Further developments of a neural network speech fundamental period estimation algorithm

I Howard

Nokia Mobile Phones, UK.

Abstract
This work describes a speech fundamental period estimation algorithm that estimates the time of excitation of the vocal tract using a pattern classifier, the multi–layer perceptron (MLP). The pattern classifier was trained using speech semi–automatically labelled by means of an algorithm that makes use of the output from a Laryngograph. Various issues arising in the training of the system were explored. Three basic configurations of the system were compared using different pre–processing strategies. It was found that processing the sampled speech time–waveform directly with the pattern classifier gave better results than using one of two filterbanks. The performance of the algorithm was evaluated against that of a simple peak–picking algorithm and the well known cepstrum algorithm using quantitative frequency contour comparisons. The performance of the new algorithm on a difficult set of test data was shown to be better than the peak–picker and comparable to the cepstrum algorithm. The advantage of the scheme is that fundamental period estimates are made on a period–by–period basis, thus preserving the irregularity in the speech excitation that is lost by techniques that produce as average period estimate. In addition, its simple structure lends itself to real–time implementation (Howard & Walliker, 9; Walliker & Howard, 14).

Introduction

Speech fundamental frequency estimation has long been the subject of much investigation, and two basic approaches to the problem can be classified as those that use short–term or time domain analyses (Hess, 1983). Short–term analyses typically result in an average estimate of fundamental frequency over a short–time (for example, 20ms). However, with some time–domain approaches, it is possible to perform speech fundamental period estimation, in which the estimate of the time of occurrence of excitation takes place on a period–by–period basis. Such an approach is desirable in those cases when one wishes to retain the irregularity that is inevitably present in the speech excitation which are lost by short–term analysis.

This paper presents a brief account of a system that has been developed to detect the cycle–by–cycle excitation of the vocal tract, which thus provides a cycle–by–cycle of speech fundamental period. A full account of the work presented in this paper appears in Howard (6) and in a more compact cut–down form in Howard (7).

Basic principle of operation

The algorithm described here formulates speech fundamental period estimation as a pattern recognition problem. The overall system schematic diagram is shown in figure 1. The input speech signal is first conditioned (using anti–alias filters and then a 12–bit A/D converter running at 8kHz). Next there is a pre–processing stage, to format the system to suit the pattern classifier (that is, to scale in input range to +–1.0) and emphasise the important aspects of the input. In this paper, three different pre–processing schemes were considered. A 6–channel filterbank with filter bandwidths equivalent to a wideband spectrogram (300Hz) covering the frequency range of 50Hz–3kHz, a 19–channel filterbank with filter bandwidths based on auditory bandwidths (Holdsworth et al., 4) and another scheme using no filtering at all. The frame rate for the filterbank schemes was 2kHz, whereas it remained at 8kHz for the direct speech scheme. The pattern classifier itself has the task of making the decision as to the presence or absence of an excitation in the input. Finally, the post–processing stage detects the output from the classifier if it is above a threshold value and converts the result into the desired format

We shall now concentrate of the function of the pattern classifier stage. Consider the case when we observe a series of speech samples ahead in time of the current one, as well as the same number past in time from the present one. This is the same scenario as in a symmetrical FIR filter, in which the filter has access to both past and future samples of the input signal. The input vector to the pattern classifier is defined in a similar way as resulting from the input samples (or elements in the input frames) of the data over an equivalent input window. In this case, the output from the pattern classifier is either one of two possible classes: Either there is a period excitation present at a the current frame, or there is not. During recognition mode, it is the task of the classifier to detect the excitation points in the input speech signal. During the training phase of the algorithm, it is necessary to specify the occurrences of the vocal fold excitations so that the pattern classifier can be trained, and this is done by means of an interactive algorithm that uses the output from a Laryngograph (Fourcin & Abberton, 2). This is a simple device that measures the electrical admittance across the larynx at the level of the vocal folds. Changes of vocal fold contact show up well in its output, and this provides a means of the identification of the point of maximum vocal tract excitation using only simple processing.

Labelling the training and test data

This algorithm used to identify the points of vocal fold excitation operates in two phases. Firstly, an automatic procedure differentiates the Laryngograph waveform and then applies a simple local threshold analysis to it. This results in a first estimate of the location of the vocal fold excitation points. An interactive algorithm is then used, whereby an operator can zoom in and view the input speech, Laryngograph and differentiated Laryngograph waveforms and edit the estimated excitation points. In addition, a facility is provided so that unreliable sections of waveform can be rejected and not used for further analysis. The labelled data can then be

used to train the MLP pattern classifier.

The database

To train and test the MLP algorithm, two phonetically balanced passages were used. These comprised "The story of Arthur the Rat " (Abercomie, 1) and "The Rainbow Passage " (Mermelstein, 10). For training, both passages were used in their entirety for 4 female speakers. For testing, one paragraph (about 15 seconds of speech) for each of 20 female speakers was used. Females were used because earlier work had established results for male speakers (Howard & Huckvale, 8). Recordings were made in naturally reverberant and noisy conditions (such as offices, lounges) at a range of distances between 30cm and 200cm.

The training database was recorded so that the effect of head movements were minimised. This manifests itself as a non–constant delay between the speech pressure waveform and the Laryngograph waveform. Unless this delay is constant, it is very difficult to correct for it. It is clearly necessary to compensate for this time delay because it is different for different recording distances and to ensure that all the training data is self–consistent, all training data must have the excitation labels aligned consistently with the corresponding speech pressure waveform.

Alignment of the training data was carried out using a "bootstrap" procedure. This involved initially aligning one section of speech and Laryngograph data such that the peaks in the speech aligned with the point of maximum gradient in the Laryngograph waveform. This data was then used to train a speech–input MLP–Tx algorithm. The partially trained MLP–Tx algorithm was then run on the remained of the speech training data. The cross–correlation between the local peaks in the MLP output and the local peaks in the differentiated Laryngograph waveform were then computed. The results appears in figure 2. The location of the corss–correlation maximum provides the delay between the speech and Laryngograph waveform and this was used to align the two waveforms. Notice that the cross–correlation also provides a reliable method of speech polarity determination. Figure 3 shows the effect of speech inversion of the corss–correlation. In this case, the maximum is much less pronounced.

After alignment has been achieved, the MLP algorithm generated an output that is time–aligned to the point of maximum gradient in the Laryngograph waveform, as shown in figure 4.

It was not necessary to use constant distance recording for the testing data because the frequency contour comparisons employed were not significantly affected by the movement of the subject's head.

Selective emphasis training of the MLP

Training the MLP classifier was carried out using the back–propagation. learning rule (Rumelhart et al., 13). However, it was found that normal training was exceedingly slow and certain additional procedures were used. These all make use of the idea of selectively emphasizing training pattern vectors depending upon their importance. This was done with regard to two main criteria: Firstly, a pattern that occurs at the boundary of the excitation period is de–emphasised , because of the similarity that it may have (in patten space) with its neighbours. If such a pattern was not de–emphasised, forcing it into one of two classes would affect the recognition of other patterns that are similar, but in the other class. Secondly, patterns are only used if they are wrongly recognized (that is, if they evoke a response from the classifier during training that is further from the target value than a preset threshold). In this was training can concentrate on those patterns that are falsely classifier and ignore those which are correctly classified. In this way computation is not wasted on making insignificant weight changes.

Processing the test data

The MLP algorithm was run in its three different configuration on the test data set, together with a cepstrum algorithm and a peak–picker (Howard, 5) algorithm (Noll, 11). The cepstrum algorithm operated with an input window of 30ms and was chosen because it is an established standard. The peak–picker is a simple time–domain algorithm that operates by locating the peaks in the speech waveform, due to the excitation whilst suppressing those due to harmonics of formant resonances. It was used because it is the algorithm the MLP–Tx system is to replace. The output period marker was found by a simple threshold algorithm that located the local maximum over a range of 10ms forward and 10 back from the current sample in the output from the MLP classifier. The reciprocal of the period value was then computed to give as estimate of the local equivalent fundamental frequency.

Frequency contour comparisons

The algorithms were evaluated using frequency contour comparisons (Rabiner et al., 12). This paper reports the voiced–to–unvoiced errors, the unvoiced–to–voiced errors and the chirp and drop gross errors generated by the algorithm, where a chirp error is defined as one in which the test frequency contour exceeds the value of the reference in frequency at a frame by more than 10%, whereas a drop error is defined as one in which the test frequency contour is lower than value of the reference in frequency at a frame by more than 10%.

Conclusions

The results to the frequency contour comparisons appear in figures 4–7. It can be seen that the direct speech pre–processing gave fewer chirp and drop errors than the other filterbanks. In addition, the auditory filterbank gave better results than the 6–channel wideband filterbank. In terms of voicing determination, it can be seen that the MLP algorithm gave better results than the peak–picker or cepstrum algorithms. The drop and chirp errors for the cepstrum were lower for the cepstrum than for the MLP algorithm, but is must be borne in mind that the former included averaging over the input window and a gross error correction routine, whereas the MLP algorithm used a simple threshold algorithm on the output of the classifier.

References

1. Abercomie, (1964), English phonetic texts, London, Faber & Faber.

2. Fourcin, A.J., & Abberton, E., (1971), First application of a new Laryngograph, Med. and Biol. Illust., Vol. 21, pp172–182.

3. Hess, W., 1983, Pitch determination of speech signals, Springer–Verlag.

4. Holdsworth, J., Nimmo–Smith, I., Patterson, R., & Rice, P., (1988), Implementing a Gammatone Filterbank, Annex C of SVOS Final; Report, MRC, APC, Cambridge.

5. Howard, D., (1986), Digital peak–picking fundamental frequency estimation, Speech, Hearing and Language,; Work in progress 2, Department of Phonetics and Linguistics, UCL London.

6. Howard, I., 1991, Speech fundamental frequency estimation using pattern classification, PhD Thesis, Submitted, University of London.

7. Howard, I., 1991, Speech fundamental frequency estimation using pattern classification, Work In Progress, Department of Phonetics & Linguistics, University College London.

8. Howard, I.S., & Huckvale, M.A., (1988), Speech fundamental period estimation using a trainable pattern classifier, FASE88, Edinburgh.

9. Howard, I., & Walliker, J.R., (1989), The implementation of a portable real–time multi–layer perceptron speech fundamental period estimator, Proc. Eurospeech, Paris.

10. Mermelstein, P., (1977), On detecting nasals in continuous speech, JASA, Vol. 61, pp581.

11. Noll, A.M., (1970), Cepstrum pitch determination, JASA 41, pp293–309.

12. Rabiner, L.R., Cheng, M.H., Rosenberg, A.E., McGonegal, C.A., (1976), A comparative study of several pitch detection algorithms, IEEE Trans. ASSP–24, 5, pp399–413.

13. Rumelhart, D.E., Hinton, G.E., & Williams, R.J., (1986), Learning internal representations by error propagation, Parallel distributed processing: Explorations in the microstrucrure of cognition, Volume 1, Cambridge, MA, MIT Press, pp318–362.

14. Walliker, J.,R., & Howard, I., (1990), Real–time portable multi–layer perceptron voice fundamental –period extractor for hearing aids and cochlear implants, Speech Communication 9, Elsevier Science Publishers B.V., (North Holland), pp63–71.

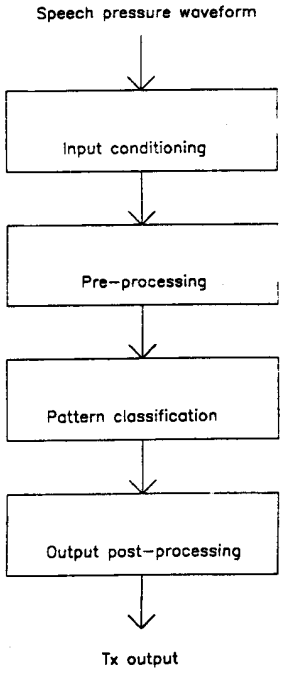

FIGURE 1
SCHEMATIC DIAGRAM FOR STAGES IN MLP–TX ALGORITHM

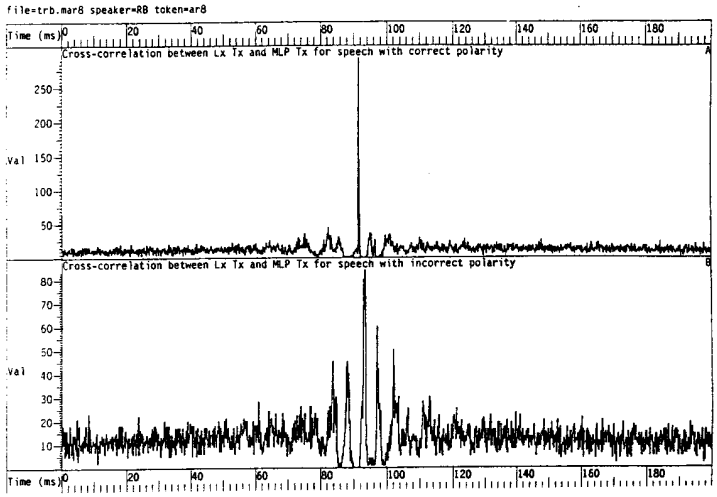

FIGURE 2
TOP TRACE: EFFECT OF CROSS–CORRELATING PERIOD MARKERS FROM PARTIALLY TRAINED MLP–TX ALGORITHM AND PERIOD MARKERS FROM LARYNGOGRAPH WITH SPEECH OF CORRECT POLARITY. BOTTOM TRACE: EFFECT OF CROSS–CORRELATING PERIOD MARKERS FROM PARTIALLY TRAINED MLP–TX ALGORITHM AND PERIOD MARKERS FROM LARYNGOGRAPH WITH SPEECH OF INVERTED POLARITY.

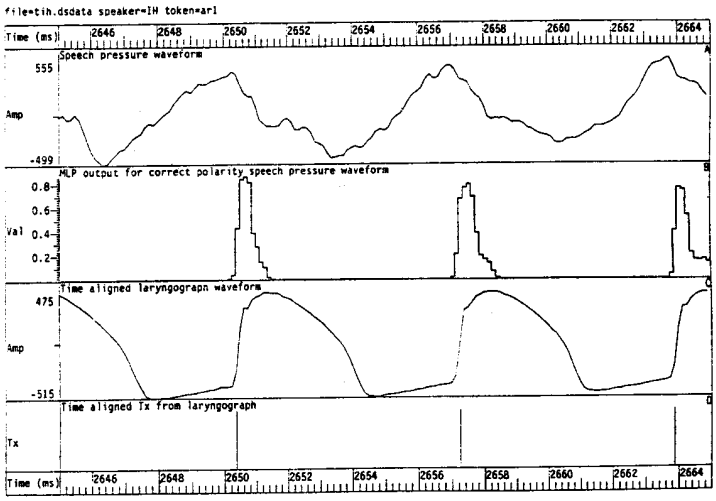

FIGURE 3
DIAGRAM SHOWING SPEECH PRESSURE WAVEFORM IN TRACE A, OUTPUT FROM MLP–TX ALGORITHM IN TRACE B, ALIGNED LARYNGOGRAPH WAVEFORM IN TRACE C AND ALIGNED PERIOD MARKERS IN TRACE D.

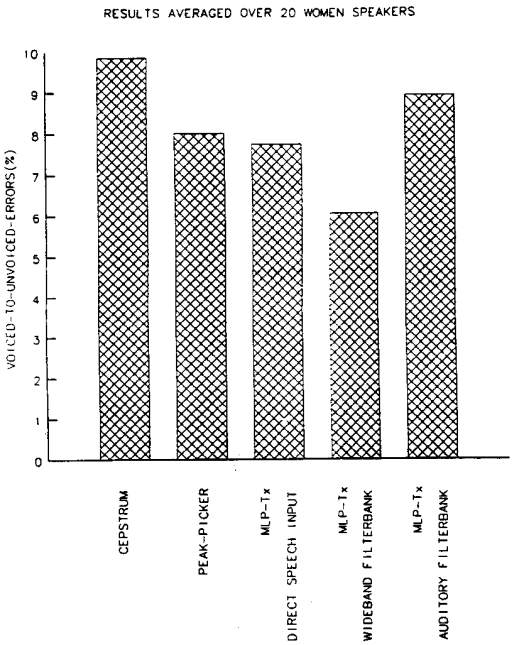

FIGURE 4
VOICED–TO–UNVOICED ERRORS FOR ALGORITHMS UNDER TEST.

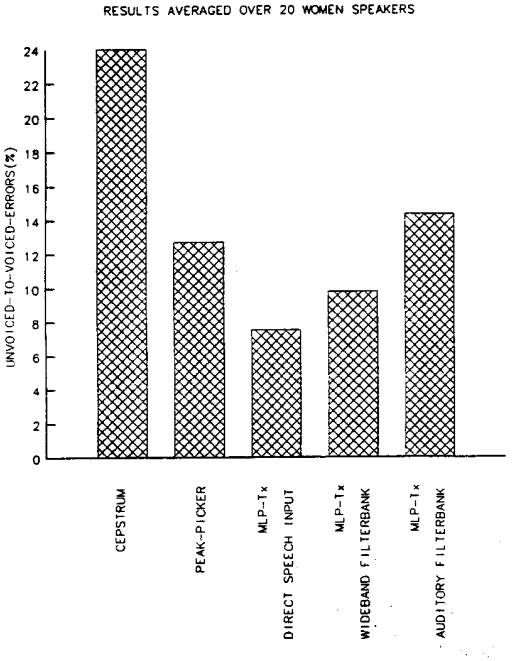

FIGURE 5
UNVOICED–TO–VOICED ERRORS FOR ALGORITHMS UNDER TEST.

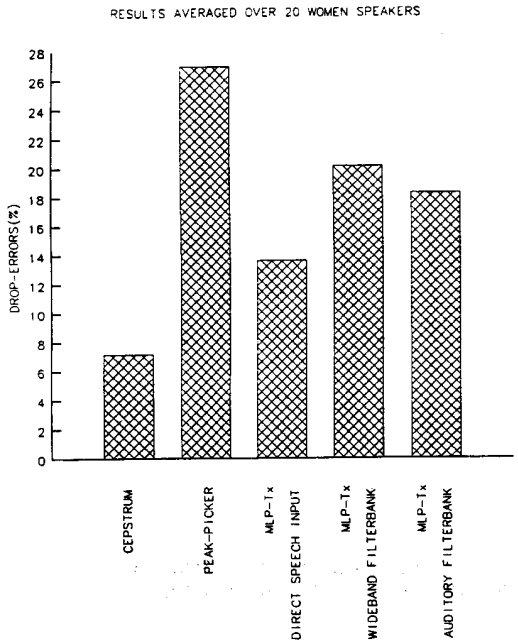

FIGURE 6
DROP GROSS ERRORS FOR ALGORITHMS UNDER TEST.

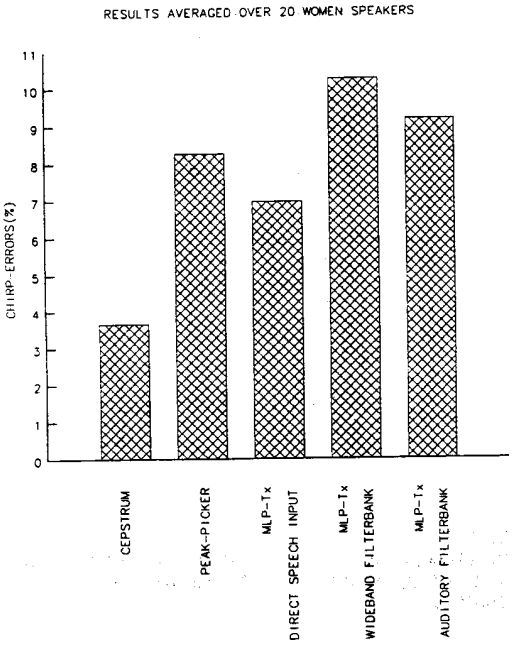

FIGURE 7
CHIRP GROSS ERRORS FOR ALGORITHMS UNDER TEST.