PERFORMANCE COMPARISON OF FIVE PITCH DETERMINATION ALGORITHMS ON THE LINEAR PREDICTION RESIDUAL OF SPEECH

E.H.S. Chilton*, B.G. Evans*

ABSTRACT

Accurate performance of pitch determination algorithms (PDAs) is essential to obtain good quality speech coding with linear prediction at low bit-rates. In this study, five pitch determination algorithms, representative of the range of short-term analysis methods, are applied to the residual from a linear prediction inverse filter. Four of these algorithms are well known (autocorrelation, amdf, cepstrum and maximum likelihood) while the fifth is a novel harmonic analysis technique applying the frequency domain autocorrelation to the power spectrum of the residual. Other comparative studies have generally used added gaussian noise to test robustness. In this study, both additive and multiplicative noise is combined with the speech at various levels and used to test the algorithms. Results indicate that multiplicative noise can have severe consequences on the accuracy of the pitch determination algorithms and that the novel harmonic analysis method performs well under adverse conditions.

INTRODUCTION

The work presented in this paper is being developed as part of a study of the design of good quality linear predictive coders (LPC) at bit-rates below 4kb/s for application to mobile telephony. The presence of pitch in voiced segments of speech provides one of the major sources of redundancy which can be exploited in low bit-rate coding. Significant improvement in coder performance can be achieved if pitch extraction is included (ref 1) and is essential at very low bit-rates. One of the major problems is the poor performance of the pitch estimation algorithms particularly when operating on degraded speech.

The speech input to a coder may have become degraded in one of two situations. Where the input is directly from a microphone, e.g. in a mobile telephone, then noise will be additive at the microphone. However, the situation could also arise where the input is a speech signal carried some distance over an analogue telephone line before entering the digital network. In this situation, noise corrupting the speech could be either additive or multiplicative.

PDAs can be broadly divided into two classes, time-domain and short-term analysis methods. The time-domain methods measure the pitch pulse separation on a pulse-by-pulse basis while the short-term methods perform some transformation on short segments of speech to provide an average pitch value over the block (ref 2). The latter have generally, but not exclusively, been used in speech coding since they provide a measure of the pitch redundancy present. It is this latter type of algorithm that was considered in this work. All the PDAs under test were applied to the residual of the speech signal at the output of the linear prediction inverse filter.

Five algorithms were tested, representative of the range of short-time analysis PDAs. Four of these are well-known techniques (autocorrelation, amdf, cepstrum and maximum-likelihood) while the fifth is a novel method which examines the harmonic structure of speech segments by applying an autocorrelation function to the power spectral density function. The application of the autocorrelation and amdf methods to the LPC residual has appeared previously in papers by Kwon & Goldberg (ref 3) and by Un & Yang (ref 4). The cepstrum method developed by Noll (ref 5) is not normally applied in this way. The maximum-likelihood method, first proposed by Noll (ref 6) and later by Wise et al. (ref 7) was developed by Friedman (ref 8) and applied to the LPC residual in a multi-dimensional form (ref 9).

Generally most comparative tests of PDA performance have been limited to either clean speech or speech with added noise (ref 10–12). Our own experience indicates that multiplicative noise even at relatively low levels, can have a severe effect on PDA performance.

*Dept. of Electronic and Electrical Engineering, University of Surrey, Guildford, Surrey, GU2 5XH
OVERVIEW OF METHOD

The process of pitch estimation can be divided into three parts: pre-processor, pitch estimation algorithm and post-processor (ref 2). The core of the process is the algorithm which performs some transform on the speech to provide an output sequence containing an enhanced pitch value. The preprocessor usually serves to flatten the speech spectrum before applying the algorithm. The postprocessor computes an estimate of the pitch value from the algorithm output.

In this study the speech was initially preemphasised before being passed through a 10-pole inverse linear prediction filter. The filter coefficients were obtained using the autocorrelation method. The residual obtained had most of the formant structure removed, leaving a good approximation to the excitation function of the speech. This residual was applied to each of the algorithms.

For each algorithm, the postprocessor was different but in order to make a fair performance comparison the complexity of each postprocessor was made as nearly equal as possible, usually a simple peak picking operation. The performance of any one algorithm could be enhanced by increasing the postprocessor complexity but this would tend to obscure the performance of the algorithm under test. Two further restrictions were placed on the postprocessor design. No arbitrary thresholds were applied to aid discrimination of peaks since these would tend to be speaker dependent and thus reduce the general applicability of the algorithms. No use was made of pitch smoothing nor of prior pitch information in making a decision.

The algorithms were tested on two four second sentences, male and female, and then on the same sentences first corrupted by additive noise and, then (in separate tests) on multiplicative noise set at various predefined levels. The noise source was synthesized from a pseudo-random sequence with approximately gaussian characteristics. The long term average signal power in the clean speech was measured and the level of the noise signal adjusted to give a measured long-term signal/noise ratio. In the case of the additive noise, noise levels were increased from -34dB to +22dB in 6dB steps.

For the multiplicative noise, the following model was used:

\[ T_n = (1+N_n)S_n \] (1)

where \( N_n \) is a dimensionless random noise sequence and \( S_n \) is the speech sequence.

The long-term signal noise ratio is defined as:

\[ \frac{S}{N} = 10 \log \frac{1}{\sigma_{N}^2} \text{dB} \] (2)

where \( \sigma_{N}^2 \) is the variance of the noise sequence \( N_n \).

We note that in this case the signal-noise ratio is independent of the signal level. The signal-noise ratio was reduced by increasing the noise sequence by a scaling factor. The corresponding values in decibels were 21, 17.3, 14.8, 11.3 and 8.8.

ALGORITHM DESCRIPTION

Each algorithm was applied to the residual of the speech files as described in the previous section. A standard window size of 256 samples, shifted by 160 samples was used in each case to provide pitch values every 20mS at 8kHz sampling.

Standard implementations of the autocorrelation, amdf, cepstrum and mlhd methods were used as described elsewhere in the literature (refs 3, 4, 5 & 8). In each case the most likely pitch value lying between 10 and 160 samples was found by applying a suitable constant bias to the time-domain output sequence of each transformation and then picking the most prominent peak (maximum or minimum as appropriate).

The harmonic analysis method requires further explanation. Harmonic analysis methods calculate the separation of the harmonics present in the spectra of voiced speech rather than the value of the fundamental pitch. A well defined harmonic structure is required for these methods to work satisfactorily.
In this case, an FFT is applied to the speech and the linear power spectral density function is calculated from:

\[ P_k = X_k \cdot X_k^* \]  

(3)

where \( X_k \) is the complex FFT of the time-domain sequence and \( X_k^* \) is its complex conjugate.

A frequency-domain autocorrelation is then applied to \( P_k \) in the form of:

\[ R_\kappa = \frac{1}{N} \sum_{k=0}^{N-\kappa} P_k \cdot P_{k+\kappa} \]  

(4)

where \( \kappa \) represents the frequency domain shift.

A highly smooth function is obtained in which most spectral features are suppressed except for the dominant harmonic structure. The appropriate value can be readily identified by a simple peak picking operation as with the other algorithms. Random variations in the spectrum appear as smooth variations in the amplitude of the frequency domain autocorrelation. Under noisy conditions, these variations could mask important peaks. Significant improvements were obtained by using the second order differential of the function and discarding negative values. The exponential variations were reduced while the periodic portion of the function, representing the harmonic spacing, remained prominent.

RESULTS

The outputs of each algorithm were post-processed and the most likely pitch values were found using appropriate peak picking routines. A reference file of pitch values was compiled for each sentence from time domain plots of the residuals. For each speech sequence, the measured pitch values were compared to the reference set and errors recorded. The criteria for an accurate match was set for agreement within 10% of the reference value or if pitch halving or doubling occurred, otherwise an error was recorded. The graphs of figure 1 indicate the measured performance of the five algorithms for male and female speech with additive and multiplicative noise by plotting percentage gross error as a function of signal-noise ratio.

![Fig.1a Male speech with additive noise.](image1a)

![Fig.1b Male speech with multiplicative noise.](image1b)
CONCLUSION

In both male and female speech, the errors introduced by multiplicative noise are significantly greater than for the corresponding level of additive noise. The relative performance of the algorithms differed markedly, with the cepstrum method performing least well in both male and female speech. For the male speech the autocorrelation and the spectral autocorrelation methods both performed well, with the spectral autocorrelation showing significant advantage in multiplicative noise. For the female speech, the amdf function and the autocorrelation both performed better than the others. However, overall performances were poorer than for the male speech. The spectral autocorrelation showed slight advantage at low noise levels but deteriorated rapidly as the signal/noise ratio decreased. It was observed that the harmonic structure in the spectra of the female speech was much poorer than for the male, particularly at higher frequencies. The relative performance of the amdf function appeared to improve under multiplicative noise. Because of the small amount of speech used in this study, it is unclear whether the differences noted between the male and female speech are general or due to the individual characteristics of those recordings. Further work is needed using a larger speech base.

REFERENCES

1. T.Araseki et al., 7th Int. Con. on Dig. Sat. Comms., p 785, Munich, May 1986.