Proceedings of The Institute of Acoustics

A PITCH DETECTION ALGORITHM OPTIMISED FOR MICROPERTURBATION ANALYSIS

A.M. Sutherland, M.A. Jack, J. Laver

Centre for Speech Technology Research, University of Edinburgh.

INTRODUCTION

Much attention has been directed to the problem of accurate measurement of the pitch, or fundamental frequency, of the voice [1]. The solution to this problem, however, is complicated by many factors. These include 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 the condition of the speaker's vocal cords, and its accurate quantification has been proposed for such tasks as medical screening [2] and speaker recognition. The extreme accuracy and robustness of pitch measurement required make this task one of the most demanding possible for a pitch detection algorithm.

Irregularities in the glottal excitation function are difficult to identify for several reasons. Firstly, some irregularities may occur at voicing onset or offset, points at which many pitch detection algorithms are at their least accurate. Further, many existing pitch detection algorithms depend upon regularity of the pitch pattern to such an extent they may fail when presented with an irregular signal - this is particularly true in the case of spectrally based algorithms. Most importantly however, is the difficulty in distinguishing between a genuine irregularity in pitch period duration, and a simple pitch detection error. The design of any pitch detection algorithm capable of identifying irregularities must take these problems into account.

The first requirement of such an algorithm is that it supply a localised estimate of pitch period. In this way, any cycle-to-cycle variations in pitch period duration may be identified. It is preferable that the pitch estimates be in synchrony with the pitch periods, otherwise omission or over-reporting of individual pitch periods may take place. The above criteria rule out the use of frame-oriented short-term analysis methods of pitch detection (such as correlation or spectral transformation). Such techniques produce a more global estimate of pitch period duration and thus effectively smooth out and correct any irregularities. The problem of distinguishing between any genuine irregularities and simple pitch detection errors must be addressed from two angles. Firstly, the intrinsic workings of the algorithm must be designed in such a way that pitch detection errors are highly unlikely. Secondly, the designer must be allowed operational parameter flexibility during the algorithm optimisation phase. Such flexibility is afforded most by time domain algorithms, particularly those employing temporal structure investigation of the speech waveform [1]. Finally, the algorithm must be capable of fast operation and be easily realisable in both hardware and software, once again implying the use of a time-domain algorithm based upon temporal structure investigation.

Several pitch detection algorithms based upon temporal structure investigation of the speech waveform have been reported. The most popular of these is the parallel processing method [3]. Although this algorithm is capable of robust operation, its design includes a measure of smoothing of the output pitch period values. As has been previously noted [4], it is thus unable to accurately measure irregularities in the vibration of the vocal cords. In a macro fashion however, it remains a reasonable standard against which to compare new methods of pitch detection. Other temporal structure investigation algorithms include the complex data-reduction method [5] and a simpler, but error prone, multi-feature algorithm [6].

This paper presents an improved pitch detection algorithm based on the multi-feature investigation of the speech waveform [6]. It is optimised for the accurate measurement of irregularities in the duration of pitch periods i.e. micro-perturbation analysis. The improved algorithm is evaluated using both subjective and objective measures. Finally, the performance of the algorithm is evaluated in the context of automatic speaker verification.

ALGORITHM DESCRIPTION

The improved algorithm employs feature extraction of the speech waveform peaks to accomplish accurate pitch detection. The basic philosophy is common to all algorithms of this type. Firstly, a given event in the speech waveform (in this case a peak) is identified and examined. Secondly, a selection process is carried out whereby only those events likely to represent a pitch period delimiter are retained. Finally, the time periods between such delimiters are calculated and presented as pitch period durations.

In order that the effects of the vocal tract resonances in the speech waveform be minimised, the first stage of the improved algorithm is adaptive centre clipping (Figure 1). This results in substantial data reduction, and an improved environment for waveform peak identification. For each peak identified, three features are extracted: the width of the peak at the centre clipping limits, the energy of the peak and the shape factor (amplitude divided by the square-root of the energy). It is the task of the decision processor to match a 'current' waveform peak to the corresponding peak in a previous pitch period, and so successfully delimit a pitch period. This is completed in two stages.

In both the selection stages, 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 (Figure 2). The cuboid volume, which is related to the probability of accepting a point, varies with the cube of the time (measured from the 'current' peak) according to (1).

$$V(t) = \begin{cases} R_1R_2R_3(t - b)^3 & b < t < Tex \\ R_1R_2R_3(e - t)^3 & Tex < t < e \end{cases} \qquad (1)$$

The cuboid volume reaches a maximum at time Tex. This value is a precalculated estimate of pitch period, obtained from the output of an infinite impulse response

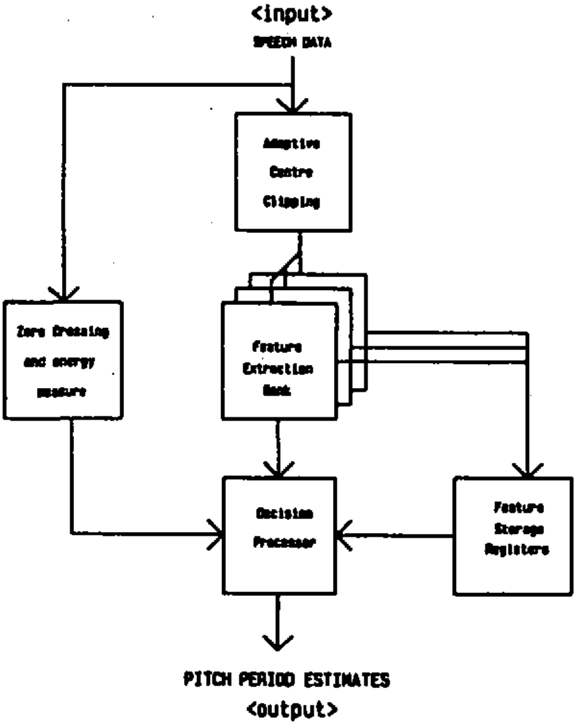

Figure 1

filter acting upon previous period duration values. The parameters 's' and 'e' refer to the start and end of the time window. It has been found that best results are obtained when these values are proportional to Tex. Finally, the parameters R_1,R_2 & R_3 refer to the allowable rate of change (with respect to time from the current peak) of each of the features. They thus dictate both the maximum value and rate of change of the cuboid volume. This selection stage represents the only use of time-dependent information in the peak selection process. It is thus essential that the effect of this action be minimal, otherwise the ability of the algorithm to accept irregular pitch period durations will be compromised. The final stage of peak selection involves the calculation of a (weighted) Euclidean distance between each of the remaining points and the origin. The closest point, corresponding to the previous peak which exhibits greatest similarity to the 'current' peak, is selected to represent the start of the pitch period. (This makes the reasonable assumption of a monotonic relationship between the similarity of two peaks and the probability of them being a matching pair of period delimiters.) The 'current' peak is defined to be the end of the pitch period.

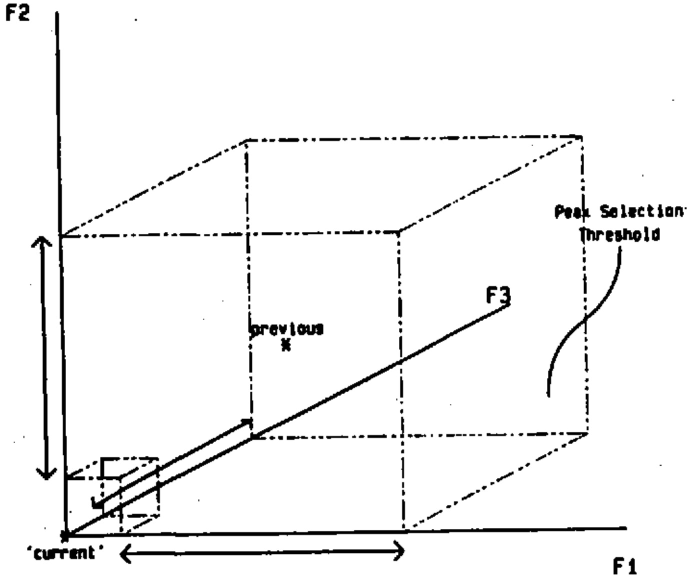

Figure 2

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.

ALGORITHM EVALUATION

The task of pitch detection algorithm evaluation demands a standard against which the output of the algorithm may be compared. Standards which have been considered useful range from the visual examination of the speech waveform and the use of other (known) pitch detection algorithms, to the direct measurement of the glottal movements by means of a laryngograph. Subsequent comparisons may be either of an objective or subjective nature. In this paper, several approaches have been adopted. Firstly, a subjective assessment of the macro (sentence) behaviour of the algorithm is made, relative to the well proven parallel processing method. The laryngograph is also used to measure the cycle-to-cycle pitch period tracking ability of the improved algorithm. The stability of the algorithm when presented with input data which varies in time origin is then assessed, and finally the application-specific measure of 'equal error rate' in an automatic speaker verification system is calculated.

The pitch against time contours for a given sentence (a) formed by both the improved algorithm (b) and the parallel processing method (c) are shown in Figure 3. The sentence, "More than twenty ships were tied up at the quay.", was uttered by a 19 year old male. Visual comparison of the two pitch contours shows a high level of agreement, although the parallel processing method (c) exhibits occasional period doubling which, it was determined from close waveform examination, is erroneous. The mean fundamental frequency of this utterance calculated by the improved algorithm (128.9Hz), compares well with both the visual calculation (132Hz) and the parallel processing method (127.1Hz).

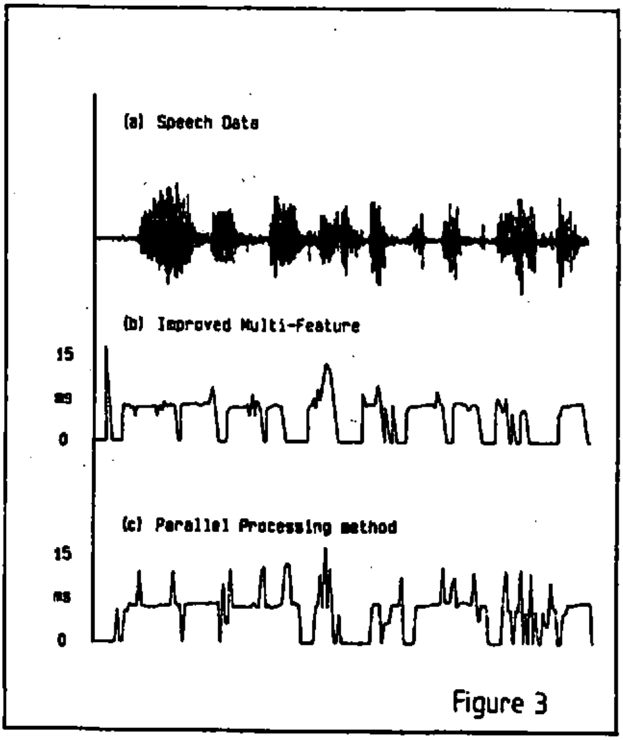

Figure 3

An improved standard for pitch measurement accuracy is the laryngograph [7]. In order to measure the cycle-to-cycle (micro) accuracy of the improved algorithm, the points of glottal closure within the word "meat" were obtained from a laryngographic recording. These instants were plotted on a time axis (Figure 4a). The positions of the speech waveform peaks selected by both the improved algorithm (b) and the parallel processing method (c) to represent the beginnings of pitch periods were plotted in a similar manner. The peaks of the cross correlation functions (r) between each of the pitch detection algorithm outputs and that of the laryngograph are shown in figure 4. The higher value of correlation achieved by the improved algorithm confirms the visual assumption of its accurate cycle-to-cycle pitch period measurement.

The third method of pitch detection algorithm comparison used, that of a limited measure of the robustness, is presented here. A sentence, of duration 3s, was

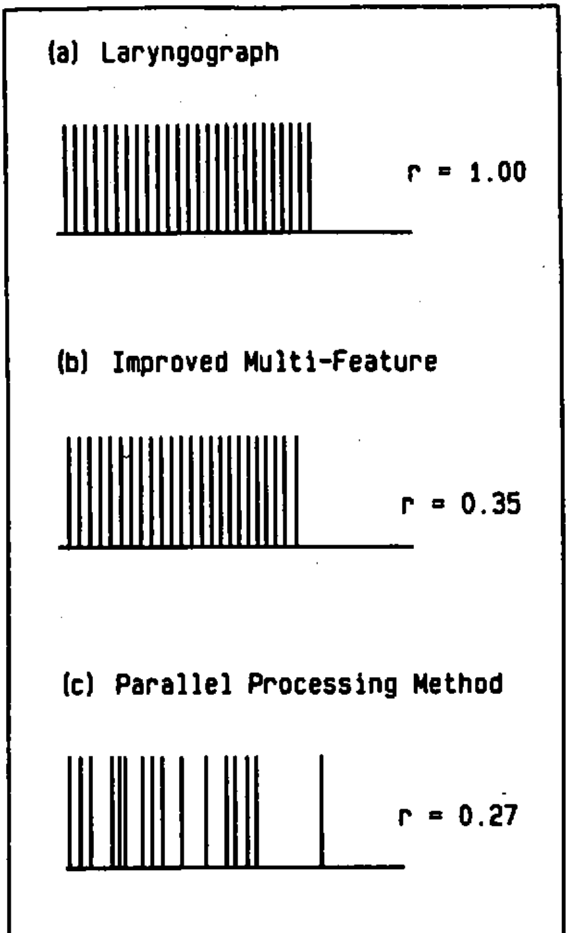

Figure 4

passed through each pitch detection algorithm five times. On each occasion, the algorithm began 1ms further into the data than the previous. It was ensured that the speech was delimited by silent intervals, and thus no pitch periods were lost or gained in this process. The maximum and minimum mean fundamental frequencies obtained from each of the algorithms are shown in Table 1. It can be seen that the improved algorithm exhibits complete stability under such conditions.

Table 1. Changes in mean fundamental frequency (Hz) with input offset.

Algorithm	Minimum	Maximum	Spread
Improved Multi Feature	100.83	100.83	0.00
Parallel Processing	99.22	101.03	1.81

The real measure of the quality of an algorithm is how well it performs in a 'real world' application. The final measure used to evaluate the performance of the improved pitch detection algorithm is thus application-specific. It involves the calculation of the 'equal error rate' when the algorithm is used to form the input of an automatic speaker verification (ASV) system. A database (all male) of 5 'clients' and 24 'imposters' was formed. A vocal profile for each client, consisting of 9 micro-perturbation measures, was formed from over 20 utterances, and the dissimilarities between it and each of the clients individual utterance profiles calculated. The dissimilarities between each of the imposter utterance profiles and the client profiles were also calculated. The most common way of describing the performance of an ASV system is the 'equal error rate'. This relates to the error rate when the threshold of dissimilarity is set (a-posteriori) such that the probability of falsely accepting an imposter is equal to the probability of falsely rejecting a client. It was found that this parameter was reduced by 25% of its original value through the use of the improved algorithm when compared with the parallel processing method.

CONCLUSIONS

An improved multi-feature pitch detection algorithm for use in micro-perturbation analysis has been described. This algorithm offers high speed operation and its simple construction lends well to hardware implementation. Its high levels of accuracy, evaluated against both an existing algorithm and the laryngograph standard, show the potential of this type of algorithm for the task of micro-perturbation analysis.

REFERENCES

[1] W. Hess, Pitch Determination of Speech Signals, algorithms and devices, Springer-Verlag, (1983).

[2] J. Laver, S. Hiller and J. Mackenzie, 'Acoustic Analysis of Vocal Fold Pathology'. Proceedings of the Institute of Acoustics, 6, 235-452, (1984).

[3] B. Gold and L.R. Rabiner, 'Parallel processing techniques for estimating pitch periods of speech in the time domain'. J.A.S.A, Vol 46, 442-448, (1969).

[4] D.M. Howard and A.J. Fourcin, 'Instantaneous voice period measurement for cochlear stimulation'. Electonics Letters, Vol 19, 776-777, (1983).

[5] N.J. Miller, 'Pitch Detection by Data Reduction'. IEEE Trans. A.S.S.P., Vol 23, 72-79, (1975).

[6] W.H. Tucker, R.H.T. Bates, 'A Pitch Estimation Algorithm for Speech and Music'. IEEE Trans. A.S.S.P., Vol 26, 597-604, (1978).

[7] D.M. Howard, J.A. Maidment, D.A.J. Smith, I.S. Howard, 'Towards a comprehensive quantative assessment of the operation of real-time fundamental frequency extractors', IEE Conf. Publ. 258, 172-177, (1986).