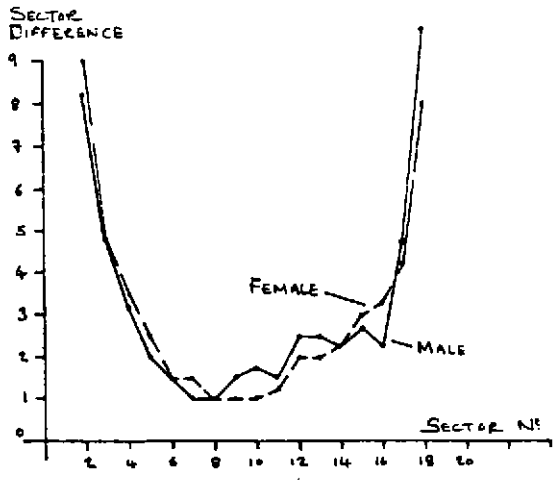


FIG 3(b). Sector differences for the two frames shown above. The vertical scale is times ten to emphasise differences. Four out of the eighteen values are identical.