QUANTITATIVE COMPARISON OF SPEECH FUNDAMENTAL PERIOD ESTIMATION DEVICES

DAVID M HOWARD IAN S HOWARD

DEPARTMENT OF PHONETICS AND LINGUISTICS

UNIVERSITY COLLEGE LONDON, U.K.

ABSTRACT

Speech fundamental frequency estimation devices are usually designed to suit the application for which they are intended. A technique is described which enables the operation of such devices, which operate in the time domain, to be quantitatively compared. It is shown that the use of this technique enables device operating parameters to be fine-tuned in a rigorous manner.

INTRODUCTION

There are many methods available for the estimation of fundamental period in speech, and these can be separated into the following categories as devices which operate: in the time domain on the speech pressure waveform (Sp), in the frequency domain on Sp, in hybrids of the time and the frequency domains on Sp, and directly from an input gained at the level of the larynx (see [1] for a review). To date no one device exists which reliably estimates fundamental period from speech for all speakers in all conceivable operating conditions. Thus the choice of a device for a particular application must be made with due attention being paid to errors which are not acceptable against those which can be tolerated.

Generally this procedure will involve the implementation of the devices under consideration, in hardware or software, and it is not always clear whether the result is operating as intended with a speech input. Further, many designs require an elaborate optimisation procedure for the particular speakers and set of operating conditions for which the final device is destined. These areas are most time consuming and they often leave the designer the formidable task of weighing up the beneficial effect of altering a parameter to, for example, reduce output frequency doubling errors when it is found that this adjustment also causes an increase in voicing onset deinition errors.

Such problems require a quantitative method which enables device performance to be compared using a speech input, of the type one expects when the device is in use, against a standard. Then the setting up of a device could be achieved with reference to a quantity defined, ideally by the designer, for a particular recorded speech input, and optimisation could be carried out with quantified feedback as to the effects that altering parameters has on device performance. Indeed, if appropriate controls are made available and the requirements of the users can be rigorously defined, then this optimisation procedure could become an automated process. This paper describes such a quantitative technique for the assessment of time domain fundamental frequency estimation device performance, and an illustration is given as to how it can be used to optimise device parameters automatically.

DEVICES STUDIED IN THE TESTS

The technique described below [2] can only be used with devices which are designed to produce a pulsatile output where each pulse corresponds to an epoch of acoustic excitation due to vocal fold closure. Such devices usually operate in the time domain, and here an already established time domain device is made the subject of study. This is a peak-picking device [3] which has been developed as the input speech processing stage of the EPI group [4] hearing prostheses for the totally deaf and profoundly deaf. It is a small battery-powered device which operates in the time domain producing a pulsatile output suitable for these tests. The version used in these tests is a software implementation [5] which is written in C which runs under "UNIX" on the department's Masscomp 5500 computers.

This work also requires a 'standard' against which the operation of the device is based on the laryngograph [6], and the algorithm used to detect period epochs is described in [7]. The laryngograph gains its input directly from the vocal folds by measuring the current passing between two electrodes placed on either side of the throat at the level of the larynx. When the vocal folds vibrate the current flow between the electrodes changes and this is clearly shown in the output waveform from the laryngograph (Lx), and an example is shown in figure 1b. The main advantages of using the laryngograph as a standard, a practice also used in [8], is that it is unaffected by competing acoustic noise, and that the Lx waveform conveys the periodicity associated with voiced sounds in a clearly defined manner which can be simply processed to give a suitable pulsatile output.

DESCRIPTION OF ANALYSES

The methodology used is composed of two parts designed to investigate the one-to-one deviations of the pulse markers generated by the test and reference devices -- thus it can be thought of as a 'micro' level comparison. It is complimentary to a 'macro' level (whole passage input) methodology which is being investigated, and the initiation of these is described in [7]. The two stages, described in detail in [8], are as follows:

1) the jitter distribution which is a histogram of the differences in the times of occurrence of output pulses from the reference and the corresponding time-aligned pulses from the device under test, and

2) the receiver operating characteristic (ROC) which is a plot of the probability of successful detection of a vocal fold closure on comparison with the reference (a HIT) against the number of pulses generated with no corresponding pulses in the reference output (FALSE ALARMS).

The ROC enables a quantitative measure to be gained as device operating parameters are altered. The peak-picking device, under test in this case, has a user-adjustable gain control which essentially determines the threshold level for the generation or non-generation of an output pulse. When this is altered there may be a change in the number of HITS and FALSE ALARMS, and this is shown by the ROC for the device. Each point on the ROC is plotted as the percentage of HITS generated against the number of FALSE ALARMS. As the gain is altered the points on the ROC trace out a curve (see figure 3). As the gain is lowered the number of HITS will increase, but so will the number of FALSE ALARMS. In general just one point for a particular device will specify the position of the ROC curve which indicates how detectable the signal is to the algorithm/device. Device operation can be ranked since those producing outputs highly similar to the reference will have some point on the ROC more closely approaching the perfect performance point (FALSE ALARMS = 0,HITS = 100%).

QUANTITATIVE COMPARISON METHODOLOGY

The data for this work was taken from a passage recorded by a male speaker (JM) in the anechoic room at UCL. A two channel digital (pcm) recording (Sp and Lx) was obtained, and the sentence "We can learn a little something from the birds, he said"
was transferred onto a Masscomp 5500 computer at a sampling rate of 12800Hz using a 12 bit ADC via a suitable anti-aliasing filter. The Sp and Lx waveforms are shown in figure 1a and 1b.

RESULTS

The reference, based on Lx, produces the period markers, and the reciprocal of these are plotted to give a fundamental frequency with time (Fx) trace in figure 1c. The peak-picker also produces period markers, which are not shown here due to lack of clarity on this scale, its outputs for a series of gain settings being shown as Fx contours (see figures 1d to 1h which corresponding to gains of 0.03, 0.1, 0.25, 0.5 and 1.0 respectively). In this manner a visual comparison can be made between the operation of the peak-picker with different gain settings, and the reference, and it can be seen that the gain appears optimum around a value of 0.25 .

This value of gain has been used for the peak-picker in both the jitter histograms shown in figure 2. They are plotted for the peak-picker (test device) against the laryngograph-based method (reference device) for (a) anechoic speech (figure 2a), and (b) anechoic speech degraded with white noise, SNR = 6dB (figure 2b). It can be seen that there is greater deviation from the zero jitter point with noise contaminated speech.

The ROC curves for these two speech input conditions are shown in figure 3. As the peak-picker gain is increased, a curve is traced away from the origin. Ideally optimum gain would result in a point at (hits = 100%,false alarms = 0). In practice, however, the optimum will only

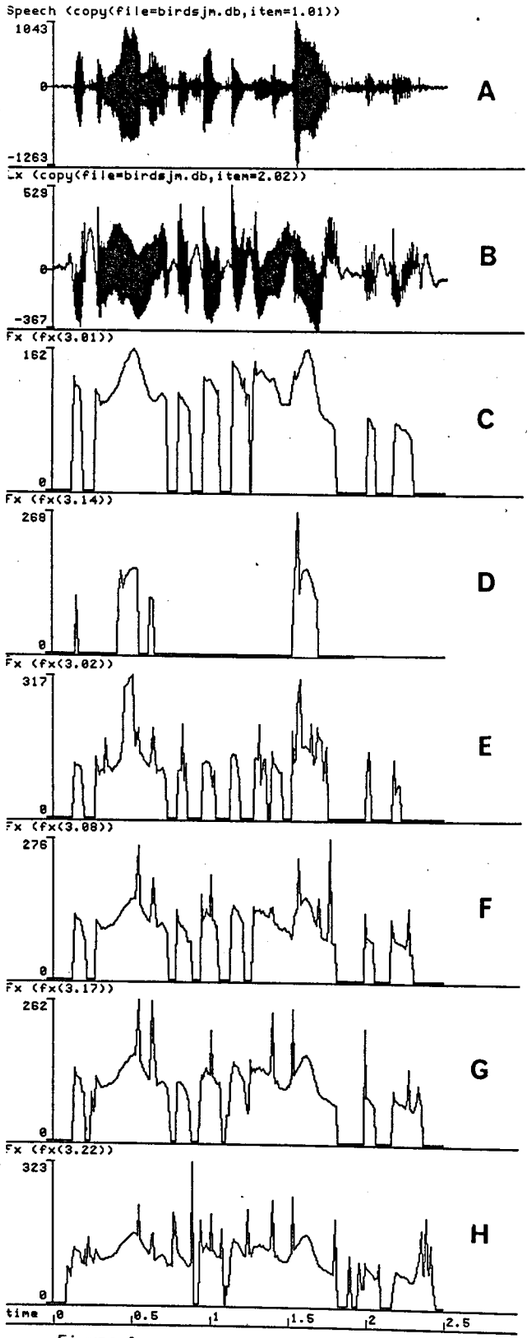

Figure 1
A) speech, B) Lx and C) Fx from Lx, D) to H) Fx from peak-picker.

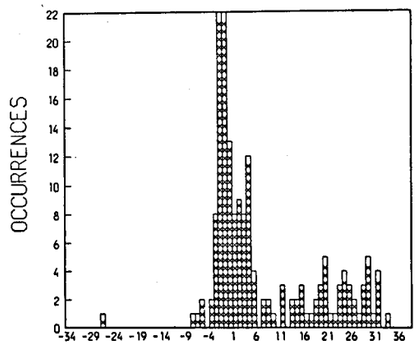

Figure 2a.
Jitter histogram for recording room quality speech .

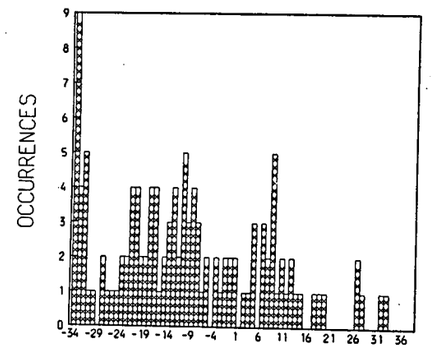

Figure 2b.
Jitter histogram for speech contaminated with uniform density noise (SNR = 6 dB).

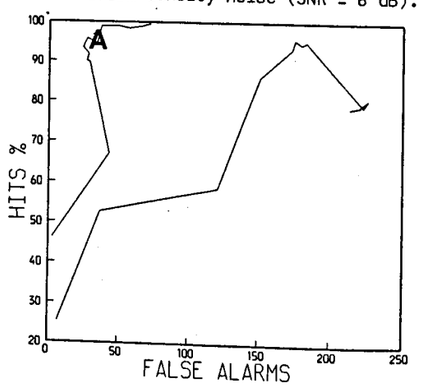

Figure 3
Top curve ROC for peak-picker with recording-room quality speech, lower curve for noise corrupted case.

approach this point and will depend on a trade-off between number of hits required against the error rate. It can be seen, in this case, that point A on the top curve is a good choice for optimum gain (point A corresponds to a gain of 0.25) because a higher gain only results in a marginal increase in the number of hits for a considerable increase in the false alarms. This value also corresponds to the value determined above for optimum gain from the Fx contours (see figure 1c to 1h). With the addition of noise device performance is degraded, and this is shown by its ROC which is below the other curve for all gain settings.

From these results, it can be concluded that the ROC gives a basis for an automated optimisation and assessment technique. In the particular case discussed above optimum gain has been selected by observation of the ROC and of the Fx contours. In practice any parameter could be optimised automatically using the ROC method for the particular application for which the fundamental period device is intended.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Hess, W., "Pitch determination of speech signals", Springer-Verlag, Berlin, (1983).

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

[3] Howard, D.M. and Fourcin, A.J. (1983), "Instantaneous voice period measurement for cochlear stimulation", Electronics Letters, 19, 19, 776-778.

[4] Fourcin, A.J., Douek, E.,Moore, B.C.J., Rosen, S.R., Walliker, J.R., Howard, D.M., Abberton, E.R.M., Frampton, S., "Speech perception with promontary stimulation", An. New York Acad. Sci., 405, 280-294, (1983).

[5] Howard, D.M., "Digital peak-picking fundamental frequency estimation". Speech hearing and language; Work in progress, 2, London: UCL, (1986).

[6] Fourcin, A.J., and Abberton, E.R.M., "First applications of a new laryngograph", Med. and Biol. Illust. 21, 172-182, (1971).

[7] Howard, I.S., and Howard, D.M. (1986). "Quantitative comparisons between time domain speech fundamental frequency estimation algorithms", Proc. Inst. Acoust., 8, 7, 323-330.

[8] Hess, W. and Indefrey, H., (1984). "Accurate pitch determination of speech signals by means of a laryngograph", Proc. ICASSP-84, 1-4.