Proceedings of The Institute of Acoustics

QUANTITATIVE COMPARISONS BETWEEN TIME DOMAIN SPEECH FUNDAMENTAL FREQUENCY ESTIMATION ALGORITHMS.

Ian S. Howard and David M. Howard

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

ABSTRACT

Two techniques are presented here to enable quantitative comparison of time domain fundamental frequency estimation algorithms against a reference, that makes use of the output from a laryngograph. These measures are carried out on the pulsatile outputs produced by the devices, where each pulse corresponds to an epoch of acoustic excitation due to a vocal fold closure. The results given here are for a peak-picking algorithm. The comparison techniques are:

1) Receiver operating characteristic.
This is a plot of the probability of successful detection of a vocal fold closure, as compared to the reference, against the number of false alarms. It is shown that this measure gives a clear indication as to how well the device under test performs with respect to the reference, as well as providing a quantitative method for device parameter optimisation.

2) Jitter distribution.
This is a histogram of the differences in the times of occurence of output pulses from the reference and the corresponding time-aligned pulses from the device under test. This measure gives an indication of how precisely and consistently devices are able to locate vocal fold closure instants.

INTRODUCTION.

This work is aimed at establishing a technique which gives an automatic quantitative evaluation of the performance of speech based fundamental frequency estimation devices which operate in the time domain. Typically the output from such devices takes the form of excitation epoch positions, each being signalled by a pulse. In order to make a successful evaluation, a reliable 'standard' is required, and in this work a laryngograph [1] is employed for reference purposes - a practice endorsed by Hess [2]. Automatic assessment strategies not only give a fast and convenient result whilst giving a quantitative basis for improving confidence in the device under test, but also they lead to the possibility of using iterative optimization procedures to adjust parameters in order to improve the device operation.

DEVICES USED FOR FUNDAMENTAL FREQUENCY ESTIMATION.

Devices and algorithms for estimating speech fundamental frequency can be broadly categorized as: those which operate on the speech pressure waveform in either 1) the time domain, 2) the frequency domain, or 3) hybrids of 1 and 2;

and 4) those which derive their input directly at the level of the larynx (see [3] for a comprehensive review).

In a time-domain device, the fundamental frequency estimate is made directly from the speech waveform by measuring its periodicity. Typically, this can be enhanced by the use of various pre-processing stages, and the output is obtained as either a pulse train or a series of durations relating to each detected period.

Frequency domain devices are designed to perform a spectral analysis on successive portions of the speech waveform. They may take advantage of the harmonic structure associated with voiced excitation. The fundamental frequency estimate arrived at by such devices is thus typically specified at equally spaced time intervals. This format is not suitable for the comparisons described here, as they require data in the form of the time of occurence of excitation events (Tx), which can then be directly compared with the laryngograph reference (see below).

A device which makes a direct measurement of vocal fold activity is the laryngograph. A laryngograph works by measuring the electrical impedance across the larynx at the level of the vocal folds. Movement of the vocal folds results in an impedance change which can be detected by the laryngograph. The appropriately polarized laryngograph output waveform (Lx) thus gives a direct measure of voicing and its time domain representation is much simpler than the corresponding speech pressure waveform. Therefore, by means of relatively simple time-domain processing, a good estimate of the fundamental period of the speech can be obtained. It should be noted however, that the Lx waveform does not always give a strong indication of vocal fold activity in all cases when observation of the speech waveform and its spectrum indicates that voiced excitation is indeed present. This mainly occurs towards the end of unstressed voiced segments, when the vocal folds are still vibrating, but no firm closures are made, and therefore the measured larynx impedance shows little change.

The algorithm used to estimate the period epoch positions from the Lx waveform makes use of the fact that the closing phase of the vocal folds gives rise to a point of maximum gradient of the Lx waveform. This point is unique and well defined in each cycle, thus making it a good feature to define period epochs. This position is located by first differentiating the Lx, and then searching for the best local maximum, subject to the constrains that it must exceed a predetermined threshold value in amplitude and that only one Tx pulse is allowed to be found within a minimum predefined interval. Any isolated single pulses are then rejected on the grounds that they were probably not acoustically significant, but rather related to precursive larynx adjustments prior to phonation.

The peak-picker under test here [4] is a time domain device that generates a pulse every time it finds what it considers to be a period epoch marker in the speech pressure waveform. The version used here is a software model of a small

battery powered device developed as part of the EPI group cochlear implant and acoustic protheses [5]. Its operation is essentially instantaneous, with no long-time constraints being applied to its output, for it is in this way that it can best benefit both the speech perception and production of prothesis users [6]. The only memory in the system is due to the time response of the low pass filter and secondary peak suppressing circuitry, which are fundamental to the operation of the device.

TYPES OF COMPARISON.

The comparisons investigated here are based on the one-to-one deviations of the Tx pulses from the device under test to those due to the reference. These comparisons are:
1) The relative pulse jitter of the reference Tx to the corresponding pulse (if it exists) in the test Tx.
2) Receiver operating characteristic (ROC). This measure has its foundations in statistical decision theory. To understand the basic principles of this measure, consider the operation of the time-domain fundamental frequency estimation algorithm as being conceptually broken down into a pre-processing stage followed by a threshold stage. The function of the pre-processing stage is to produce, from a noisy speech signal, a waveform that gives the maximum possible discrimination (to the following threshold circuit) between the period-epoch-marker-present case and the period-epoch-marker-absent case. The threshold section then merely decides into which of the two categories to place the input data, that is, whether or not to generate a Tx pulse. If only the decision criterion of the algorithm is changed, by raising or lowering the threshold, then there will probably be a change in the output. However no fundamental changes to the algorithm have been made and the signal is just as detectable to the algorithm in each case. It is this inherent detectability that the ROC of the device will show. A point on the ROC is a plot of the probability of the device correctly identifying a period epoch marker (a hit) versus the probability of the device commiting an error, that is, by indicating that a period epoch marker was present when this as not the case (a false alarm). Clearly as the threshold becomes lower, there will be more hits but also more false alarms. If the probability of hits vesus probabilities of false alarms are plotted for different threshold criteria, the points will trace out a curve. The position of this curve (which for one particular device can be specified the definition of a single point) is indicative of how detectable the signal was to the algorithm.

It is legitimate to plot the percentage of hits versus the number rather than the probability of false alarms for the ROC, provided that they are all plotted for the same input data (that is, with the same speech and reference Tx) on the same graph with the same scales. The probability of a false alarm would be estimated by the ratio (false alarms)/(maximum possible number of false alarms), but since the denominator is constant for all the points plotted on the graph, it can be ignored without changing the relative shape of the curves. The reference Tx gives the basis for a decision as to whether or not a pulse in the test Tx is a hit or a false alarm, a hit occuring when there is a

corresponding test Tx pulse for a given reference Tx pulse, and a false alarm occurs when there is a test Tx pulse and no corrisponding reference Tx pulse.

In order to carry out the two comparisons it is necessary to know, for the ROC, for each reference Tx pulse, whether or not there is a corresponding test Tx pulse (a hit), as well as the number of false alarms, and for the jitter distribution to also have access to the relative time shift of established hits. The algorithm used to find the "optimum" correspondence of the reference Tx pulses and the test Tx pulses employs a dynamic programming approach.

Pulse corespondence routine.
This routine is designed to decide which, if any, pulses in the test Tx pulse train correspond to those present in the reference Tx pulse train. There will in general be a time delay between the two Tx pulse trains due not only to the different time delays in processing, but also due to the fact that the Lx signal is derived ahead of the speech pressure waveform due to the propagation delay the latter incurs in reaching the microphone. The overall time delay is estimated by cross correlating the two Tx pulse trains and locating the time at which the maximum in the result occurs. This delay is then used to time align the two Tx pulse trains. For each of the reference Tx pulses it is then assumed that the corresponding test Tx pulse is either 1) the nearest, 2) the second nearest, or 3) there is no corresponding test Tx pulse.
A dynamic programming algorithm is then employed to optimize the correspondence, the criteria of optimality being to minimize the sum of the magnitudes of the individual pulse jitters. The solution is constrained to only allow correspondence of pulses within a predetermined range and that the correspondence must be monotonic. Also a penalty (corresponding to the maximum allowed jitter) is paid for deciding there is no corresponding pulse for a given reference Tx pulse. This is to prevent the trivial solution of finding no pulses and therefore having no global jitter. The algorithm thus works as follows:
1) For each reference Tx pulse, the nearest two test Tx pulses are found.
2) If a test Tx pulse is further away from the reference Tx pulse than a predetermined limit, then it is rejected.
3) Starting with the last reference Tx pulse, the "optimum" two alternatives are noted. Then moving backwards by one reference Tx pulse, for each of its two possible local correspondances, the "optimum" of the two possible paths forward from each one in turn is then chosen. Moving backwards again by one reference Tx pulse, for each of the two local possibilities the "optimum" path forward is again chosen. This proceeds until all the reference Tx pulses have been investigated and the result is a list of the "optimum" correspondences of each reference pulse either a test Tx pulse or not. From this list the number of "hits" is then obtained along with their jitter values. The number of false alarms is then calculated from as the difference between the total number of test Tx pulses and the number of hits.

QUANTITATIVE COMPARISONS BETWEEN TIME DOMAIN SPEECH FUNDAMENTAL FREQUENCY ESTIMATION ALGORITHMS.

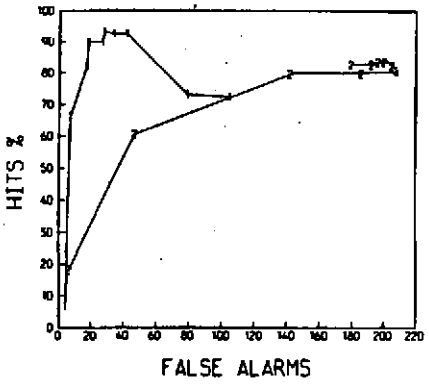


Figure 1: ROC for the peak-picker
CURVE 1: Different peak-picker gain settings for recording room quality speech.
CURVE 2: As curve 1, but speech is decraded by addition of uniform density random noise.

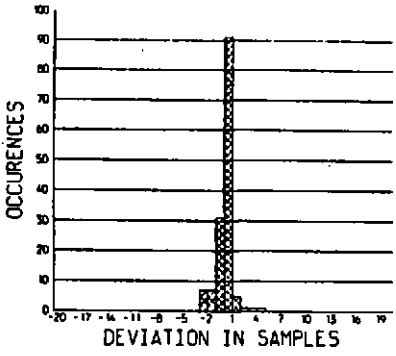


Figure 2: Jitter histogram for two Lx to Tx algorithms.

RESULTS

The results presented here are for the utterance "speech patterns on view", spoken by a male speaker. The fundamental frequency estimator algorithms as well as the comparison algorithms were implemented on a Masscomp 5500 series computer and the speech pressure waveform and Lx were sampled at 12800 Hz with a 12 bit ADC, with appropriate anti-aliasing filtering.

The ROC curve for the peak-picker is shown in figure 1. The curve with points marked "1" correspond to the performance with natural speech, and the curve with the points marked "2" corresponds to using the same speech as before but this time corrupted with uniform density random noise. In each case moving along a curve away from the origin corresponds to increasing the peak-picker gain. It can be seen that as the gain of the peak-picker was increased it was more likely to detect an excitation (up to a limit) but at the same time there was a greater likelihood of generating a false alarm. By comparing the two curves it is evident that adding the noise degraded the device performance such that its repective curve was always below the other, as is to be expected. The noise present case can be thought of as equivalent to the performance of an inferior device. A 'better' design will have a curve which is further towards the top left of the ROC. Saturation of the device is probably the cause for the curve folding back on itself in the noise present case, where the amplitude compression stage is presenting signal and noise peaks to the peak-pickers at much the same level, and the random nature of these is causing increased hits and fewer false alarms.

Histograms of the jitter between the hits in the reference Tx and test Tx for speech with different levels of uniform density random noise contamination are shown in figures 3,4,5 and 6. It can be seen that the jitter also became considerably worse as the noise amplitude was increased. This indicates that the peak-picker was less able to precisely define the fundamental period epoch as the noise level was increased. Figure 2 shows the jitter histogram for another Lx to Tx routine. The difference between this routine and the reference is that the Lx to Tx routine under test only thresholds the differentiated Lx waveform to determine the epoch period marker, whereas the reference Lx to Tx located the local maximum in the differentiated waveform. Notice how well the two different Lx to Tx algorithms compare.

CONCLUSIONS AND FUTURE WORK

Two new measures have been described which have been implemented on a Masscomp 5500 computer with a view to make quantitative comparisons between speech fundamental frequency estimation devices which operate in the time domain. For time-domain devices these comparison techniques have clealrly demonstrated some useful aspects of performance, as well as providing the basis for device parameter optimization. Thus the possibility exists to utilize these techniques in the development of new time domain algorithms. Further work is aimed at ensuring that the very best reference Tx system is being used, and therefore other Lx to Tx algorithms are being investigated and compared with the current reference. Fundamental frequency estimation algorithms based on frequency domain or hybrid techniques cannot be subject to a ROC analysis since their outputs are not in the form of acoustic excitation epoch markers, upon which the analysis depends. However in cases where a threshold is used with frequency domain devices for voiced/voiceless decisions, the ROC technique may be applicable. At present other techniques are being investigated with a view to including any acoustically based devices in future comparisons.

ACKNOWLEDGEMENTS

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

QUANTITATIVE COMPARISONS BETWEEN TIME DOMAIN SPEECH FUNDAMENTAL FREQUENCY ESTIMATION ALGORITHMS.

Figure 3

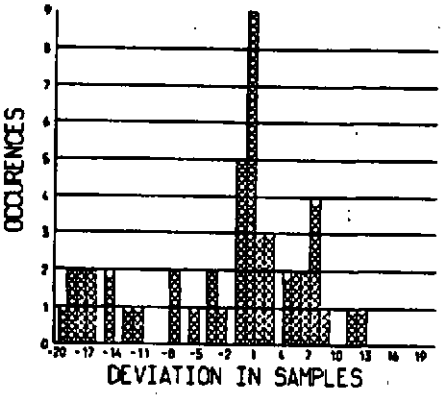


Figure 4

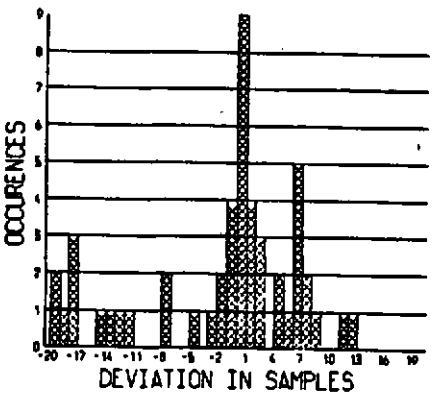


Figure 5

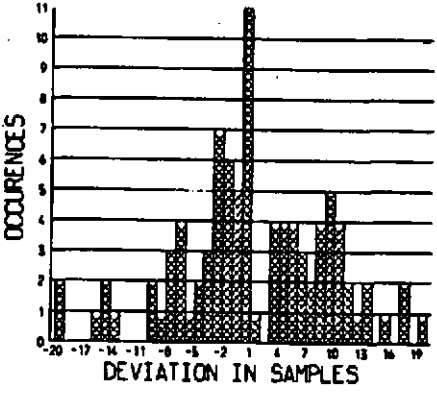
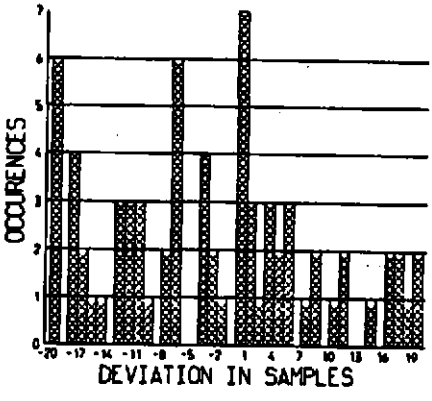


Figure 6



JITTER HISTOGRAMS FOR THE PEAK-PICKER

Figure 3: Recording room quality speech.
Figure 4-6: Speech degraded by increasing uniform density noise.

QUANTITATIVE COMPARISONS BETWEEN TIME DOMAIN SPEECH FUNDAMENTAL FREQUENCY ESTIMATION ALGORITHMS.

REFERENCES

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

[2] Hess, W. and Indeffry, H., Proc. ICASSP-84, 1-4, (1984).

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

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

[5] 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).

[6] Abberton, E., Fourcin, A.J., Rosen, S., Walliker, J.R., Howard, D.M., Moore, B.C.J., Douek, E.E., and Frampton, s., "Speech perceptual and productive rehabilitation in electro-cochlear stimulation"., In Schindler, R.A., and Meryenich, M.M., (eds), Cochlear Implants, New York: Raven press, 527-537, (1985).