

SPEECH FUNDAMENTAL FREQUENCY ESTIMATION BY MULTI-CHANNEL PEAK-PICKING

David M Howard * Andrew Faulkner * and Ian S Howard *

ABSTRACT

A variety of methods have been implemented for speech fundamental frequency (F_x) estimation from an acoustic input, but none can reliably estimate F_x in typical acoustic environments. The purpose of this paper is to describe developments, based on current theories of human pitch perception, to the design of a device which is used in speech processing hearing aids, speech training aids and speech research.

INTRODUCTION

The estimation of the fundamental frequency (F_x) of speech has been carried out by a variety of methods, which if suitably engineered can be tailored to meet the requirements of the intended application (ref 1). At UCL work is in progress in the EPI group (ref 2) to develop speech processing hearing aids for the totally and profoundly deaf. The presentation of voice pitch information to supplement lipreading forms the backbone of our technique, a specially developed peak-picking F_x estimation device (ref 3) is used.

The motivation behind the present work is twofold and is gained from contemporary theories of human pitch perception which are based on multi-channel temporal processing (e.g. ref 4). Firstly a multi-channel approach to F_x estimation should lead to enhanced performance of EPI prostheses in the presence of competing acoustic inputs and/or reverberation. Secondly there is a need for clear illustrations to support the descriptions of pitch perception used in teaching.

IMPLEMENTATION

The system has been developed in 'C' on a Masscomp 5500. It is designed to interface with the 'SPAR speech filing system' (ref 5). The multi-channel implementation consists of a number of stages, the output waveforms from each being stored with the original digitised input data for subsequent verification.

The first stage consists of a bank of band-pass filters. These cover the range 50-5000Hz, and each has a quality factor of 10. Their adjacent 3dB points intersect, resulting in a 45 channel filterbank. The companion paper (ref 6) gives a detailed description.

A software peak-picking algorithm acts on each channel of filtered speech. In the lower frequency bands the waveforms will be essentially sinusoidal, whilst in the higher frequency bands the output waveforms will be 'beat-like', where multiple harmonics interact. The basic algorithm (ref 7) has been modified to enable the system to establish appropriate period epochs for each channel. The resulting pulses are treated to a blanking interval such that if another pulse occurs within some preset

* Department of Phonetics and Linguistics, University College London, U.K.

time after any given pulse, then the first is ignored. For a periodic input the output period markers (Tx) will be related to the period of the resolved harmonics in the lower channels, and directly to the fundamental period in the higher channels.

The final two processing stages take the Tx items from each channel and process them to give an estimate of the fundamental frequency contour (Fx). Firstly each Tx is time windowed in 10ms frames and for each a mean Tx value is found. If this value falls outside the allowed range for Tx, 1.66ms to 25ms, then it is discounted. Thus a two-dimensional array (filter channel against window) of Tx values is built up. The second stage "period sieves" these Tx values by relating potential fundamental period values and their sub-multiples to the actual values found, to assess which Tx value gives the "best fit" for that window. This is then converted to a final Fx value for each frame.

PROGRESS AND RESULTS SO FAR

This system, in common with most Fx estimation techniques, has various parameters which will require optimisation as further experience is gained. Thus the results presented show progress to date, and do not make a definitive statement as to the capabilities of such a system.

The output from Fx estimation devices can be referenced (ref 8) against the laryngograph (ref 9). The laryngograph provides a 'standard' (ref 10) against which acoustically based devices can be referenced. Figure 1 shows the speech pressure (Sp), the Fx contour (logarithmic scale) derived from it after processing by the multi-channel system, the laryngograph output waveform (Lx), and the Fx contour (logarithmic scale) derived from it. These waveforms are for the vowel "a" spoken with a falling intonation by an adult female.

It can be seen that the detailed overall shape of the Fx contours is similar, and the falling intonation pattern is clearly shown in both contours. The most notable differences, in this example, are at the points of voicing onset and offset, where the output from the multi-channel system switches from voiceless to voiced earlier than the output from Lx, and the output from Lx switches from voiced to voiceless earlier than the output from this system. The reason for this can be seen by comparing Sp with Lx, especially at the offset of voicing, where the amplitude of the closure peaks in Lx drop suddenly to zero, whilst the last few in Sp reduce more gently in amplitude.

CONCLUSIONS

The initial stages in the implementation of a multi-channel peak-picking fundamental frequency estimation system have been described. It is based upon ideas which have been developed to explain human pitch perception, and consists of a filter-bank analysis followed by peak-picking, windowing, mean period detection and a cross-channel period-sieve analysis to give a final windowed fundamental frequency estimate. At present no output smoothing is employed.

The initial results with the system are promising and confirmation by further comparisons with the laryngograph and other acoustically based devices is planned. Comparison work with the laryngograph is being extended towards the establishment of semi-automatic optimisation procedures for acoustically based Fx estimation devices (ref 11) and at present the peak-picking device is under investigation. It should be possible to extend this to sections of this multi-channel system which have much in common with the peak-picking device.

Initial results already give a basis for useful teaching illustrations of current theories of human pitch perception. Eventually it is hoped that they will lead to the improved operation of the Fx estimation device used in EPI group hearing prostheses, especially in the presence of competing acoustic inputs and local environmental effects.

ACKNOWLEDGEMENTS

The authors would like to thank SPAR and EPI group colleagues for their help and support. This work is supported by the SPAR SERC-Alvey grant MMI/056 and MRC grants PG-8220657 and 209050.

REFERENCES

1. W Hess, Pitch determination of speech signals, (Springer-Verlag, Berlin, 1983).
2. A J Fourcin, E E Douek, B C J Moore, S M Rosen, J R Walliker, D M Howard, E R M Abberton, and S Frampton. An. New York Acad. Sci., 405, 280-294, (1983).
3. D M Howard, and A J Fourcin. Electronics Letters, 19, 76-78, (1983).
4. B C J Moore, and B Glasberg. In B C J Moore (ed) "Frequency selectivity in hearing (Academic Press, London, 1986), 251-308.
5. M A Huckvale, D M Brookes, L T Dworkin, M E Johnson, D J Pearce, and L W Whitaker. Proc. Eur. Conf. Sp. Tech., (these proceedings), (1987).
6. I S Howard. Proc. Eur. Conf. Sp. Tech., (these proceedings), (1987).
7. D M Howard. Speech Hearing and Language: Work in Progress, (London: UCL), 2, 153-163 (1986).
8. D M Howard, J A Maidment, D A J Smith, and I S Howard. IEE Conf. Pub., 258, 172-177, (1986).
9. A J Fourcin and E R M Abberton. Med. and Biol. Ill., 21, 172-182, (1971).
10. W Hess and H Indefrey. Proc. ICASSP-84, 1-4, (1984).
11. D M Howard and I S Howard. Proc. 11th Int. Cong. Phon. Sci., in press, (1987).

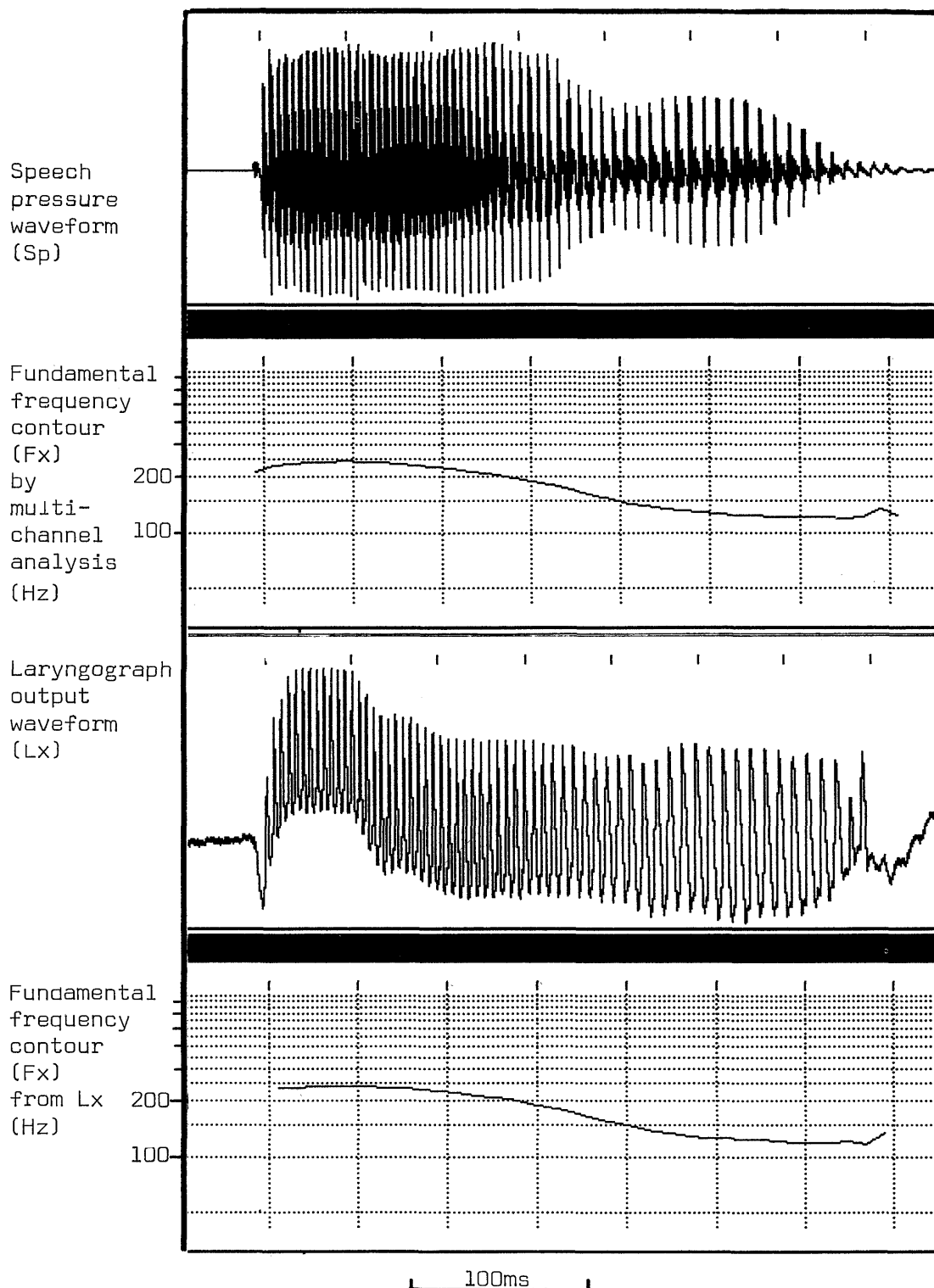


FIGURE 1: Speech pressure, laryngograph output, and fundamental frequency contours derived from the laryngograph output waveform and from the multi-channel peak-picking system.