SIGNAL PROCESSING IN ACOUSTIC PHONETIC ANALYSIS OF SPEECH.

M.A. Jack, G. Duncan, A.M. Sutherland and J. Laver

Centre for Speech Technology Research, University of Edinburgh.

ABSTRACT - This paper describes improved formant estimation and pitch tracking signal processing algorithms. Formant estimation is based on linear predictive coding techniques (LPC) using off-axis spectral estimation coupled with a progressive increase in vocal tract model order. Pitch tracking is based on a multi-feature investigation of the time-domain speech waveform, optimised for accurate measurement of individual pitch periods.

IMPROVED FORMANT ESTIMATION.

Introduction.

LPC-based spectrum estimation (Markel and Gray, 1976) is well-suited to the characteristics of the speech signal, but performance is generally impaired both from inadequate modelling of low-intensity formants which are heavily damped and by an inability to separate formants which are closely-spaced in frequency. The technique described here employs off-axis spectral estimation coupled with progressive over-estimation of model order in any single analysis frame, This method of formant extraction, enables the detection of weak formant features, greatly improves the ability of the LPC estimator to resolve formants which may have merged together and offers improved noise immunity.

The pole enhancement technique.

In LPC-based speech analysis, the choice of model order has a critical effect on the performance of the spectral estimator, and usually the model order is chosen to be between $p=12$ and $p=16$, inclusive. With fewer coefficients, LPC analysis fails to separate formants which merge in certain speech sounds; too large a value of model order clutters the smooth spectrum with spurious peaks and is considered to degrade the signal-to-noise performance of the estimator. The new pole enhancement technique outlined here, Figure 1, avoids the problem of masking of weak vocal tract formants by more intense formant features by employing off-axis spectral estimation. For any given analysis frame, a series of off-axis spectra are calculated at decremented values of z-transform radius. Formant candidates are then extracted from each off-axis spectrum using peak detection and parabolic interpolation. The improvements afforded by the technique are achieved by the deliberate progressive over-estimation of the model order with each decrease in z-transform radius. In this way, the estimate of the vocal tract frequency response becomes more likely to exhibit events relating to low-intensity formants both by providing for weaker eigenvalues in the solution to the vocal tract transfer function (by increasing the model order) and by enhancing their effect on the frequency response (by decreasing the z-transform radius). The simultaneous use of off-axis spectral estimation, when coupled with progressive increase in model order produces a deterministic pattern of formant movement on the short-time stationary characteristics of the speech spectrum, namely the formants associated with the vocal tract, as the z-transform search path approaches the positions of transfer function poles. This implies that each formant value in the set of formant values which relate to any single formant feature will occupy slightly different frequency positions in each of the off-axis spectra. The final step in the formant estimation process therefore requires the use of an averaging filter using formant Q-factor as a weighting criterion. The objective of this final stage is to extract those spectral features which have remained consistent across several pole-enhanced spectra as model order has changed.

Performance.

An example of the application of the pole enhancement technique is illustrated in Figure 2, which compares formant extraction using a standard 16th-order LPC analysis (Figure 2(a)) against results obtained using the new pole enhancement technique (Figure 2(b)). The speech used was from a male speaker uttering the word "nana". The nasalised consonant /n/ presents a difficult challenge to LPC analysis since several formant features associated with nasality are characteristically of high bandwidth and low intensity, that is, their characteristic poles in the vocal tract transfer function are deeply-embedded inside the z-plane unit circle. Weak nasal formant events at approximately 0.9kHz (N2) and 2.4kHz (N4), which have remained undetected using standard LPC analysis (see Figure 2(a)), are plainly visible in the enhanced formant estimation (see Figure 2(b)), particularly during the second occurrence of the consonant /n/. It is also interesting to note that the pole enhancement technique has succeeded in detecting the low-frequency voice bar characteristically associated with low back vowels such as /a/, together with some apparent nasalisation of each of the occurrences of /a/ (note the pattern of formant movement around the second formant, F2). All of the above weak features detected by the pole enhancement technique are discernable in wideband spectrographic analysis of the speech segment. In terms of noise performance, Figure 3 shows the comparative performance of normal LPC analysis (using a 16th order model), and the pole enhancement technique in the presence of additive white noise. Here the percentage of total frames for each technique in which a value of second formant (F2) was correctly found has been plotted as a function of signal-to-noise ratio, showing the new technique to offer improved resistance to noise-induced errors.

IMPROVED PITCH TRACKING.

Introduction.

Accurate pitch determination is complicated by many factors including the non-stationary nature of the speech signal, and the large frequency range of a typical voice. A further problem is irregular glottal excitation, giving rise to perturbations in the pitch/time contour. This feature is however, a useful source of information regarding speaker characteristics and the condition of the speaker's vocal cords, and its accurate quantification has been proposed for such tasks as medical screening (Laver, Hiller and Mackenzie, 1984) and speaker recognition. Extreme accuracy and robustness of pitch measurement are required for this application. The improved time domain pitch tracker considered here is based on extraction of "anchor points" in the speech waveform, (Tucker and Bates 1978), such as waveform peaks. Various attributes, which include waveshape, peak energy, peak width and amplitude are estimated and used to determine which two peaks best delimit any single pitch period.

Algorithm description.

In order that the effects of the vocal tract resonances in the speech waveform be minimised, the first stage of the algorithm, Figure 4, involves adaptive centre clipping. This results in substantial data reduction and an improvement in the signal environment for waveform peak identification. For each peak identified, three features are extracted: the width of the peak at the centre clipping points, the energy of the peak and the shape factor (amplitude divided by square root of energy). The three features of each peak may be represented as a point in three-dimensional space. The origin of this space is defined to be the "current" peak, and the axes correspond to the three features. In this way, the "distance" between a given point and the origin is related to the similarity between the corresponding peak, and the "current" peak. The initial stage of peak selection is carried out as follows. All speech waveform peaks prior to the "current" peak (within a given time window) are plotted in the space. As each point is plotted, a cuboid is superimposed on the feature space and the point rejected if it falls outwith this volume. The swept cuboid volume, which is related to the probability of accepting a point, varies with the cube of the time (measured from the "current" peak). The cuboid volume reaches a maximum at time which corresponds to a precalculated estimate of pitch period, obtained from the output of an infinite impulse response filter acting upon previous period duration values. The final stage of peak selection involves the calculation of a (weighted) Euclidean distance between each of the remaining points and the origin. In order that the above operations are not carried out during unvoiced or silent sections of speech, an energy measure, combined with a zero crossing measure, is used to identify voiced intervals.

Performance.

The task of pitch detection algorithm evaluation demands a standard against which the output of the algorithm may be compared. For our purposes here, the new algorithm is compared with the well proven parallel processing method (Gold and Rabiner, 1969) using the output from a laryngograph as reference to measure the cycle-to-cycle pitch period tracking ability. Figure 5 compares the error performance of the two algorithms. The x-axis corresponds to the points of glottal closure within the word "meat" as obtained from a laryngographic recording and the errors in the two algorithms at these points are plotted on the y-axis. The new algorithm displays improved performance. Comparative results for the noise performance of the pitch trackers are shown in Figure 6, for the case of "crowd noise" or the "cocktail party effect", indicating the improved performance of the new technique. Similar improved performance has been noted for other noise sources including air conditioning, keyboard clicks, traffic noise and broadband Gaussian noise.

CONCLUSIONS.

A formant estimation algorithm has been described which offers improved accuracy. The technique has been demonstrated to offer improved performance over standard LPC analysis. Similarly an improved pitch tracker has been described which offers high accuracy in cycle-to-cycle pitch estimation. Both techniques have been evaluated under conditions of noise.

REFERENCES.

Gold, B. and Rabiner, L.R. (1969), "Parallel processing techniques for estimating pitch periods of speech in the time domain", J. Acoust. Soc. Amer., 1969, 46, pp442-448.

Laver, J., Hiller, S. and Mackenzie, J. (1984), "Acoustic analysis of vocal fold pathology", Proc. Inst. of Acoustics, 1984, 6, pp235-242.

Markel, J.D., and Gray, A.H. (1976), "Linear Prediction of Speech", Springer- Verlag, 1976.

Tucker, W.H. and Bates, R.H. (1977), "Efficient pitch estimation for speech and music", El. Lett. 1977, 13, pp357-358.

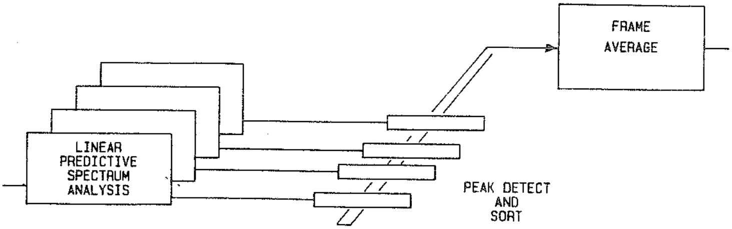

Figure 1. Scheme of improved formant estimator.

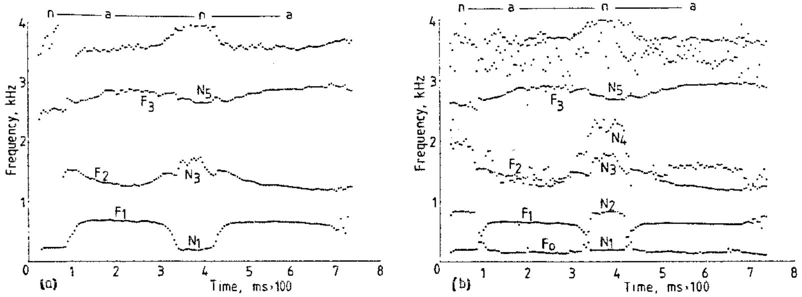

Figure 2. Formant estimation for utterance "nana".
(a) 16th order LPC. (b) pole enhancement method.

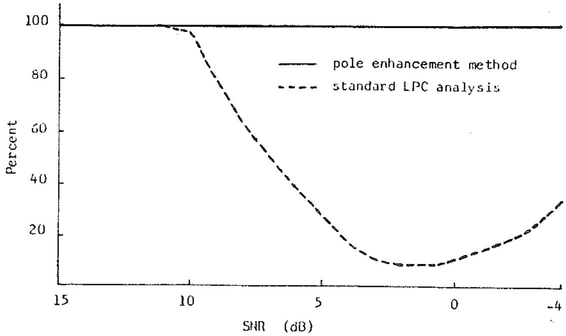

Figure 3. Noise performance of formant estimator.

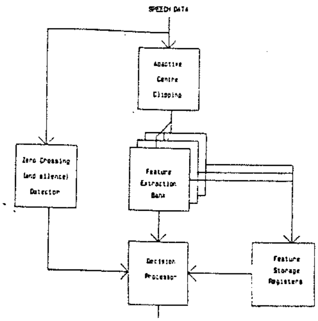

Figure 4. Scheme of improved pitch tracker.

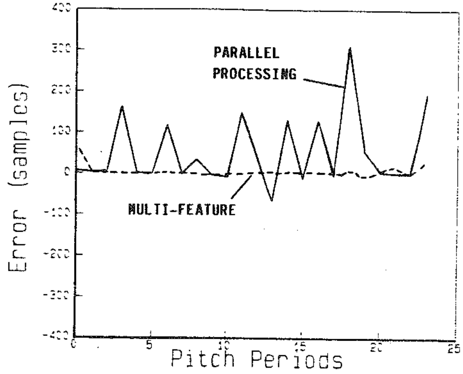

Figure 5. Accuracy measure of pitch tracker.

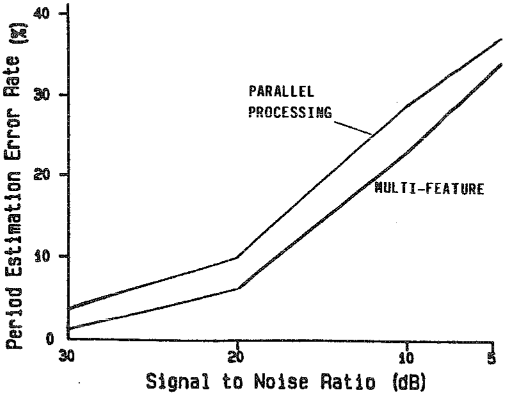

Figure 6. Noise performance of pitch tracker.