DETECTION OF PERIODIC PITCH IN SPEECH SIGNALS

W MILLAR & R LINGGARD
DEPARTMENT OF ELECTRICAL & ELECTRONIC ENGINEERING
THE QUEEN'S UNIVERSITY OF BELFAST

1. INTRODUCTION This contribution describes a method of extracting the pitch frequency from the harmonic structure of the short-time spectrum of speech. A harmonic histogram is formed and pitch is detected as the frequency of the largest component. The spectra used are obtained from a digital filter bank (implemented in soft-ware) having 187 channels covering the frequency range 50 Hz. to 3200 Hz., on a log. frequency scale designed to include all 16 harmonics of a nominal 50 Hz. to 200 Hz. male pitch range.

2. FILTER CHARACTERISTICS Each digital filter is a realisation of a second-order, sampled-data, band-pass filter. The input speech waveform is sampled at 15 kHz., thus enabling a complete spectrum to be obtained from the rectified outputs of the individual channels every 67 μS. Since they are second-order, the frequency responses of adjacent channels overlap to a certain extent. The filter bank is designed such that the attenuation at the overlap point of adjacent filters is the same for all filters. However, the actual level of this overlap attenuation may be varied to increase or decrease the spectral resolution.

The frequency scale of the filter bank is chosen so that the harmonics of possible pitch values coincide to a large extent, thereby minimising the number of channels required. This implies that the frequency scale is such that harmonics of each value in the scale are also values in the scale, similar, in fact, to a musical scale. Thus one solution would be to use the 12 note chromatic scale but this does not give sufficient resolution. A frequency scale with greater resolution is required, say N values or notes per octave, where $N > 12$. On a logarithmic scale, the frequency separation x, between individual notes in the scale is given by $\ln x = \ln 2/N$.

Thus if f_0 is the base note, then all other notes in the scale are given by the relation

$$f_n = x^n f_0,$$

where f_n is the frequency of the nth note in the scale. The mth harmonic of f_n will be $mf_n = mx^n f_0$. If this is also a note in the scale (say f_r) then

$mf_n = mx^n f_o = f_r$ so that $mx^n f_o = x^r f_o$ and $m = x^{r-n}$. Hence $(r-n) = N \frac{\ln(m)}{\ln(2)}$.

The number $(r-n) = h$, is the harmonic interval, ie the number of notes between a note f_n and its mth harmonic mf_n. This in turn means that h must be an integer, if mf_n is to be a valid note in the scale. If values of h are calculated for various values of N and m, it is found that it is not possible to meet this requirement and the criterion is relaxed to allow h to be an integer plus or minus a small fraction. An optimum solution exists for the case $N = 31$. Thus if the pitch range is taken to be a nominal 50 Hz. to 200 Hz., then for 16 harmonics of pitch to be included, the frequency range of the filter bank is 50 Hz. to 16 x 200 = 3200 Hz. The 6 octave range is covered by a total of 6 x 31 + 1 = 187 filters.

3. HARMONIC HISTOGRAMS The spectra produced from the filter bank can be used to calculate the harmonic histogram by means of the calculation

$$H_p = \sum_{m=1}^{16} F\ (mf_p),$$

where H_p is the histogram component for a possible pitch value f_p and $F(mf_p)$ is the spectral component at frequency mf_p. This component may be either the linear or logarithmic value of the filter output.

The resolution of the individual filters has a considerable effect on the shape of the spectra and hence on the resulting harmonic histograms. The nature of the harmonic histogram results in peaks being produced at frequency values corresponding to pitch, twice the pitch frequency and half the pitch frequency. It was found that the best histograms ie those with the clearest indication of pitch, were produced when the frequency responses of adjacent filters overlapped at 3dB below their central maximum and the histograms were calculated from the log. outputs.

4. SHORT SPECTRA It is well known that the pitch frequency of a section of speech may be perceived even though the fundamental frequency is absent. This is particularly evident in the case of a male speaker communicating over a telephone line. It seems reasonable to suggest that the ear uses the higher harmonics present in the speech signal in order to obtain the pitch. Thus it is interesting to examine the performance of the harmonic histogram in this respect.

The easiest way to band limit the speech is to truncate the spectra at the low frequency end, leaving only the 4 octave range from 200 to 3200 Hz. Harmonic histograms calculated from such "short" spectra gave a good indication of pitch, although this was less marked than with the "long" spectra. The advantage of this technique is that the number of filters required is now 125. (Ref. 1).

5. INFLUENCE OF FORMANTS In some cases the harmonic histogram is distorted by the presence of formant peaks in the spectrum. These act to give extra emphasis to lower harmonics, so that even the troughs between harmonic peaks at low frequencies are greater than the peaks themselves at high frequencies. To overcome this it is necessary to remove the formant structure from the spectrum, which in turn requires a knowledge of the formant structure.

An approximation to the formant structure may be obtained from the spectrum by a process of smoothing. The easiest way to do this is to "view" the spectrum through an aperture several channels wide. In this way the smoothed spectrum consists of components B_i such that

$$B_i = \frac{1}{a} \sum_{j=i-\frac{a-1}{2}}^{i+\frac{a-1}{2}} A_j\ .$$

where A_j are the components of the original spectrum and a is the width of the aperture. It is convenient to make a an odd number, tests showed that a = 9 gave best results.

An enhanced spectrum is formed by subtracting, on a log. scale, the smoothed spectrum from the original spectrum. The harmonic histogram calculated from this, gives a peak at the frequency corresponding to the pitch frequency.

6. PITCH CONTOURS The histograms discussed so far have been obtained from instantaneous spectra produced at 67 μS. intervals. It is more useful to have pitch contours of an utterance and in this respect it is neither necessary nor desirable to have the pitch information available at so high a rate.

There are two possible methods of reducing this information rate to a more acceptable figure of, say one pitch estimate every 10 ms. The pitch histograms may be calculated as before and an average value output every 10 ms. or the spectra may be averaged over a similar time interval and a single histogram calculated from this. A comparison of the two methods shows that the latter is marginally superior. This is a convenient result, since the processing load is dramatically reduced if this method is adopted.

7. COMPLEX SOUNDS WITH NARROW BANDWIDTH It is well-known that when the ear is presented with a narrow bandwidth signal, as in the case of an amplitude modulated tone, a low frequency pitch may be perceived, even though no corresponding low frequency is physically present. Schouten (2) has also reported the phenomenon, known as the first effect of pitch shift, that if the carrier frequency is varied, then the perceived pitch also changes. This effect may be due to a perceptual mechanism which uses harmonic information to estimate pitch.

The pitch extraction method described above was used to calculate histograms of an amplitude modulated tone. The output from the period histogram produced similar trends to those described by Schouten.

8. CONCLUSION A method of extracting pitch from a short-time frequency analysis of speech has been presented. This method works equally well with band-limited speech or with speech containing the fundamental. Occasional errors are produced by the presence of the formant structure in the spectrum and a method of overcoming this has been presented. The performance of this pitch detection algorithm with amplitude modulated tones, shows similar results to the well known "first effect of pitch shift".

REFERENCES

1. R. LINGGARD and I. BOYD 1979 Proc. Inst. Acoust. Autumn Conf. 45-48.
Real-time Spectral Analysis of Speech.

2. J.F. SCHOUTEN 1940 Philips Techn. Rev. 5 286-294
The perception of pitch.

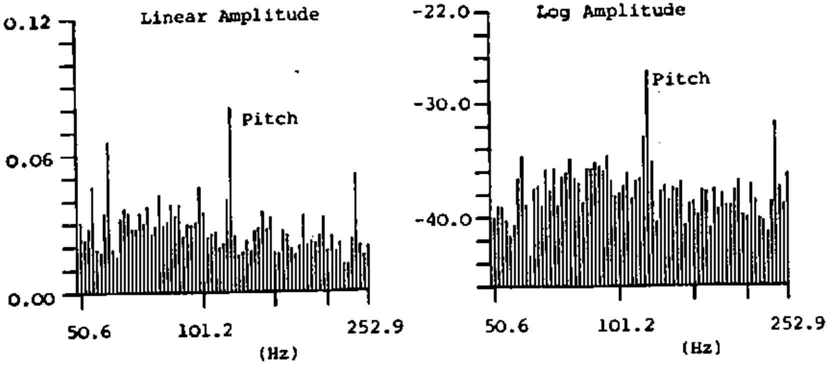

Fig 1 Comparison of Linear and Log amplitude weighted Histograms.

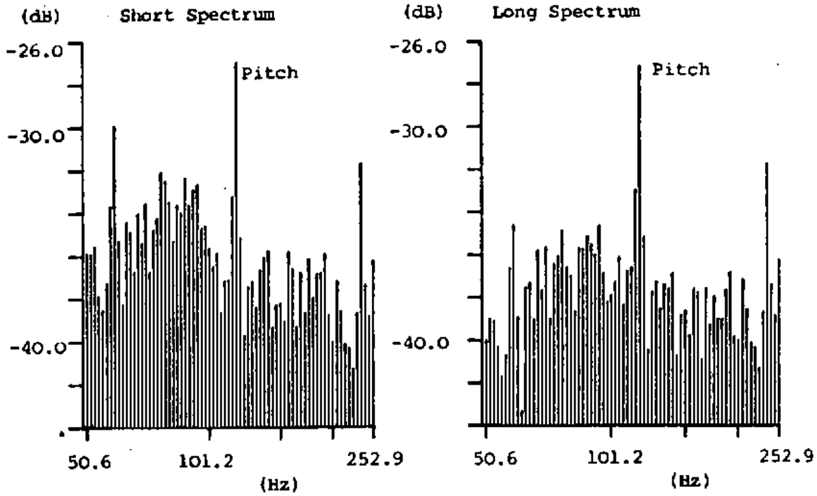

Fig 2 Comparison of Log weighted Histograms for 'Short' and 'Long' Spectra of same voiced utterance.