Speech Communication 9 (1990) 63–71
North-Holland

REAL-TIME PORTABLE MULTI-LAYER PERCEPTRON VOICE FUNDAMENTAL-PERIOD EXTRACTOR FOR HEARING AIDS AND COCHLEAR IMPLANTS

J.R. WALLIKER

Department of Phonetics and Linguistics, University College London, UK and Department of Clinical Physics and Bioengineering, Guy's Hospital, London, UK

and

I. HOWARD

Department of Phonetics and Linguistics, University College London, UK

Received 30 October 1989

Abstract. Previous work has shown that it is possible to train a multi-layer perceptron to estimate the voice fundamental period (Tx) for multiple speakers in the presence of high levels of background noise. The algorithm has been implemented in real-time on a TMS320C25 based development system. A prototype pocket-sized portable device has been constructed and the real-time software transferred to it. This will provide the basis for a new generation of signal processing hearing aids for the profoundly and totally deaf. Power supply current is sufficiently low for battery operation for periods of 12 hours between charges. The basic algorithm has been adapted to provide the higher time resolution which will make it applicable to a wide range of other applications.

Zusammenfassung. Wie frühere Arbeiten gezeigt haben, ist es möglich ein mehrschichtiges Perzeptron darauf zu trainieren, die Grundperiode von Sprachsignalen sprecherunabhängig bei hohen Geräuschpegeln zu bestimmen. Der Algorithmus wurde in Echtzeit auf einem Signalprozessorsystem TMS320C25 entwickelt. Diese Software wurde auf einen tragbaren Prototypen in Taschenformat portiert. Dieser Entwurf liefert die Basis für eine neue Generation signalverarbeitender Hörhilfen für schwerst hörgeschädigte und völlig taube Patienten. Der Energiebedarf ist hinreichend klein, um einen 12-stündigen Akkubereich zu ermöglichen. Der Grundalgorithmus wurde dahingehend erweitert, eine bessere Zeitauflösung herzustellen, damit er für zahlreiche andere Anwendungsgebiete eingesetzt werden kann.

Résumé. Travaux antérieurs ont démontré la possibilité d'entraîner un perceptron à niveaux multiples pour l'estimation de la période fondamentale (Tx) de signaux produits par différents locuteurs parlant dans des conditions de bruit intense. L'algorithme a été implanté en temps réel et réalisé sur un système de développement TMS320C25. Un prototype portatif incorporant le logiciel a été construit. Il servira de base à une nouvelle génération de prothèses auditives à traitement de signal pour les sourds profonds. Grâce à une très faible consommation, l'autonomie de l'appareil est de 12 heures sans recharge. L'algorithme a été amélioré afin d'augmenter sa résolution temporelle ce qui le rendra intéressant pour une large série d'autres applications.

Keywords. Fundamental frequency, multi-layer perceptron, real-time, cochlear implant, hearing aid.

1. Introduction

The purpose of this work was to develop the basis for a new generation of speech feature extracting hearing aids for the profoundly and totally deaf. The External Pattern Input (EPI) group has already developed first generation devices which use a combination of analogue and digital signal processing techniques (Walliker et al., 1986; Rosen et al., 1988).

These devices are intended for two classes of patient:

(1) Those who have a severe audiometric loss combined with poor frequency selectivity. Conventional hearing aids which provide amplification, compression and filtering are

often of limited use to them. We have shown that it can be beneficial to simplify the speech signal and present the most important components to the deaf lipreader in a form which is matched to the needs and receptive capabilities of the damaged ear.

(2) The totally deaf who cannot respond to any acoustic input can often be given a sensation of sound by electrical stimulation of the auditory nerve endings in the cochlea. The same principles of speech pattern extraction and synthesis of matched stimuli are used, but the detailed current waveform which needs to be applied to the electrodes is different.

Our existing battery powered portable devices extract the voice fundamental frequency with an analogue peak picking circuit (Howard and Fourcin, 1983). This drives a microprocessor-based waveform synthesiser which carries out amplitude and frequency mapping matched individually to the needs of the particular patient before generating the stimulus waveform.

Voice fundamental frequency is the most important component of speech which is inaccessible to the deaf lipreader. Others, such as frication, amplitude envelope and formant frequencies could also be detected and usefully presented to some patients. However, the ability of the severely hearing-impaired listener to detect small frequency changes is generally much poorer than that of normal subjects and degrades rapidly with increasing frequency. A problem with our existing systems has been the degradation in performance of the fundamental frequency extractor in background noise and reverberation. This difficulty can be overcome in most applications by the use of noise cancelling microphones placed very close to the speaker. This option is not conveniently available to the deaf listener, however, where distances of several metres are typical of normal conversation. It is also important to detect voicing on a period-by-period basis, as otherwise the deaf speaker will not be aware of irregularity in his voice. Processing time delays must be minimised when the extracted signal is used in conjunction with lipreading.

Consequently, the requirements for a fundamental frequency extractor to be used in this application are somewhat different from those in other areas of speech technology, but the techniques described here are being adapted to meet the differing needs of other applications.

2. Initial developments

It has been shown that a multilayer perceptron (MLP) can be trained to detect voice fundamental period in adverse noise conditions (Howard and Huckvale, 1988). Further development, however, was needed to adapt the original algorithm to portable real-time operation. This was undertaken in the following stages; (1) investigation of the use of integer weights and calculation to allow the use of a lower power processor, (2) investigation of the use of lookup tables to replace the non-linear compression and output functions, (3) the simplification of the algorithm to a point where existing hardware could carry out the required computation and demonstration that it could be trained, (4) implementation of real-time code on a development system, and (5) design and construction of low-power hardware and transfer of program code.

3. Original network configuration

The objective was to detect, from a voiced speech waveform, the moments at which the speaker's vocal folds snap together and excite the vocal tract. Figure 1 shows the speech waveform and the desired output under good conditions. There is a clear peak in the acoustic waveform which can be detected by relatively simple means. When the waveform is contaminated by noise and reverberation, however, this task becomes very difficult.

The solution was to filter the speech signal with a bank of 9 second order broad-band filters (300 Hz) which do not unduly smear the temporal information. The outputs are then rectified and low-pass filtered with second order Butterworth filters to reduce aliasing in the following stage where they are down-sampled to a 2 kHz rate and logarithmically compressed. Half-wave rectification, with the positive pressure component being retained, was used to avoid doubling of the output

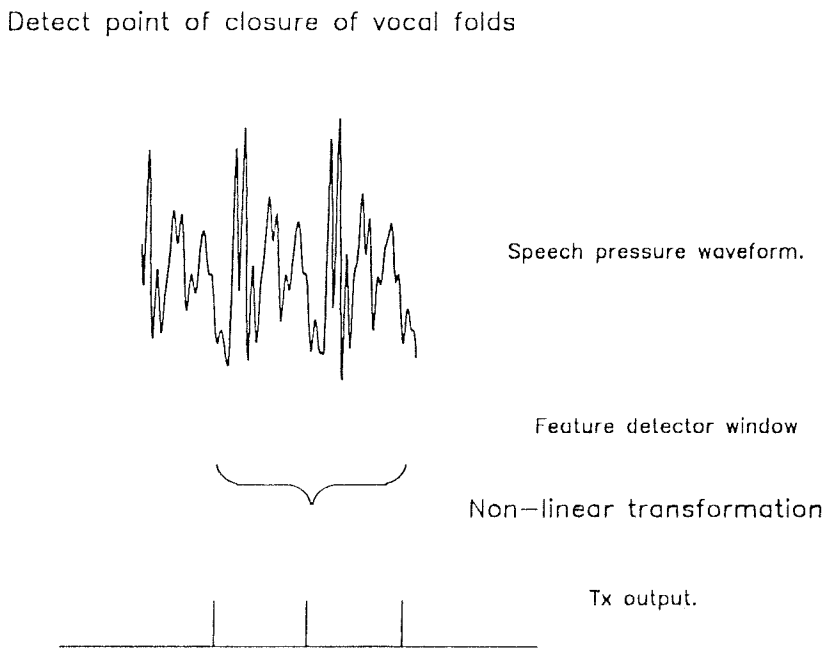

Fig. 1. Speech pressure waveform and idealised *Tx* output.

frequency. The outputs drive 9 delay lines of 41 stages each which form the input vector for the MLP.

The MLP has 2 layers of 10 hidden units and one output unit, with full interconnection between adjacent layers. The delay lines and the whole network are clocked at 2 kHz. This configuration was obtained by trial and error, and is probably not optimal (Howard and Huckvale, 1988). The choice of clock frequency was a compromise between temporal resolution and computational load. It corresponds to a frequency quantisation of +/−2.5% at 100 Hz. Because our profoundly deaf patients have much poorer frequency discrimination abilities than normal listeners they will barely be able to detect this degree of quantisation. It was found that a network with two hidden layers was much easier to train and gave better results than a one hidden layer network. The input time window needs to be at least 20 ms long so that it spans the period between successive voice periods in low-pitched male speakers. Figure 2 shows the configuration of the filterbank and MLP used for the initial tests.

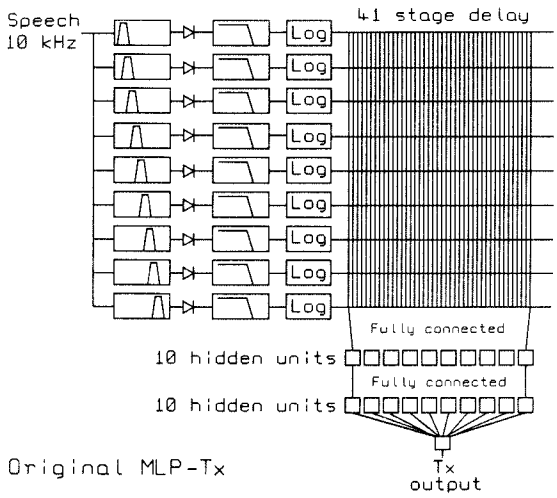

Fig. 2. Original filterbank and MLP configuration. This is too complex to operate in real-time with current low-power technology.

4. Network training

Large amounts of accurately annotated and noise contaminated training data were needed.

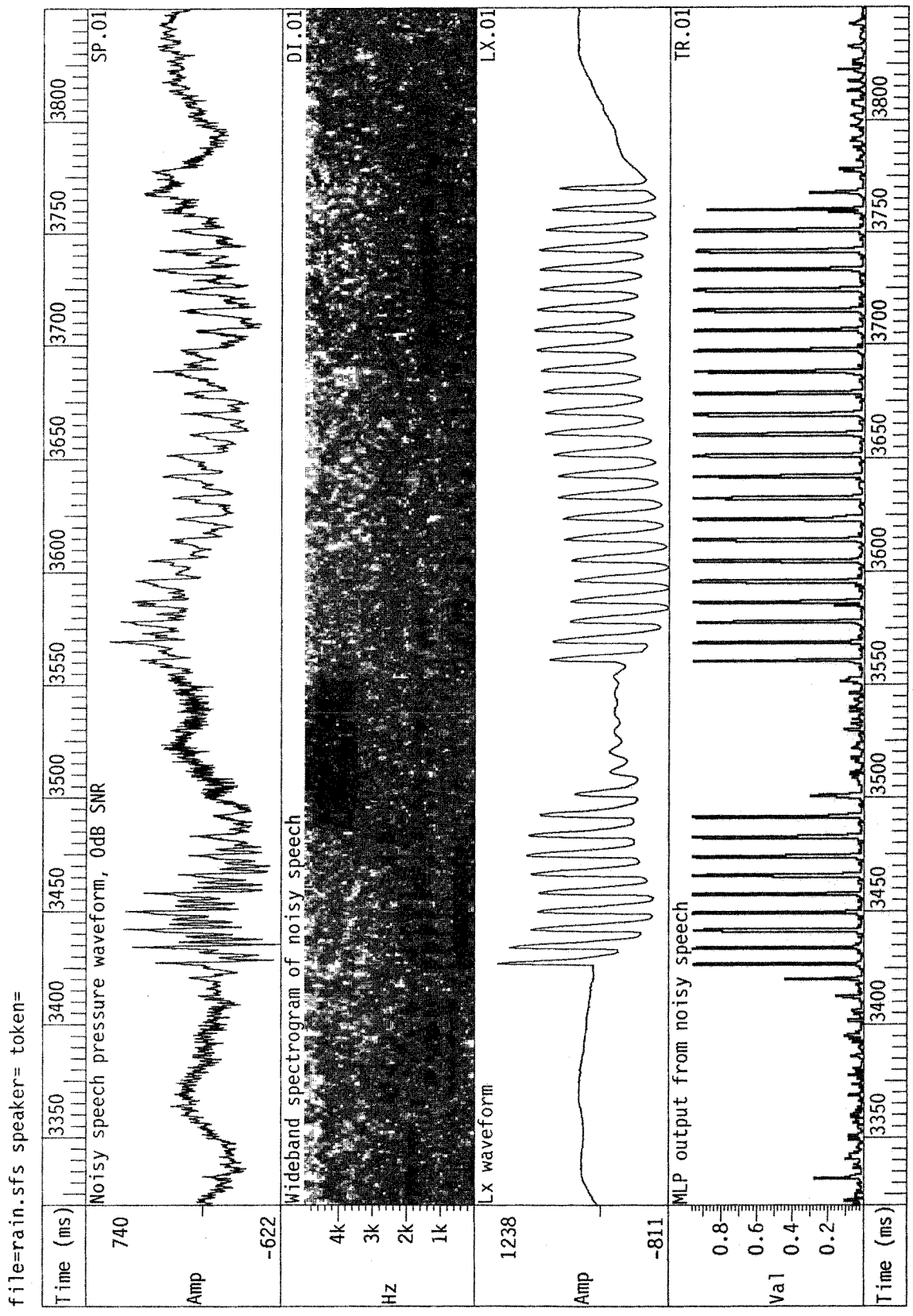

Fig. 3. Example of 0 dB noise contaminated speech with wide-band spectrogram, Laryngograph waveform and MLP output. The rising edges of the *Lx* waveform indicate vocal fold closures which are used for training and evaluation.

This was obtained by making recordings of the "Rainbow" passage (Mermelstein, 1977) in an anechoic chamber at 15 cm microphone distance. A Laryngograph® was simultaneously used to obtain the timing of vocal fold closures (Fourcin and Abberton, 1971). Five male speakers were used, and two recordings made from each. The first was used for training, and the second for evaluation. Speech and laryngeal data were digitised at 10 kHz. An automatic annotation program derived the vocal fold closure points from the Laryngograph channel. A noise signal consisting of multiple conversations and impulsive noise was recorded in the College Refectory at lunchtime and mixed in 0 dB and −20 dB ratios with the clean speech. Two networks were trained by back propagation on a Masscomp MC5600, using the UCL Pattern Recognition Workbench software, one with each noise ratio. Figure 3 shows sample noisy speech waveforms with the MLP output at 0 dB signal to noise.

The evaluation criterion used for gauging the performance of the algorithm in different situations was the equal error point of the receiver operating characteristic for the device. This was used because it is relatively easy to determine automatically and provides a good indication of system performance (Levine and Shefner, 1981).

5. Effect of quantisation of weights

It is highly desirable to avoid the use of floating point arithmetic if at all possible because hardware which supports it typically draws several times more supply current than that for integer computation. Furthermore, the integer precision should be no greater than necessary because the complexity, and hence power consumption of multipliers (which dominate the computational load) increases roughly as the square of the word size while the power consumption of memory is in proportion to the word size.

Training of a network with highly quantised weights has been shown to cause problems (Chong and Fallside, 1988). We therefore investigated the effect of quantising the weights *after* training in the usual way with 32 bit floating point hardware. Table 1 shows that down to 16 levels there is no measurable degradation at 20 dB signal to noise ratio. We therefore specified a minimum of 8 bit representation for the design of real-time hardware, because this provides a reasonable safety margin and is convenient to implement on most processors.

Table 1
Quantisation of weights. Speech at 20 dB s/n

Quantisation	Equal error hit rate
None	92.1%
None, 0 dB s/n	87.9%
3 levels	0.0%
4 levels	74.1%
6 levels	91.5%
8 levels	88.8%
16 levels	93.5%
32 levels	92.0%
64 levels	92.1%
128 levels	92.1%
256 levels	92.2%
512 levels	92.1%
1024 levels	92.1%
2048 levels	92.1%

6. Effect of non-linear lookup tables

For an MLP to be useful, it is essential to apply a non-linear output function to each unit. A commonly used sigmoid function is:

$$\text{output} = (1 + e^{-\Sigma \text{ weighted inputs}})^{-1}. \quad (1)$$

Division and exponentiation are both very slow operations on most processors and would prevent real-time operation. We therefore generated a set of look-up tables by computing a range of values for the function and quantising them to varying numbers of levels. A series of simulations were run to determine the degradation in performance with these tables substituted for the original sigmoid function. Values outside the range of the tables were clipped. Table 2 shows that a lookup table with only 16 levels can provide results as good as the original sigmoid function. However, the effect of table size seems also to depend on the complexity of the network in relation to the number of speakers for which it has been trained.

Table 2
Quantisation of sigmoid table. Speech at 20 dB s/n, 8 bit weights

Quantisation	Equal error hit rate
3 entries	84.9%
4 entries	83.1%
6 entries	83.7%
8 entries	92.8%
16 entries	91.7%
32 entries	91.6%
64 entries	92.0%
128 entries	92.0%
256 entries	92.2%
512 entries	92.0%
1024 entries	91.9%
2048 entries	91.9%

It is prudent, therefore, to allow a large safety margin, if sufficient memory is available.

7. Simplification of network for real-time operation

An estimate was made of the largest network which could run on a digital signal processing (DSP) chip together with the filterbank. The number of filter bands was reduced by combining the highest channels because these were considered to be proportionally rather narrow. It was necessary to reduce the hidden units in proportion. This led to a design with 6 filter channels, 6 first layer hidden units and 6 second layer hidden units which is shown in Figure 4. The input sampling frequency was reduced to 8 kHz. Initial tests which were carried out with speech recorded in a noisy computer room indicated similar performance to the original network.

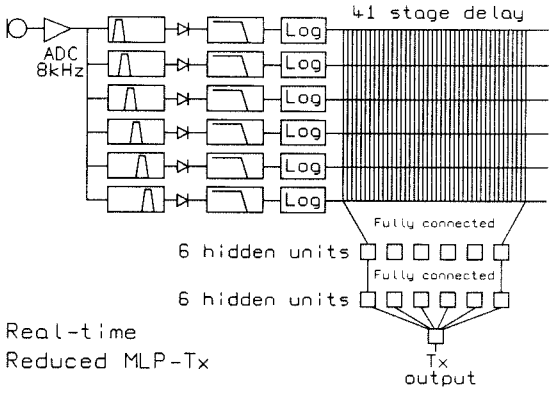

Fig. 4. Simplified network for real-time operation.

8. Choice of processor

There were many conflicting factors to balance when selecting a processor on which to implement the MLP algorithm. Firstly, there must be sufficient computing power. The estimate described above assumed a device able to carry out a multiply accumulate and data move operation (MACD) in a single processor cycle of 100 ns or less. Because the application required a pocketable device, able to run for 12 hours before recharging the batteries, power consumption and system size were dominant considerations. Low power high speed CMOS technology was essential for the active components.

The experiments on quantisation of weights and lookup tables had shown that a word size of 8 bits would suffice for the MLP itself, so long as an extended accumulator and saturation arithmetic were available. The filterbank software, however, needed 16 bit operation to obtain an adequate dynamic range and stable operation of the IIR filters. The minimisation of external components ("glue" logic and memory) was important, as was the quality of technical support from the manufacturer and the prospect of future devices which might match the application better. The Texas Instruments TMS320C25, Motorola DSP56001 and several other European, American and Japanese processors were carefully considered, but the Texas device was finally chosen.

9. Real-time software implementation

The choice of the TMS320C25 dictated the organisation of the program code. The dominant operation in executing an MLP is multiplication of the input vector of each unit by the associated table of weights and summing the results. This can only be done efficiently if variable data is placed in "data" memory on the processor chip and weights are stored in fast "program" memory

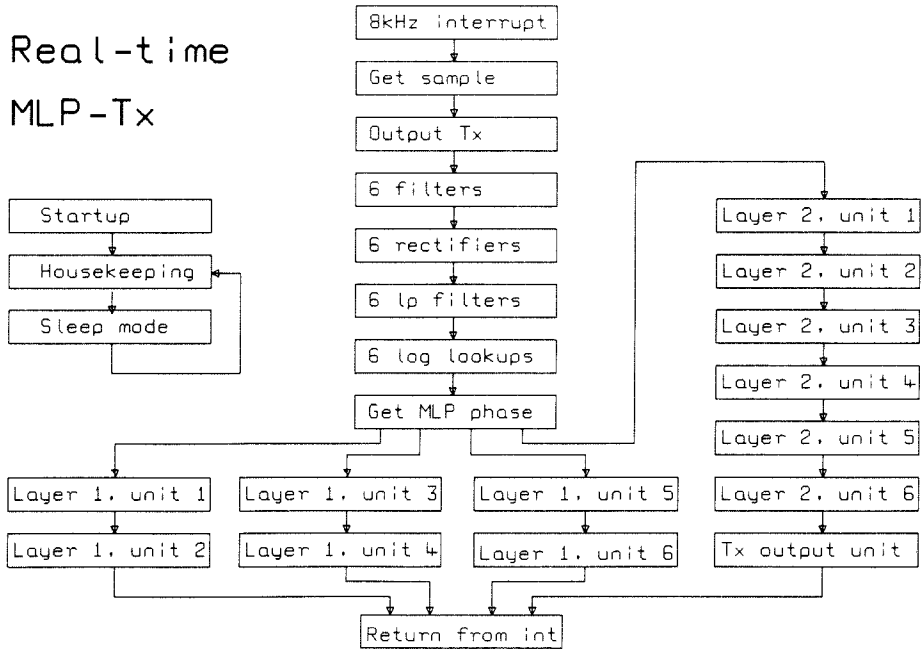

Fig. 5. Software flow diagram for real-time system.

which may be an external eprom. Under these conditions, a single multiply accumulate instruction can be automatically repeated so that the multiplier operates at its full bandwidth. Furthermore, it is possible to copy data words to adjacent memory locations in parallel with the MAC operation thereby allowing the delay line to be implemented with zero overhead.

Analogue samples are acquired and fed directly into the filterbank software every 125 μs. The MLP code is interleaved in the remaining time before the next sample. Its execution is broken up into four phases, three of which contain two input units each, while the third contains all the hidden units and the output unit. Figure 5 shows this partitioning in detail.

The software was developed on a Loughborough Sound Images development board in an IBM PC/AT compatible computer, using the Texas COFF format assembler and linker. Debugging presented some difficulty, mainly because the algorithm was so robust that even quite serious errors did not completely prevent it from operating. A display program was written which captured internal data from the development card while it was running the MLP in real-time and graphically displayed the outputs of the filterbanks and of each unit. This helped considerably in locating problems.

10. Hardware implementation

The key to a compact, low power implementation was the use of a fast, low-power CMOS EPROM for the storage of the program, weights and lookup tables. The device chosen was a Wafer Scale Integration WSI57C257-55 which contains 16 k words of 16 bit data with an access time of 55 ns and a maximum supply current of 40 mA. It was possible to interface the EPROM and a 12 bit analogue-to-digital converter with very little external "glue" logic, thereby saving power and space. Surface mounted components were used wherever possible (Fig. 6). The TMS320C25 was clocked at the reduced frequency of 32 MHz to allow zero wait state access to the EPROM. This slowed the execution time of MACD instructions to 125 ns, which was just sufficient to run both the filterbank and the re-

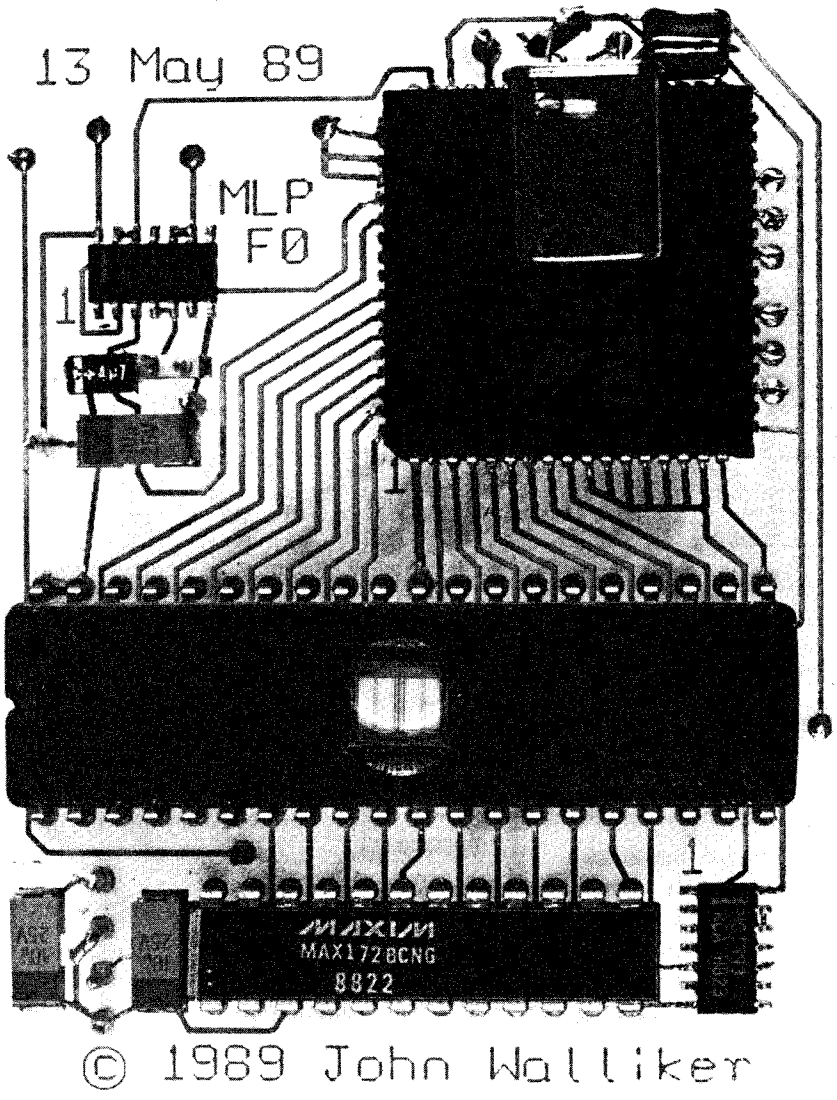

Fig. 6. Prototype of portable real-time hardware. The main components are (from top) TMS320C25 dsp chip with quartz crystal overlying it, WSI57C257 eprom, MAX172 12 bit analogue-to-digital converter. The circuit is 70 mm × 55 mm in size.

duced MLP in real-time at 8 kHz sampling frequency. The power supply current for the complete board was 80 mA @ 5 V. (It was found that this dropped to 60 mA @ 4 V and 50 mA @ 3.5 V. Below 3.5 V the clock oscillator stopped. Operation below 4.75 V, however, was not recommended by the respective manufacturers and could prove to be unreliable.)

11. Performance of real-time hardware

The current real-time implementation uses 16 bit arithmetic, 16 bit weights and lookup tables with an 11 bit input range. This degree of quantisation does not appear to degrade the classifier performance when compared with 32 bit floating simulations. Figure 7 shows some results from such a comparison.

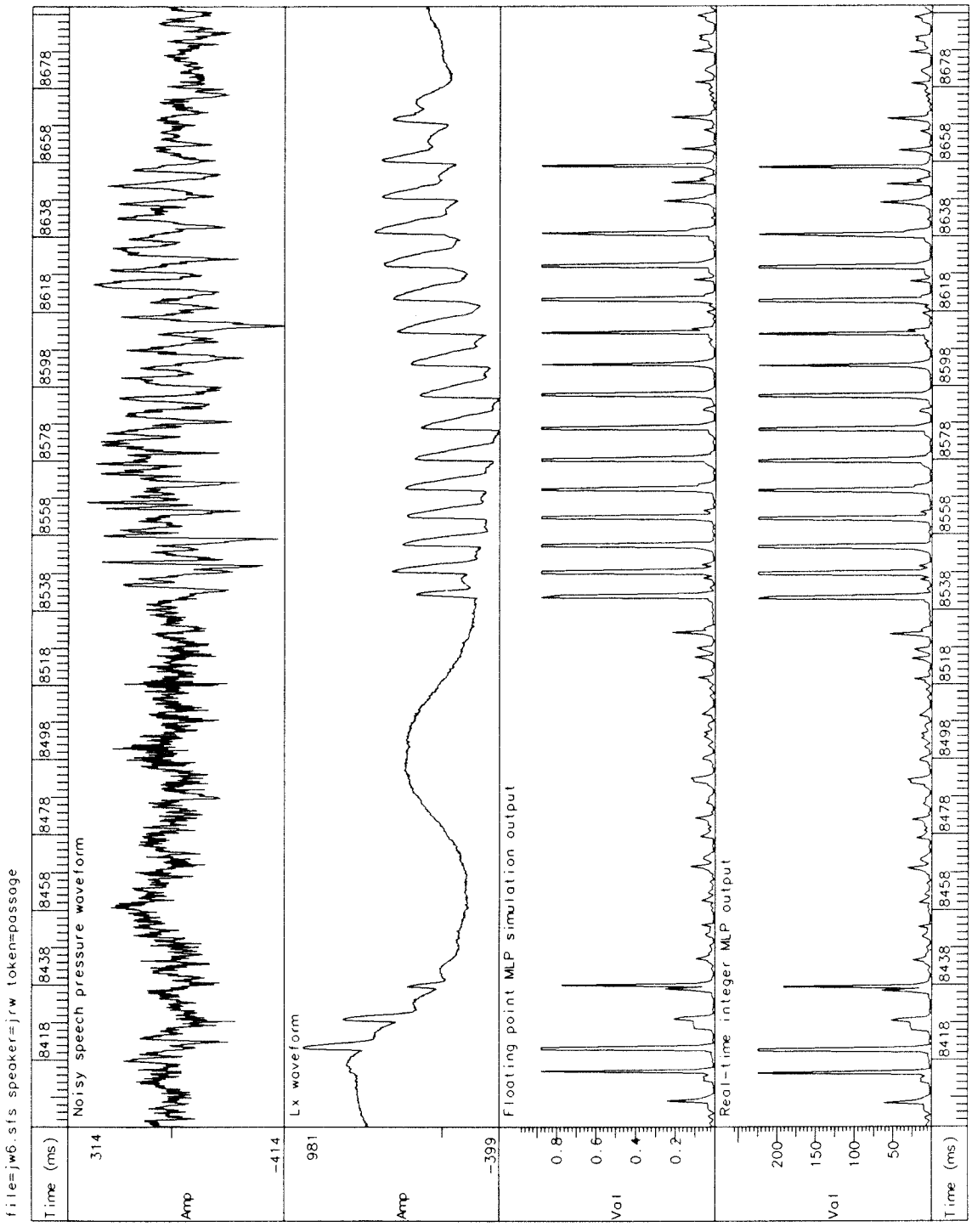

Fig. 7. Examples of noisy speech and Laryngograph signals together with results of floating point simulation and real-time hardware operation.

12. Conclusions

We have demonstrated that a multi-layer perceptron can be trained to extract voice fundamental period, and hence frequency, for multiple speakers in very noisy conditions in real time. The algorithm avoids the long processing delays and loss of fine structure associated with other techniques such as cepstral analysis. The MLP has been demonstrated running in real-time on battery powered portable hardware which will form the basis for a new range of hearing aids for the profoundly and totally deaf. By increasing the sampling rate and accepting higher power consumption we expect to improve the temporal resolution to a point where the real-time algorithm will be useful in a much wider range of speech applications.

Acknowledgements

The pattern recognition workbench software was written in conjunction with Dr. Mark Huckvale at University College London. We are grateful to Micro Call Ltd. for the loan of a WS6000 EPROM programmer. This work was supported by the Medical Research Council of the United Kingdom.

References

M. Chong and F. Fallside (1988), "Implementation of neural networks for speech recognition on a transputer array", Cambridge University Engineering Dept. CUED F-IN-SENG\TR8.

A.J. Fourcin and E. Abberton (1971), "First applications of a new laryngograph", *Med. Biol. Illust.*, Vol. 21, pp. 172–182.

D.M. Howard and A.J. Fourcin (1983), "Instantaneous voice period measurement for cochlear stimulation", *Elect. Lett.*, Vol. 19, pp. 776–778.

I. Howard and M.A. Huckvale (1988), "Speech fundamental period estimation using a trainable pattern classifier", FASE88.

M. Levine and J. Shefner (1981), "Fundamentals of Sensation and Perception" (Addison Wesley, Reading, MA).

P. Mermelstein (1977), "On detecting nasals in continuous speech", *J. Acoust. Soc. Am.*, Vol. 61, p. 581.

S. Rosen, J.R. Walliker, A. Fourcin and V. Ball (1988), "A microprocessor-based acoustic hearing aid for the profoundly impaired listener", *J. Rehab. Res. Dev.*, Vol. 24, pp. 239–260.

J.R. Walliker, S. Rosen and A. Fourcin (1986), "Speech pattern prostheses for the profoundly and totally deaf", *IEE Conf. Pub. No. 258, Int. Conf. Speech Input/Output, London, England*, pp. 194–199.