PERIODICITY ESTIMATION IN THE PRESENCE OF NOISE

M J IRWIN

JOINT SPEECH RESEARCH UNIT

INTRODUCTION

The reliable estimation of the periodicity of a time series is an important requirement in many branches of science. Most practical signals of interest (for example speech signals, which are normally represented in sampled data form as a time series) are either only stationary over a fairly short time or are only available for a short duration. The techniques for making measurements of the periodicity fall into two main categories. For periodic signals where an accurate understanding of the underlying physical process is available, direct modelling of the spectrum, taking into account all prior information about the system, is possible and indeed desirable (for example: all pole modelling). When the spectral model is not so well defined,however, or the harmonic content of the signal is extensive and not amenable to simple analysis, the estimation is normally based either directly or indirectly on the autocorrelation function of the time series. For speech signals it is this latter approach which is normally used. The question to be considered in the following sections is what is the optimum method for estimating the periodicity of a speech signal when little prior information is available or being made use of. In this context we are excluding the use of 'conditioning' filters and will assume that the necessary band limiting or simple spectral shaping of the signal has already been done.

GENERAL COMMENTS

There are two aspects to the problem of determining the periodicity of a speech signal, namely recovering the fundamental frequency of the voiced speech and estimating the degree of voicing inherent in the speech. These two aims are not necessarily equivalent. A method which is optimal in estimating the periodicity of the signal may not necessarily be optimal in recovering the actual period. This is due to the ambiguity present in selecting the correct period from one of a series of multiples and submultiples of the fundamental. This problem occurs to some extent in all autocorrelation based methods and can be minimised, at the expense of altering the noise spectrum, by enhancing the overall harmonic structure of the signal. For speech signals this entails removing the influence of the vocal tract transfer function and the spectral effects of the glottal pulse shape.

These difficulties lead us to consider the periodicity estimation process to be of three stages. In the first stage, which may not always be necessary, some form of preprocessing is used to give a better estimate of the desired signal. This step depends on having some limited prior knowledge of the mechanism which generated the signal. For speech this is frequently the approximation that the signal can be described as the convolution of the vocal tract transfer function with the excitation function. The second step consists of optimally estimating the periodicity of the processed signal on a frame by frame basis.

Finally, these individual frame measures are suitably combined to form an estimate of the 'time track' of the fundamental frequency.

As lack of space precludes a full discussion of tracking algorithms a few words about tracking are appropriate at this point. One simple and powerful technique for fundamental frequency tracking is to use a dynamic programming approach to produce and monitor possible trajectories back through preceding frames. This relies on a suitable periodicity measure combined with a continuity score which takes account of the percentage changes in pitch from one frame to the next. The 'best' trajectory through the preceding frames is then automatically available.

OPTIMAL SIGNAL PREPROCESSING

In a large variety of signal processing problems the quantity we wish to investigate, p(t), (the periodicity in this case) is convolved with some system function, v(t), (the vocal tract transfer function) to produce the signal that we actually measure. If we have additive corrupting noise present, n(t), this system covers a wide range of speech processing problems. Such a system is defined by the equation

$$s(t) = p(t) \circledast v(t) + n(t)$$

where $\circledast$ denotes the convolution operation.

Weiner (1949) and others have shown that a better approximation to p(t) in cases such as this can be obtained by preprocessing the signal with a linear filter h(t). The new signal y(t) is governed by

$$y(t) = h(t) \circledast s(t)$$

and the criterion for selecting h(t) is to minimise the expected value of the square of the error term y(t) - p(t) with respect to the filter h(t), otherwise known as the Weiner deconvolution filter. The optimum filter h(t) is defined by

$$H(w) = P(w)S^{*}(w) / |S(w)|^{2}$$

where H(w), P(w) and S(w) are amplitude spectra and * denotes complex conjugate.

The filter H(w) is the optimal filter to use for recovering the excitation function of the speech. For low noise levels this filter is the inverse filter $V^{-1}(w)$. There are several practical methods for approximating the effect of the inverse filter $V^{-1}(w)$. These include linear prediction analysis to directly estimate this filter or alternative approaches which directly 'flatten' the spectrum, for example: by forming its logarithm (in the Cepstrum method) or by using non-linear operations on the signal itself to achieve the same effect. In separating out the vocal tract response from the excitation function all of these methods are essentially approximating the use of a deconvolution filter.

However,when the noise power is not negligible we can in general only make crude approximations to H(w). Two commonly used approximations are: no filter is applied or the filter is set equal to an estimate of the inverse filter $V^{-1}(w)$. Even assuming that we have an accurate estimate of V(w) which is not always true, by applying this approximation for the optimal deconvolution filter we may have increased the overall level of noise present in the system since the spectrum of the preprocessed signal y(t) will now be governed by

$$Y(w) = P(w) + N(w)/V(w)$$

Although the excitation function has been separated out from the vocal tract transfer function the noise spectrum has been considerably altered. Quite often this effect will far outweigh any advantage gained by using the filter. Thus, it is apparent that the use of signal preprocessing should be dependent on the corrupting noise level present. For low noise levels preprocessing should be used but for significant noise levels it is probably better to estimate the periodicity directly from the input signal.

LEAST-SQUARES PERIODICITY ESTIMATION

After the first stage of processing we will be left with a 'speech signal', s(t), which in general will be composed of a periodic part, $s_\theta(t)$, and a non-periodic or noise part, n(t). The noise-like component will contain contributions from the speech process itself and from any external noise present in the system. If we make the assumption that

$$s(t) = s_\theta(t) + n(t)$$

then an optimum least-squares method for estimating $s_\theta(t)$ from s(t), over some analysis window w(t), $0 \leq t < T$, is to minimise the integral $I(\tau)$ with respect to $s_\theta(t)$ where

$$I(\tau) = \int_0^T w(t) \left[s(t) - s_\theta(t)\right]^2 \, dt$$

and $$s_\theta(t) = s_\theta(t + k\tau) \qquad ; \text{for integer } k \text{ and } \ 0 \leq t < \tau$$

The noise component of the speech signal is then uncorrelated with the periodic component and hence minimising the integral above is equivalent to maximising the periodic energy estimate from the signal for any period τ. To find the fundamental frequency of the speech we now require to find that value of τ which minimises $I(\tau)$, with $s_\theta(t)$ replaced by its least-squares estimate. More details of this method and an efficient means of computing $I(\tau)$ for a discrete signal are given by Friedmann (1977).

We can express the above method approximately in terms of the signal autocorrelation function, $\Phi(\tau)$ as

$$I(\tau) \approx \Phi(0) - \frac{1}{K} \sum_{k=0}^{K-1} \Phi(k\tau)$$

It is now apparent that the measure is in effect exploiting the fact that for a periodic signal the autocorrelation function itself is periodic and also that the variance of the least-squares estimate will be a factor of K-1 times less than that of the autocorrelation estimator.

A frequency domain interpretation of the least-squares method can be derived using the above approximation by taking the Fourier transform of both sides of the equation. The resulting interpretation is that in using the least-squares method, rather than the standard autocorrelation function, we are replacing the cosine sampling of the power spectrum (or preprocessed power spectrum such as the log spectrum) with a much closer match to the optimum detector. In the least-squares analysis the 'sampling function' is $\sin[(2K-1)\pi\grave{w}/w] / \sin[\pi\grave{w}/w]$ As K increases this sampling function rapidly tends to a delta function and the least-squares method approaches an Harmonic Sum function similar to that given by Schroeder (1968). The Harmonic Sum function can be thought of as sampling the spectrum using a delta function as the harmonic detector.

DISCUSSION

In practice it has been found that there is little difference between the least-squares method and the Harmonic Sum method providing suitable window functions are used. However, whatever the noise level, sampling the spectrum with a series of delta functions produces superior results compared to sampling with a cosine function. This implies for instance, that the Harmonic Product spectrum (that is, an Harmonic Sum applied to the log spectrum) is superior to the Cepstrum. This result is perhaps not too surprising since in sampling the log spectrum in this way we are ignoring the unreliable information present in the troughs between harmonics.

REFERENCES

1. N. Weiner in 'Extrapolation, Interpolation and Smoothing of Stationary Time Series', John Wiley, New York, 1949.
2. D.H.Friedmann IEEE Trans., Vol.ASSP-25, No.3, pp 213-221, 1977.
3. M.R.Schroeder J. Acoust. Soc. Amer., Vol. 43, No.4, pp 829-834, 1968.