THE IMPLEMENTATION OF A PORTABLE REAL-TIME MULTILAYER-PERCEPTRON SPEECH FUNDAMENTAL PERIOD ESTIMATOR

I.S. Howard [1] and J.R. Walliker [1 & 2]

[1] Department of Phonetics and Linguistics, University College London, Gower Street, London WC1E 6BT, UK.
[2] Department of Clinical Physics, Guy's Hospital, London SE1 9RT, UK.

ABSTRACT

This paper describes the real-time implementation of a speech fundamental period estimation algorithm known as MLP-TX. The algorithm uses a multi-layer perceptron (MLP) classifier to determine the location of the points in time of vocal fold closure in a noisy speech signal. The real-time implementation was carried out in assembly language using a TMS320C25 signal processor. The system will be used in the next generation of SiVo hearing aids developed by the EPI group at UCL. These aids provide voice frequency information as a guide to lipreading and voice control for the profoundly deaf.

INTRODUCTION

Techniques that have been developed to determine speech fundamental frequency can be broadly classified into four main groups [1]. Those that operate in the time domain, those that operate by means of some short-term frequency domain analysis, those that operate as a hybrid of the first two and finally those that measure vocal fold activity electrically.

Time domain fundamental frequency determination involves looking for direct evidence of an excitation point in the speech waveform. A consequence of this is that time domain algorithms tend not to involve smoothing in their fundamental period estimates, and consequently it is easier to retain irregularities in vocal fold vibration than with short-time analysis algorithms. The MLP-TX algorithm is of this type [2]. With conventional techniques there is a trade-off between the preservation of fundamental period irregularity, the time delay between the input speech and the output estimate, and good performance in noise. The advantages the MLP-TX algorithm has over its competition are due to the small time delay with such an algorithm, its robustness in the presence of noise and its inherent preservation of irregularity in the fundamental period estimate. These features of the MLP-TX algorithm make it particularly suitable for use in feature extracting hearing aids [3] [4].

DESCRIPTION OF THE ORIGINAL MLP ALGORITHM

The formulation of the problem of speech fundamental period estimation into a pattern recognition framework can be seen from figure [1]. The basic idea is that one has a series of observations over a given window of a sampled speech pressure waveform in time, which is then processed by means of some non-linear transformation. It is required that the resulting output corresponds to the desired target pattern class. That is, a pulse is generated every time evidence of a vocal fold closure is detected. As with any classical pattern recognition problem, the task was broken down into two stages. The first stage is a feature detector stage and the second is a MLP classifier [5].

The task of the feature detector stage is to give a representation of the speech pressure waveform that makes the fundamental period more explicit than it was in the original speech. It was felt that a wide-band filterbank that approximated a wide-band spectrogram would be an appropriate choice. The filterbank was designed to ensure that the individual vocal fold closures could be resolved in its output. It comprised nine second order 10th Butterworth filters with 3dB points of 50-300Hz, 300-600Hz, 600-900Hz, 900-1200Hz, 1200-1600Hz, 1600-2000Hz, 2000-2400Hz, 2400-2800Hz and 2800-3300Hz. The outputs from the filters were half-wave rectified and then smoothed by means of a second order low-pass Butterworth filter.

FIGURE 1. The speech fundamental period estimation problem. With a cut-off frequency of 1kHz. Half-wave rectification was employed as opposed to full wave because the latter has the effect of doubling the periodicity of the filter outputs. The smoothed outputs were then down-sampled (from a 10kHz input sampling rate) to 2kHz. This decimation was carried out to reduce the computational load of the classifier.

The classifier used was a multi-layer perceptron. It has an advantage over a classical algorithm like the Bayes classifier for normal patterns [6] because it does not assume the form of the distributions of the pattern classes a priori. In recognition mode the MLP is computationally efficient, making it suitable for real-time hardware implementations. The input to the classifier consisted of 41 frames. The network used had 369 units inputs, two hidden layers of 20 units each and an output layer with 41 units. The classifier achieved a remarkable 99.8% correct classification rate over a given window of a sampled speech pressure waveform. The input data to the network generated by the filterbank is shown as a grey-level display C. It can be seen that temporal variation concerning the excitation is retained. Item D shows the reference Tx markers generated from B. The Tx markers are used to train the output of the MLP network.

ORIGINAL SYSTEM RESULTS

It was trained on five male speakers using anechoic speech contaminated with background noise from a canter. For further information concerning the original system, see reference [2]. Figure [4] shows a typical output of the trained MLP network for the two noise conditions. It can be seen that the MLP achieves a remarkable similarity to the target Tx on which it was trained. It is to be noted that the performance is not substantially degraded in the 0dB SNR case. The epoch marker locations due to the peak-picker are also shown. The peak-picker is a very simple time-domain device. The version used here for comparisons is a software model of a small...
battery powered device developed as part of the EPI group cochlear implant prosthesis [8]. It can be seen that it experiences considerable difficulty in the 0dB SNR case.

DESIGN OF REAL-TIME SYSTEM

For a real-time system, it was necessary to develop a new reduced computation version of MLP-TX that could run in real-time on currently available DSP processors. The real-time implementation of such an algorithm requires consideration of the amount of processing that can be performed. The computation required by the MLP-TX algorithm is determined by the number of filter channels, the order of the filters, and the number of nodes and interconnections in the MLP classifier. The system variables must fit into RAM and the filter coefficients, network weights, look-up tables and program code must fit into EPROM.

The decision was made on basis of what a high performance DSP processor might be able to offer. The biggest system that could be implemented on such a processor was estimated, assuming that a single cycle multiply accumulate instruction could be executed in 10ns, that a new filter channel was initiated that filters the speech into 6 frequency bands. It comprised six second order IIR Butterworth filters with -3dB points of 50-500Hz, 300-600Hz, 600-900Hz, 900-1200Hz, 1200-2000Hz, and 2000-3000Hz. The outputs from the filters were again half-wave rectified and then smoothed by means of a second order low-pass Butterworth filter with a cut-off frequency of 1kHz. This time the network employed used 41 input frames, that is (41*6) input units, two layers of 6 hidden units, and one output unit. The schematic for the real-time system appears in figure [5].

In a practical implementation of the MLP-TX algorithm the arithmetic can be carried out by means of integer multiplications. The use of a look-up table to calculate the sigmoid non-linearity and log compression is essential. Simulations were carried out to investigate the effects of such operations. This involved quantizing the input data and the weights to a specified number of levels. Additionally, the use of a look-up table was investigated. It was found that 8 bit representation for the weights and a look-up table of 2K entries for the sigmoid gave no noticeable reduction in recognition performance.

These decisions were all made before the processor was selected and its code written.

The selected processor had to have the following features:

1) Single cycle multiply-accumulate data-move in a loop instruction.
2) Integer processor, because floating-point operation uses at least twice the power and it is unnecessary.
3) Must be CMOS for as low as possible power consumption.
4) Preferably as much on-chip memory as possible.

The TMS320C25 processor was finally chosen. A 12 bit, 10 microsecond A/D converter is used to digitize the output from the microphone.

The algorithm for the reduced MLP-TX was written for the TMS320C25 in assembly language. The computation is interleaved with the filterbank running every sample and the MLP every 4 samples. The system was initially implemented using a Longborough Sound Images development board in a PC/AT computer. The final system consists of the TMS320C25 with on-chip RAM, EPROM, A/D and support circuits on a small printed circuit board measuring 7cm by 6 cm. A reduced clock frequency of 32MHz both to reduce power consumption and to be compatible with the timing requirements of the 55ns EPROM WSI572C25-5S.

The new reduced MLP-TX was initially trained on a Masscomp 5600 system and the weights obtained from training were then downloaded into the EPROM in the small portable system. At present only initial training has been carried out, and further training on a wider range of data is being undertaken to give full speaker independent performance over a wide range of environmental conditions.

The final system is portable and runs in real-time with a power consumption of 400mW. This will allow the complete pocket-sized hearing aid to run for at least 12 hours powered by a rechargeable lithium battery.

IMPROVEMENTS

Further developments of the basic algorithm have also been carried out. These are at present implemented on the Masscomp mini-computer system, but will be implemented using the TMS320C25 at a later stage.

The use of a non-symmetrical window on the speech that uses a lot of past evidence, but little future evidence, has been investigated. Such a scheme retains the high performance of the original system, but with a shorter delay between input and output. In addition, fundamental period doubling errors for low frequency speech are reduced. This is because an asymmetrical window is able to include evidence of an excitation point input window for lower fundamental frequencies than a comparable symmetric one.

The MLP-TX algorithm is ideally required to operate well over a wide range of speakers, utterance, environmental conditions, etc. It is clear that if information from over a wide time window is available in the classification process that the performance of the classification process could increase. However if the dimensionality of the input vector increases, the classifier will have correspondingly more degrees of freedom. Consequently a stage will be reached where there is no practical benefit from increased window size, because this will result in a classifier that cannot be trained. It is believed that its task may be facilitated by the inclusion of some parameters in the input vector that give some indication of these conditions. The use of a slowly varying context in addition to the original input vector has been investigated. Because of the slowly varying nature of this input, incorporating the context will not drastically increase the pattern dimensionality, but it will provide information over a larger time window.

This has been tried using 3 frames of vocoder data [9] as context to the wide-band filterbank data, and the results obtained using such an approach are encouraging.

One problem with the original design for MLP-TX is that the location of vocal fold closure in time is too imprecise for many applications. Ideally one would like higher resolution than 0.5ms. This is rather difficult to achieve with the old system because simply increasing the frame sampling frequency it would again result in a massive network that would be difficult to train. However by using a high resolution central zone in the input pattern vector, this problem can be largely avoided. This was done by using a 1kHz frame rate in the central region and a 2kHz frame-rate for the remainder of the input vector. In this way the lower frame-rate vectors can provide a context for the higher frame-rate vectors. Such an approach increases the resolution from 0.5ms to 0.125ms.

Tools to permit the separate training of parts of networks that are then joined together in the final system, are being developed. It is expected that this approach will enable more sophisticated system to be produced with faster training times.

ACKNOWLEDGEMENTS

This work was supported by RSRE and by the MRC. The MLP software for the Masscomp computer was written in conjunction with Mark Huckvale. The software version of the peak-picker was implemented by D. M. Howard.

REFERENCES


---

**FIGURE 2.** Schematic for original MLP-TX system.

**FIGURE 3.**

---

EUROSPEECH '89, Paris, France, September 1989
FIGURE 4. Portable real-time MLP-Tx

41 stage delay

FIGURE 5. Portable real-time MLP-Tx