

SPEECH FUNDAMENTAL PERIOD ESTIMATION BY A NOVEL PARALLEL PROCESSING METHOD

Ian Howard *

ABSTRACT

The algorithm described here is a time domain speech fundamental period estimator. Its operation involves filtering the speech with a filterbank. The filterbank outputs are then represented in terms of positive-going zero-crossing locations and their envelopes. Finally the period epoch marker locations are "optimally" selected. The algorithm is compared with a reference based on the output of a laryngograph.

INTRODUCTION

Voiced speech is characterised by the vibration of the vocal folds. In sustained periods of voicing, this gives rise to a quasi-periodic signal with essentially a harmonic spectrum. Fundamental frequency estimators generally operate either in the time domain and/or in the frequency domain, and often take advantage of these respective signal attributes (See [ref.1] for a review).

Fundamental frequency estimators that operate in the frequency domain give rise to an output in the form of frequency values and such algorithms typically have smoothing in their operation. Those which operate in the time domain give rise to an output specified in terms of the locations in time of period epoch markers (T_x). The estimator described here outputs T_x . Period epoch marker determination has to be, by its very nature, based on very short-time changes in the speech signal. As a result, irregularities in vocal fold vibration are easier to preserve with such a scheme, and their retention can be desirable [ref.2].

PREPROCESSOR

One may consider the problem of detecting the occurrence of a period epoch in terms of looking for features that signify evidence of such an event. A set of features that are useful for such a task are inherent in a filterbank. Clearly the time-domain constraint on the filters are to ensure that the location of a period epoch is not smeared-out in time, thus making its location too imprecise. The frequency domain requirements of the filters are such that one particular channel must respond to only a narrow range of periodic disturbances, that is, they require adequate frequency selectivity. Clearly, for linear filters, the time and frequency domain requirements are contradictory, and one must reach a compromise. Finally, one must have enough filters to cover the range of possible period epochs.

FILTERS

The filter bank specifications were arrived at by the following reasoning. To ensure operation for a wide range of speech fundamental frequencies, it is required that harmonic rejection of the filters be independent of fundamental frequency. Also, to permit later development of the selection algorithm, it is required that there will always be several resolved harmonics, again, invariant of fundamental frequency. This is consistent with constant Q for all the filters. To ensure that there are no neglected period values, it is required that at least the -3dB points intersect. This is the minimum acceptable condition - clearly there would be advantages for having more filters than this. Together with the constant Q specification, this leads to a logarithmic scale for filter centre frequencies.

*Dept. of Phonetics and Linguistics, University College London, UK.

To cover the range of normal speech the filter centre frequencies should be between about 50Hz - 5KHz. The result is a filterbank with 45 channels.

One clearly requires as much rejection of adjacent harmonics as possible. A figure of 60dB rejection in the stop-band was found acceptable and it is about the same as the quantization noise SNR. FIR filters were chosen to permit easy delay correction. They were designed using the Remez exchange algorithm, as available in ILS. The implementation of the filters involved a decimation / interpolation procedure to avoid the use of filters with many coefficients at low centre frequencies.

Due to the narrow bandwidths of the filters, their outputs will be, for the lowest few harmonics, essentially sinusoids with slowly varying envelopes. These outputs may be described in terms of their positive going zero crossings and their corresponding envelope values. Such a representation will still retain the required information and achieve substantial data rate reduction.

THRESHOLD TRAINING

A training phase was carried out to determine the thresholds to permit the algorithm make a voiced/voiceless (v+/v-) decision on the basis of channel envelope. The thresholds were estimated by training on speech labelled for voicing. They were calculated individually for each channel. The values chosen correspond approximately to the values of the respective channels's envelopes at the v+/v- transitions.

SELECTION OF PERIOD EPOCH MARKERS

In the simple selection scheme considered here, the assumption is that the optimum Tx output corresponds to the positive-going zero-crossings due to the output of one channel. That channel is selected on the basis of having an envelope that is the first amplitude maxima in the low frequency end of all channel envelopes, subject to certain constraints. The problems to solve are the dynamic selection of the "optimum" channel to use, and how to take account of relative delays when switching between different channels. The selection of the optimum channel is made over a window of length 20 ms. Since the rate of change of fundamental frequency and the formants is relatively slow, a condition imposed is that the choice of the next optimum channel must be either the same as the last, or an adjacent channel. This is to ensure that it is possible to achieve continuity at channel change over, and is only demanded if the channel amplitude is greater than a pre-determined threshold. This threshold determines whether or not the input is taken as voiced or voiceless. This permits discontinuities to occur at the end of voiced segments.

The problem is formulated as an optimization procedure in which it is required to minimize a penalty function. The penalty depends upon channel envelope and continuity in the selection of the optimum channels. The algorithm used to solve the optimization is based on dynamic programming.

To ensure alignment of the Tx of different channels at a cross-over, the delay between the channels is constantly estimated over the last window between a channel and its neighbours. It is then appropriately added or subtracted to give rise to an output Tx value that is consistent with the previous ones, without discontinuities.

The only smoothing in the output is that due to the transient response of the linear filters.

The algorithms were written in C and ran under the Unix operating system on a Masscomp MC5500 computer.

DATA

The data used in the trial run of the algorithm was the passage "These are steps towards some pattern that's ideal". The speaker was an adult male and the speech was recorded anechoically simultaneously with the output signal from a Laryngograph (Lx) [ref.3].

INITIAL RESULTS

The performance of the new algorithm was compared with a Laryngograph based analyser. Figure 1 shows a plot of fundamental frequency derived from the period epoch marker locations for the multi-channel fundamental frequency estimator and also for a reference that makes use of a Lx signal [ref.4].

Figure 2 shows the speech pressure waveform and Lx signal from a small portion of the utterance, together with the period epoch marker estimates, displayed as pulses, for the multi-channel and Lx based estimators. It can be seen that the main difference in the Fx contours is due to the fact that they are defined for longer with the multi-channel estimator. This is due to the fact that the Lx waveform does not always give a good indication of voicing when observation of the speech waveform clearly indicates that it is indeed present. The general shape and values of the two curves can be seen to show close similarity.

FUTURE WORK

Presently no use of harmonic relationships is used in the selection of the output Tx. A scheme that makes use of the harmonic relationships between channels in the optimization procedure for best channel selection is currently being implemented. At the higher frequencies the uncertainty in the zero-crossings becomes significant. This may be important when the higher harmonics are investigated. Greater precision could be achieved by incorporating interpolation before the zero-crossing detector. Related work on a period sieve based system is reported in [ref.5]. Work is continuing to assess the performance of the algorithm in a more rigorous manner [ref.6] in particular in the presence of noise. In addition an IIR filterbank is also being implemented, which should reduce computational requirements.

ACKNOWLEDGEMENTS

This work was supported by Alvey grant MMI/056 and MRC studentship RS-85-2.

REFERENCE

- [1] W Hess, Pitch determination of speech signals, Springer-Verlag, Berlin, (1983).
- [2] E Abberton, A J Fourcin, S Rosen, J R Walliker, D M Howard, B C J Moore, E E Douek, and S Frampton, In R A Schindler and M M Mertenich (eds), Cochlear Implants, New York: Raven Press, 527-537, (1985).
- [3] A J Fourcin, and E Abberton, E.R.M., Med. and Biol. Illust. 21, 172-182 (1971).
- [4] W Hess, and H Indefry, Proc. ICASSP-84, 1-4, (1984).
- [5] D M Howard, A Faulkner, and I S Howard, Pro. Eur. Conf. Sp. Tech., (these proceedings), (1987).
- [6] I S Howard, and D M Howard, Proc I.O.A., Vol 8, 323-330, (1986).

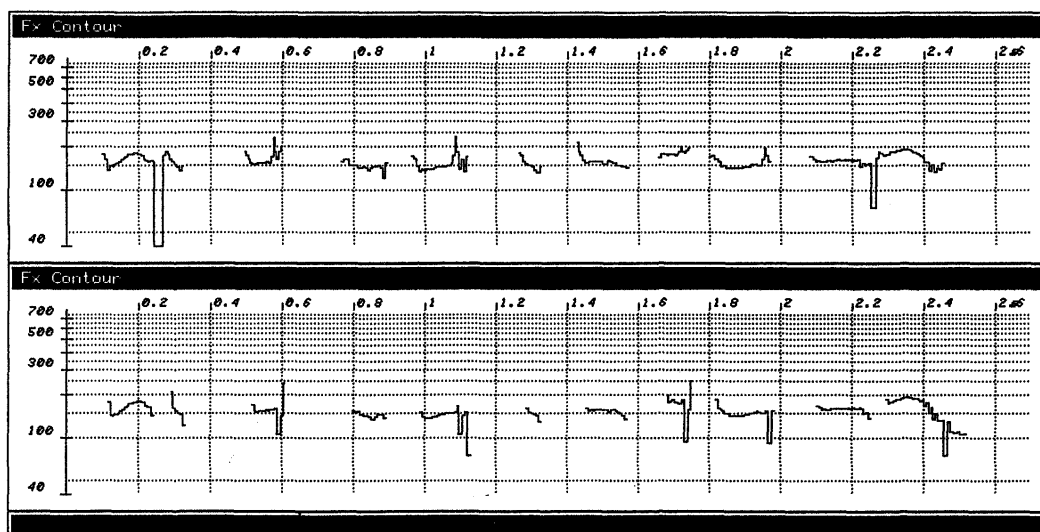


FIGURE 1. Top graph shows the Fx contour derived from the output of the multi-channel fundamental period extractor. Lower graph shows the Fx contour derived from the Lx waveform.

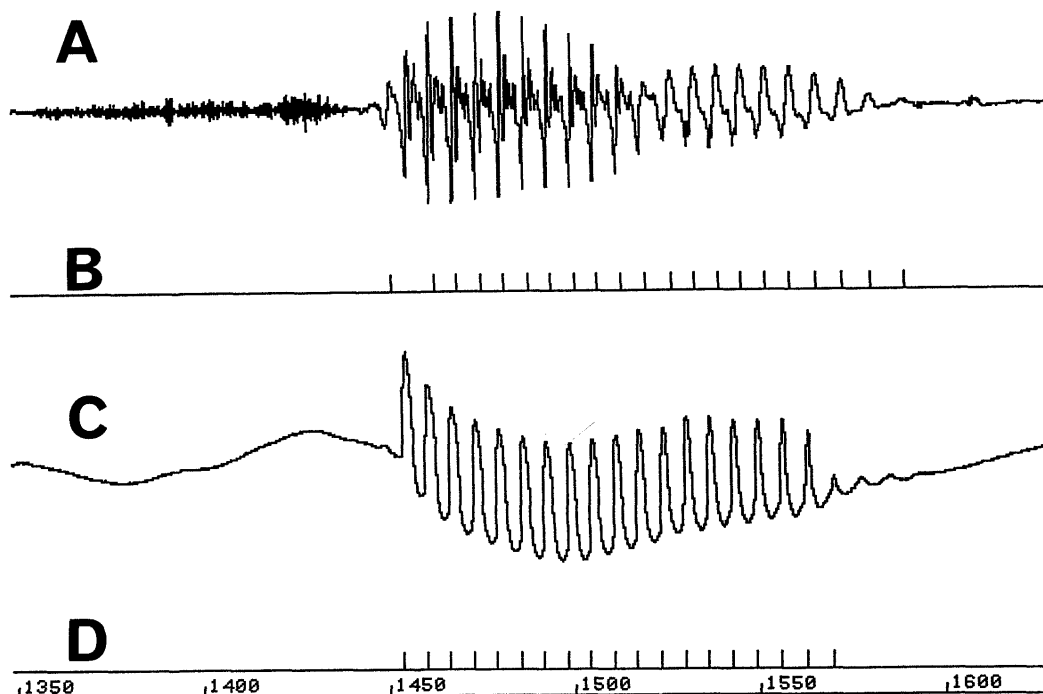


FIGURE 2. Trace A shows the speech pressure waveform for a smaller section of the passage. Trace B shows Tx markers derived by the multi-channel extractor. Trace C shows the corresponding Lx waveform. Trace D shows Tx derived from the Lx waveform.