SPEECH HEARING and LANGUAGE

work in progress 1991

Draft copy 4/5/91

University College London
Department of Phonetics and Linguistics

SPEECH FUNDAMENTAL PERIOD ESTIMATION
USING PATTERN CLASSIFICATION

IAN HOWARD

1. BASIC ISSUES IN SPEECH FUNDAMENTAL FREQUENCY ESTIMATION

1.1 Introduction

1.1.1 Aims of the work

This work is concerned with the design and development of an algorithm (MLP-Tx) that can perform speech fundamental period estimation. The algorithm has been specifically developed to perform fundamental period estimation for a signal processing hearing aid designed by the external pattern input (EPI) group at UCL which seeks to provide from the acoustic speech signal an output which corresponds to that from the laryngograph (Walliker et al., 1986; Rosen et al., 1988).

1.1.2 Novel design

The novel aspect of the work stems from the fact that the task is formulated as a pattern recognition problem and the MLP-Tx algorithm, based on a multi-layer perceptron pattern classifier (MLP), was trained-by-example to perform the required task. The data that was used to train the algorithm was composed of speech that was semi-automatically labelled for fundamental period location using a pair of algorithms that made use of the output of a laryngograph, which was recorded simultaneously with the speech. The fundamental periods were delineated in terms of the closure of the vocal folds as a function of time, as defined by the location of the maximum positive differential in the output of the laryngograph.

1.1.3 Preliminary system

The first system configuration investigated were based on a pre-processing of the speech pressure waveform by a wide-band filterbank analyzer. A brief account of this work was originally reported in Howard & Huckvale (1988). This gave an input to the classifier which consisted of a set of adjacent time frames from the output of the filterbank. The output from the classifier was defined as being in one of two classes. In the first there is a period epoch marker at a given output frame, in the second there is not. This first classifier was trained to generate an output which signified the presence of a vocal fold closure at the centre of its input window.

1.1.4 Developments of the basic technique

Developments of this first approach were then explored. These were primarily directed towards methods for reducing the training time for the MLP, improving the time resolution of the fundamental period estimates and trying out different pre-processing stages. This included direct operation on the speech pressure waveform and the use of a simplified auditory filterbank.

1.1.5 Suitability for real-time operation

The MLP-Tx algorithm is suitable for real-time implementation because of the simple uniform structure of the MLP, and the inherently small (about 10ms) input to output delay. This small delay is important to prevent a lack of synchronization between the speech signal and gestures and lip movements made by a speaker. Methods to reduce the computation load required for practical implementation were examined and these resulted in a system using a low-order filterbank together with a smaller MLP network. The last configuration was of practical interest because it had a processing load small enough to be run in real-time on a portable DSP system. A practical real-time implementation for use in hearing aids for the profoundly hearing impaired has been carried out (Howard & Walliker 1989; Walliker & Howard, 1990).

1.1.6 Quantitative frequency contour comparisons

A number of objective assessment techniques were developed and used to permit quantitative comparisons between fundamental period estimation algorithms to be carried out. These involved quantitative comparisons between frequency contours. Using these comparisons, various different configurations of the MLP-Tx algorithm were evaluated over a wide range of speakers and environmental conditions. The performance of the MLP-Tx algorithm was also compared against that of several established fundamental frequency estimation algorithms, and its performance was found to be better than that obtainable by some established techniques.

1.1.7 Strengths of the MLP-Tx algorithm

One of the strengths of the MLP-Tx algorithm is that the fundamental period estimations are made on a cycle-by-cycle basis. Consequently irregularities in vocal fold vibration can be detected by the algorithm, whereas many other algorithms would tend to smooth the period values. Creaky voice can be dealt with effectively using the MLP-Tx algorithm, whereas many other algorithms treat this important larynx excitation as being unvoiced due to its intrinsic irregularity.

1.1.8 Organization of this paper

The first section provides a discussion of applications, problems and approaches to speech fundamental frequency estimation. The second section gives the background that lead to the development of the MLP-Tx algorithm and a preliminary set of experimental results. The third and last section then examines the issues involved in more detail. A more comprehensive set of experimental results are presented in which three different configurations of the MLP-Tx algorithm are compared against two established fundamental frequency algorithms. This report is very much a 'work in progress', and for a full account of the work in more detail with many further experimental results, the reader is referred to Howard (1991).

1.2 Applications of speech fundamental period estimation

1.2.1 Cochlear implants

There are many other applications for fundamental period estimation algorithms.

One application for algorithms that can perform speech fundamental period estimation is in signal processing hearing aids (Fourcin et al., 1983). Such hearing aids are of value to the profoundly deaf. The auditory systems of this class of patients has a very restricted channel capacity which is less than that required to encode adequately the whole speech signal. Consequently presentation of the whole speech signal maybe confusing and even painful, whereas a basic fundamental period signal can be made much easier for the patient to interpret usefully (Rosen & Fourcin, 1983). The elements used by the EPI group are chosen to be simplified aspects of speech that, when suitably coded, are matched to the residual discriminative abilities of the patients (Faulkner, Ball & Fourcin, 1990; Walliker et al., 1985). Speech fundamental frequency is a particularly useful feature to present to the profoundly deaf because it is almost completely invisible to the lip-reader and consequently it provides an aid to lipreading and the development of voiced speech production. There is particular benefit to be derived from the use of an algorithm that can perform fundamental period estimation on a cycle-by-cycle basis for such signal processing hearing aid applications. This is because preservation of the irregularity present in the original speech excitation is beneficial since the patient can then hear creaky and other irregular voice characteristics (Abberton et al., 1985). For these applications the algorithm must run in real-time with a processing delay between input and output should be as small as possible, and no more that a maximum of 40ms, or the signal losses its usefulness (McGrath & Summerfield, 1985). The MLP-Tx algorithm is therefore particularly suitable for such applications, because it possesses both of these qualities.

1.2.2 Speech Coding

Many coding schemes that aim to reduce the data-rate in the transmission of speech assume a source-filter model of speech production and require the determination of speech fundamental frequency (Fant, 1970). Such schemes can be considered as speech analysis/synthesis systems and include Dudley's vocoder (1939), the channel vocoder (Flanagan, 1972) and copy synthesis (Lawrence, 1953).

1.2.3 Speech and Speaker Recognition

Another application of information relating to speech excitation is in automatic speech recognition. The incorporation of such information into the recognition process has been demonstrated as beneficial to the recognition task (Atal, 1974; Rosenberg & Sambur, 1975). In addition, fundamental period information has been shown to be useful in speaker recognition (Atal, 1972; Abberton, 1974; Abberton, 1976). Abberton also showed that transformations similar to the "mapitch" algorithm of the SIVO aids preserved important speaker identity information.

1.2.4 Glottal-synchronous speech analysis

The individual identification of speech fundamental period is also useful in providing a means to carry out glottal-synchronous analysis of speech. The idea behind this technique is that when, for example, performing a short-time spectral analysis of the speech, the window for the analysis is selected on a period-by-

period basis to include results from only one period, rather than using a fixed window size for the analyses. It has been found that such an approach can give better estimates of the vocal tract transfer function than using fixed window analysis (Hunt & Harvenberg, 1986; Pearce & Whitaker, 1986; Hunt, Zwierzynski & Carr, 1989). To carry out this task requires the identification of the vocal fold closure points, and this is the function of the MLP-Tx algorithm.

1.3 Problems in speech fundamental frequency estimation

1.3.1 General difficulties is speech fundamental frequency estimation

The determination of speech fundamental frequency is a difficult problem for many reasons. Speech is a non-stationary signal. One reason for this is due to the fact that the shape of the vocal tract can change rapidly so that it can become acoustically very different even within the space of a single fundamental period. In addition, the vocal tract can give rise to a wide variety of speech sound, with a multitude of different temporal structures. The glottal excitation of the vocal tract is often only quasi-periodic. This is particularly true in the case of creaky voice. In addition there are acoustic interactions between the excitation from the vocal folds and the vocal tract. Problems can also arise in when there is a strong first formant in the vicinity of the second harmonic. This can lead to "doubling" errors, because this leads to a significant second peak in each fundamental period. The range of the fundamental frequency of speech is quite large, an it can extend from below 50Hz for some male subject, to above 800Hz for women and children. The problems of speech fundamental frequency estimation become more difficult when there is additive noise and distortion in the communication channel.

1.3.2 Required measurement resolution and accuracy

The accuracy and resolution requirements for a fundamental frequency algorithm are determined by its intended applications. The human auditory system is more sensitive to changes in absolute frequency at low frequencies, and in general the noticeable difference in frequency is proportional to frequency. The difference limen with respect to the fundamental frequency for human listeners perhaps represents the ultimate required performance, which is typically 0.3-0.5% resolution of the fundamental frequency. Most algorithms do not meet this specification, although for most applications less accuracy can be tolerated. The frequency difference limens for the profoundly deaf patients, for whom high technology signal processing hearing aids are intended, is only 3-4% of the F_0 values, which is ten times worse than for normal listeners. For this application it is possible to use a much lower time resolution than that needed in other applications.

When sampled data is used to represent the speech pressure waveform, there is a limit to the resolution in time to which any feature in the speech cycle can be located, and this is determined by the sampling period. For example, at a sampling frequency of 10KHz, it is only possible to locate a time event to 1/10000 = 100 microseconds. For a fundamental frequency of 100Hz, this corresponds to an accuracy of 1%. At higher fundamental frequencies, this percentage error increases still further. With regard to this accuracy issue, Hess and Indefry point

out (1987) that to reduce sampling accuracies to 0.5% up to the fundamental frequency of 500Hz requires a sampling period of 10 microseconds.

1.4 Approaches to speech fundamental frequency/period estimation

1.4.1 Types of algorithm

The techniques that have been developed to determine speech fundamental frequency are broadly classified into four main groups by Hess (1983); Those that operate in the time-domain, those that operate over some short-term window of the speech, which he calls short-term analysis, those which are hybrids of the first two, and finally those that operate by direct measurement of the action of the larynx.

1.4.2 Time domain algorithms

Time-domain algorithms employ direct measurements on the speech signal and involve looking for temporal features in the speech pressure waveform, such as local maxima and minima, and gives rise to an output signal that consists of a series of excitation markers that delineate period boundaries. The main strength of time domain speech fundamental period estimation is its ability to make cycle-by-cycle measurements. This enables such algorithms to deal satisfactorily with irregular voice qualities (such as creaky voice) and with rapid changes in the frequency of vocal fold vibration. However, the price to be paid is that such algorithms can be more sensitive to noise. There are two main reasons for this. Firstly, the key feature used by a typical algorithm, such as a signal peak, may be more affected by noise than the gross waveform shape. Secondly the measurement is not averaged over several cycles, as is the case with short-term analyses.

1.4.3 Short-term analysis algorithms

Short-term analysis procedures use some form of transformation of the data within a short (for example, 20ms-50ms) time window. The nature of the transformation depends on the particular method used. As a result of the way in which evidence is combined over the observation window, short-term approaches give rise to fundamental frequency estimates that corresponds to the average value over the analysis window. Therefore short-term analysis techniques are unable to perform estimation on a period-by-period basis, and they generally do not make use of phase information. This does have the advantage that they are not sensitive to phase distortions that may adversely affect simple time-domain techniques. In addition, because of the fact that they make use of evidence from all the data within the input window, such techniques are generally robust in the presence of unwanted noise and signal corruption. Because the function of short-term algorithm is essentially to detect the periodicity of the input signal, rather than discrete events signifying excitation points, this class of algorithms run into difficulties when the signal contains irregular periods.

1.4.4 Frequency domain algorithms

Frequency domain algorithms are a sub-class of short-time analyzers that operate on some kind of spectral representation of the signal. They can take advantage

of the harmonic structure of the excitation, and in some algorithms the fundamental does not even have to be present for such schemes to function. A valuable feature of frequency domain analyzers is that the resolution of their frequency estimates can be increased relatively easily by means of interpolation.

1.4.5 Laryngeal measurements

Another approach to speech fundamental period estimation involves the measurement of larynx activity. A device of particular value in the analysis of voiced speech excitation is the laryngograph (Fourcin & Abberton, 1971). A laryngograph operates by measuring the conductance across the larynx at the level of the vocal folds. This is achieved by placing two electrodes across the larynx with a small alternating voltage at several MHz across them. Movement of the vocal folds resulting in their closure causes a change in the electrical conductance, and consequently the current, which is demodulated to recover this fluctuation. The output waveform from the laryngograph thus gives a direct measure of vocal fold activity and is temporally much simpler than the corresponding speech pressure waveform. The point of closure of the vocal folds, which gives rise to the main peak in excitation, can be easily determined from the laryngograph waveform. The manifestation of the closure of the vocal folds in the laryngograph output signal is well agreed upon (Fourcin, 1974). The point of closure is usually taken as the point of maximum gradient in the closing phase of the laryngograph signal. This point of closure of the vocal folds, which gives rise to the main peak in excitation, can be easily determined from the laryngograph output waveform. Therefore, by means of a relatively simple time domain fundamental frequency estimation algorithm, a good estimate of speech excitation markers can be obtained. One big advantage of the laryngographic techniques is that it is easy to perform period-by-period estimation and consequently averaging of fundamental period values over time is avoided.

1.5 Reference fundamental period estimation using the laryngograph

1.5.1 Need for a reference algorithm

For the purposes of assessment of other speech fundamental frequency analysis algorithms (and for the **training** of the MLP-Tx algorithm) it is necessary to have a reliable standard. Hess argues that the laryngograph can form the basis of an ideal instrument for speech fundamental period estimation because it is robust, reliable, does not interfere with articulation and is essentially immune to environmental noise (Hess,1983; Hess & Indefry, 1984).

1.5.2 Automatic fundamental period estimation algorithm using the laryngograph

The fundamental period epoch markers used in this work (as a reference for the training and testing of the MLP-Tx algorithm) were obtained using a system that uses the laryngograph signal. This involved two separate algorithms that were implemented as two programs. The first consisted of an initial automatic analysis that generated a first approximation to the period markers. The second was an interactive program which provided the user a means to alter and correct the first set of period markers.

The automatic reference period marker estimation algorithm defines the output in terms of the point of maximum positive gradient in the laryngograph waveform (Hess & Indefry, 1984). Figure 1 shows an example of the laryngograph signal and its corresponding differenced signal for a typical segment. An expert system was used to locate these points and then checks the validity of the initial estimates. The various processing stages of the first program are now described in detail.

Stage 1:
The laryngograph signal is read in from a SFS file (Huckvale, 1988) and then bandpass filtered between 40Hz and 3KHz with a 251 point linear phase FIR filter, and the delay is then corrected. This removes unwanted low-frequency as well as high frequency noise from the signal.

Stage 2:
The bandpass filtered laryngograph signal is then differenced. The laryngograph signal is then checked for correct polarity by separately summing the values of all the positive and negative peaks in the differenced laryngograph signal. If the sum of positive peaks exceed the sum of negative peaks, then the laryngograph signal is assumed to be of correct polarity; otherwise it is inverted. The correct polarity bandpass filtered signal and the differenced signal are then retained for future analysis. In addition, both these signals are written back to the SFS file to permit further analysis using the interactive program.

Stage 3:
It has been observed that there is a tendency for the laryngograph waveform to die away rapidly towards the end of voiced segments (Howard & Lindsey, 1988). Under these circumstances, it is very difficult to detect the excitation points, because the peaks in the differenced laryngograph signal become very small and comparable in size to the background noise. As a consequence to this, the threshold that best detects the period markers is as low as the background noise on the laryngograph waveform will permit. Therefore, it is the characteristics of the noise that determines how low the threshold can be set.

The level of the background noise in the differenced laryngograph signal is estimated by consideration of unvoiced regions of the laryngograph signal that are previously labelled by hand. The mean μ_{dlx} and standard deviation σ_{dlx} of the differenced laryngograph signal within the unvoiced regions are then estimated, giving a simple statistical description of the background noise for the unvoiced conditions. The use of the mean and the standard deviation is based upon the assumption that the noise is Gaussian and un-correlated. In this case, one can calculate the probability that the signal will exceed a given value in terms of the signal mean and standard deviation. The probability that the noise will exceed a given threshold corresponds to the area under the Gaussian curve for values greater than the threshold. Ideally it is required that the threshold used should give no false period markers within the voice-free region. If we specify that we want no more that 1 false period marker in 20 seconds of speech (the approximate length for a sentence) this leads to 1 error out of 160000 samples (using an 8KHz

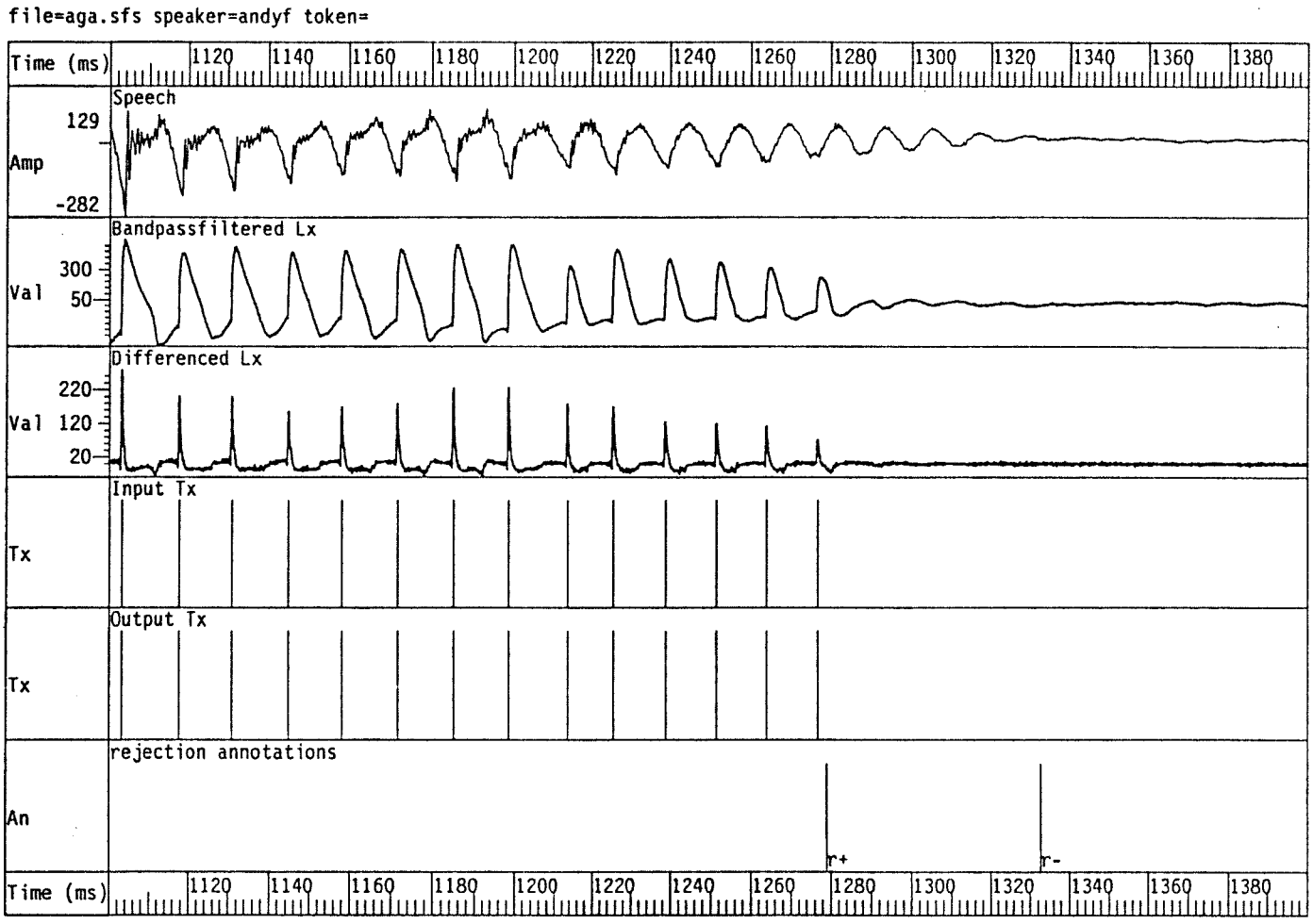


FIGURE 1

Typical operator's view when using the interactive period marker location program. The top trace shows the speech waveform. Below it are the band-pass filtered and differenced laryngograph waveforms. Next are the period markers from the automatic analysis, followed by the period markers from the present analysis. In the bottom of trace the rejection annotation labels are displayed, which can be set to label certain regions of the signal unreliable and not suitable for future analysis.

sampling rate). This correspond to a probability of error of around 10^{-5}. Writing the threshold as

threshold = $A\sigma_{dlx} + \mu_{dlx}$

Then the probability of an error is given by P_e, where

$P_e \approx (1/2)\text{erfc}(A/1.414)$

Where erfc(x) is the complementary error function for x. For a value of P_e of 10^{-5} this leads to a value of A of around 4.

Stage 4:
The generalized maxima that occur over the threshold value and over a pre-defined minimum period value (± 0.5ms) are then calculated. This avoids the detection of multiple period markers around the point of maximum gradient that can otherwise occur if the signal is very noisy. This limits the maximum operating frequency to 500Hz. A generalized maxima constitutes a point, which is above the threshold, that is greater in value than its neighboring points, and that also is the maximum value over a 0.5ms time window before and after its own location. If the threshold has been well chosen by the previous stage of processing, the generalized maxima will constitute the final period markers, with only a few exceptions.

Stage 5:
Potential period markers are then rejected if they do not have corresponding local maxima and minima in the laryngograph cycle, or if they share the same maxima and minima. In the latter case, only the period marker with the maximum differenced laryngograph point is retained.

Stage 6:
Potential period markers are also rejected if they are too isolated from other period markers by a predetermined range of 20ms. This has the effect of removing any spurious period markers that may arise. This limits the lowest operating frequency to 50Hz.

stage 7:
The period marker values are finally written out to another item in the SFS file.

1.5.3 Interactive fundamental period estimation algorithm

It is very important that the period markers used as a reference for training and testing values are reliable. To this end, another program was used to permit manually check the automatic period marker labelling and permit changes to be made if necessary. In addition, it provides the means to reject sections of speech, so that they can be ignored and not be used for training and testing of the MLP-Tx algorithm. This latter point is important, because occasionally there are regions of the signals where the speech and laryngograph apparently contradict each other. For examples, such situations arise at th end of unstressed segments where firm

vocal fold contact is not made, and in the hold phase of plosives. In the former case, there is no laryngographic evidence of voicing when there is clear acoustic evidence. In the latter case, the reverse situation occurs. These issues are discussed by Howard & Lindsey (1988). Figure 1 shows a typical operating window seen by the user whilst using the program. The program operates by displaying the speech pressure waveform, the bandpass filtered laryngograph waveform, the differenced laryngograph waveform, and the preliminary estimate of the period makers generated by the automatic analysis program. In addition there is another trace representing the output from the interactive program, which consists of a set of period markers that are initialized from the automatic analysis program. The program lets the operator zoom in and examine the waveforms in the required detail. In addition, the user may select a new threshold on the differenced laryngograph waveform, and re-run the analysis over a selected region. In this way, period markers may be added or removed as desired.

2. PRELIMINARY EXPERIMENTS IN SPEECH FUNDAMENTAL PERIOD ESTIMATION USING PATTERN CLASSIFICATION

2.1. Background to the development of the MLP-Tx algorithm

2.1.1. Introduction

This section provides the background reasoning behind the design and preliminary development of an algorithm that performs speech fundamental period estimation using pattern classification. This algorithm, called MLP-Tx, uses a multi-layer perceptron classifier (MLP) to estimate the speech excitation markers (Tx). It is shown that the estimation of the fundamental period of speech can be performed using a system with the same basic structure as one previously used to perform the task of voicing determination, providing the system parameters in the latter are suitably adjusted. Some initial experimental results generated using a preliminary configuration of the MLP-Tx algorithm are given. Limitations in this initial configuration and experiment are then discussed.

2.1.2 Initial work on voicing determination

Preliminary research by the author was concerned with voicing determination algorithms. This involved using the laryngograph as a reference against which other algorithms could later be compared. Various approaches have been employed in the past to tackle this problem. Many of these schemes involved the analysis of a single parameter of the speech signal, and often operate in conjunction with fundamental frequency estimation algorithms; for example, the cepstrum algorithm (Noll, 1964,1967), or autocorrelation (Sondhi, 1968). Such single feature schemes typically generate an estimate of the regularity of the input signal, and the speech is classified as voiced or unvoiced depending upon whether this value is above or below a preset threshold. One more sophisticated approach that appeared particularly encouraging was the statistical pattern recognition approach of Atal & Rabiner (1976), since their scheme provided the means to combine several features in an optimal and automatic way (by training the

classifier). The features they employed were the speech energy, zero-crossing rate, autocorrelation coefficient at unit sample delay, the first LPC predictor coefficient and the energy of the LPC prediction error. These were defined on a frame-by-frame basis and used to generate an input vector which was then fed to the input of a pattern classifier. The classifier was initially provided with labelled speech data and was trained to perform the desired voicing determination task. A scheme using the then newly emerging pattern recognition technique, the multi-layer perceptron (Rumelhart et al., 1986), was used by Peeling & Bridle (1986) to recognize several acoustic-phonetic qualities of the speech signal, including voicing. Their system employed input pre-processing using a 19-channel vocoder, and the performance of their scheme was shown to be high.

2.2.3 Desirable qualities of the multi-layer perceptron

The MLP has also shown itself to be a robust pattern recognition technique in many applications of speech pattern processing (Peeling & Bridle, 1986; Boulard & Wellekens, 1987) and better than many standard methods for speech pattern processing (Huang & Lippmann, 1987). Atlas et al., (1990) found that the MLP performed as well or better than classification trees.

2.2.4 Experiment using pattern classification to estimate voicing

The work by peeling & Bridle (1986) prompted the author and a colleague (Mark Huckvale) to set about to develop and test voicing determination algorithms that employed 19-channel vocoder input pre-processing, and used either a Bayes' classifier for Gaussian pattern or a multi-layer perceptron to implement the pattern classifier. This work was reported by Howard & Huckvale (1987), but it is briefly described here because it forms a good basis for the introduction of the MLP-Tx algorithm.

2.2.5 Database for voicing determination experiments

To provide a set of training data for the classifiers and testing data so that the systems could be evaluated, five male speakers were recorded in an anechoic room using a high quality B&K 4134 condenser pressure microphone, together with the output from a laryngograph. The microphone was maintained 15cm from the subjects lips and was located equally forward and to the side to avoid wind noise from the subject's breath. Each speaker was require to read "The Rainbow Passage" twice; once to provide the training data and once to provide the testing data. This text was chosen because it was phonetically balanced (Mermelstein, 1977). Both channels of the data were then low-pass filtered at 5KHz using eight-order Butterworth filters and acquired into SFS files on a Masscomp 5500 computer via a 12-bit A/D converter running at a 10KHz sampling rate (Huckvale, 1988).

The voicing regions on all the data were automatically labelled using the laryngograph based techniques described earlier. Next, the speech data was analyzed using a 19-channel vocoder, which generated 19-element output frames at 10ms intervals (Holmes, 1980). Either one or three input frames were used as the input to the pattern classifiers. The additional adjacent frames included in the input vector provide context for the recognition task. The Bayes' classifier was

trained by estimating the mean vectors and covariance matrices for the voiced and unvoiced pattern vectors. Various configurations of the MLP were investigated. The MLPs were trained using back propagation with weights updated after each pattern presentation. Several passes over the training data were made until they showed no sign of further learning. The initial results from this voicing determination experiment were most encouraging. The results for the Bayes' classifier and the MLP classifiers are given as the receiver operating characteristics for the respective detectors, and are shown in figure 2 (Howard & Huckvale, 1987). The MLP was shown to give better performance than the Bayes' classifier for this task.

2.3 Speech fundamental period estimation using pattern classification

2.3.1 Similarities between voicing determination and fundamental period estimation

The problem of detecting the points of excitation in the speech signal is somewhat akin to voicing determination, except the event to be detected (the excitation marker) is essentially impulsive rather than a steady-state region. Because the precise location of the excitation marker is essential to achieve an accurate fundamental period estimate, a much higher time location resolution is needed than in the case of voicing determination. The system used for voicing determination was modified to perform speech fundamental period estimation. The MLP-based voicing determination scheme involved the classification of the input speech signal into voiced or unvoiced frames, each lasting 10ms. By reducing the frame duration, and modifying the pre-processing, the system could be used to classify frames into those that contained an excitation marker, and those that do not.

2.3.2 Initial fundamental period estimation system structure

A schematic diagram illustrating the stages involved in this process is shown in figure 3. Two questions that arose were what would constitute a suitable set of input pre-processing filters and what should the frame rate be. The original vocoder analyzer was clearly an unsuitable pre-processor for such a task, because it was specifically designed to loose information regarding the excitation present in the input speech. That is, any temporal fluctuations in the output from a channel after the full-wave rectifier stage are smoothed out using a 50Hz cutoff low-pass filter. In addition, the output frame rate of 10ms was much too low to be of any value in fundamental period estimation. Such a frame rate would give rise to a frequency quantization error of 100% at 100Hz. In the task of deciding upon the characteristics of an appropriate set of pre-processing filters, consideration was given to the appearance of a wideband spectrogram for (male) voiced speech. Such a spectrogram shows a vertical striation whenever there is a well defined excitation point in the input speech, and it is widely appreciated that such a spectrogram can be used to give a crude estimation of the fundamental period (Borden & Harris, 1980). By using input pre-processing to the fundamental period estimation algorithm that retained temporal information in the same way, the problem can be viewed as the detection of the vertical striations (this considers the problem to simply be the image analysis of a wideband spectrogram. Other pre-processing schemes are discussed is section 3).

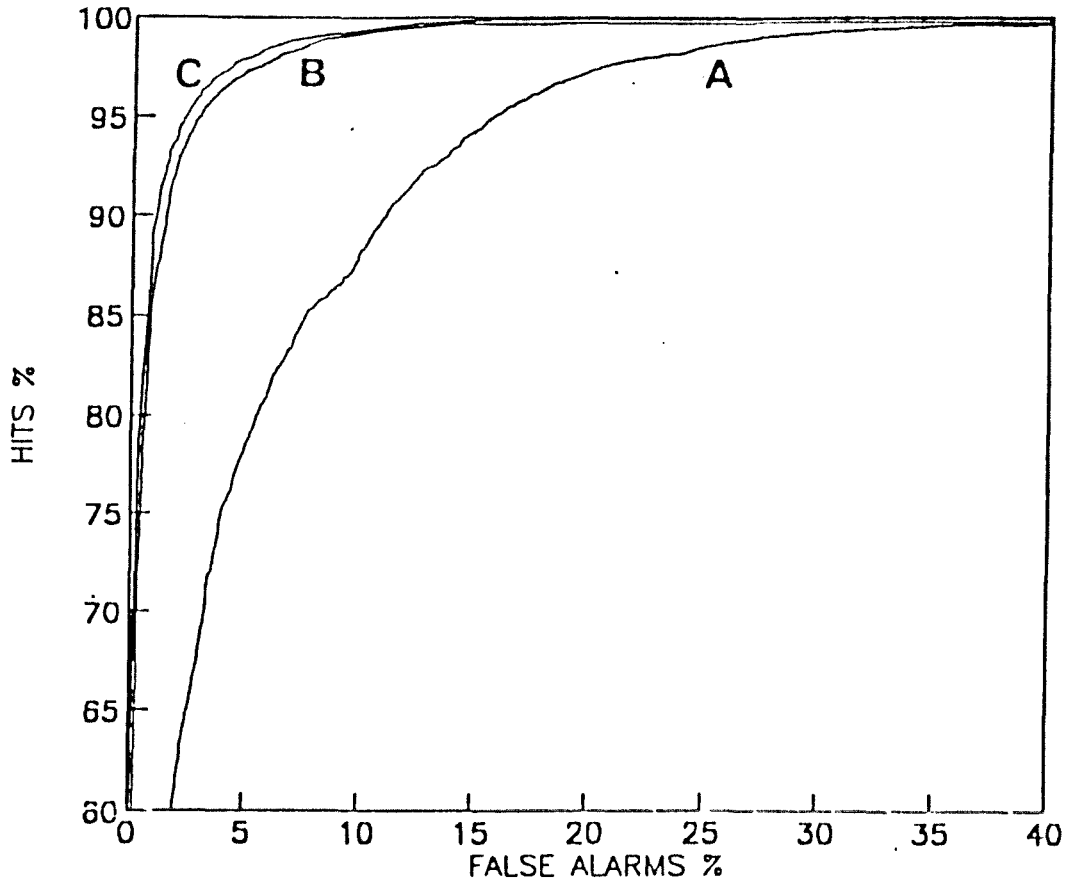


FIGURE 2
Receiver operating characteristic for voicing estimation algorithms operating on anechoic speech. Curve A is the ROC for a Bayes' classifier for Gaussian patterns using one vocoder frame in the input vector. Curve B is the ROC for a MLP with no hidden units using one input vocoder frame per vector. Curve C is the ROC for the MLP using three adjacent vocoder frames in the input vector, both in the cases with and without hidden units in the MLP. In this case, hidden units did not improve the recognition performance.

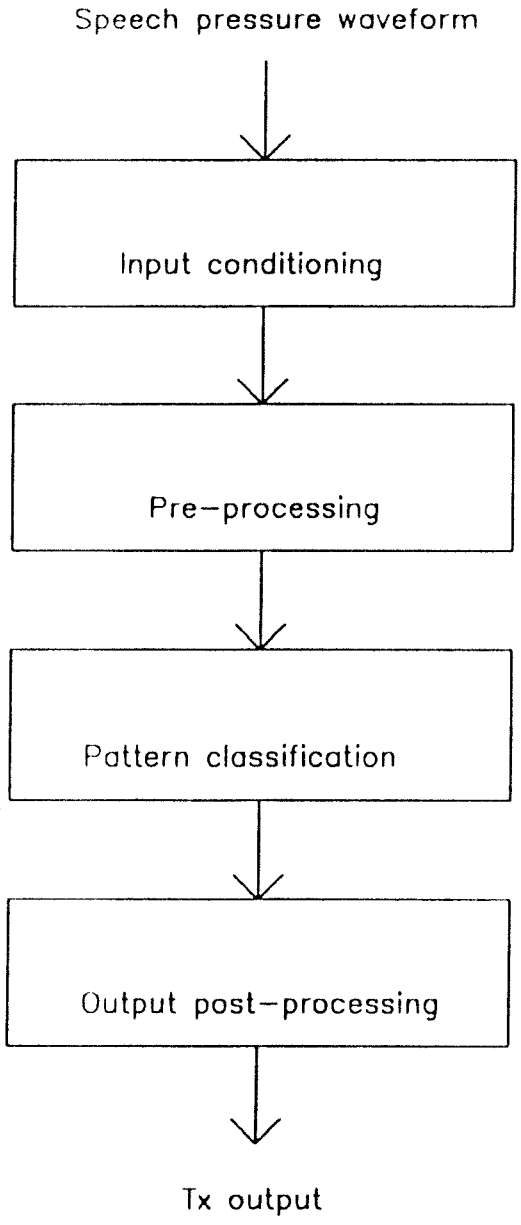


FIGURE 3
Overall system schematic diagram for the various stages in the MLP-Tx fundamental period estimation system. The first stage consists of input condition, which encompasses the microphone, anti-aliasing filtering and digitization of the input speech pressure waveform. The next stage pre-processes the raw input signal, to give an input suitable to be used as the input to a pattern classifier. Next the pattern classification transformation is applied, which results in a raw estimate of the vocal fold closures. Finally, the output is cleaned up, and converted into the overall output format by a post-processing stage.

2.3.3 MLP-Tx wideband filterbank

The filterbank that was initially used as a pre-processor for the fundamental period estimation task constituted an approximation to a wideband filterbank that was implemented using the 19-channel vocoder program. This involved selecting filter band-widths of around 300Hz. To keep the number of channels, and therefore the computational load to an acceptably low level, coarse steps between the filter centre frequencies were used. The filterbank that was arrived at comprised nine fourth order IIR band-pass Butterworth filters with -3dB points of 40-300Hz, 300-600Hz, 600-900Hz, 900-1200Hz, 1200-1600Hz, 1600-2000Hz, 2000-2400Hz, 2400-2800Hz and 2800-3300Hz.

2.3.4 Selection of frame rate

The outputs from the filters were half-wave rectified, smoothed by means of a second order low-pass Butterworth filter with a cut-off frequency of 1KHz and then down-sampled to 2KHz. Half-wave rectification was employed as opposed to the full-wave rectification originally employed in the vocoder, because the filter outputs would often be sinusoidal, and therefore full wave rectification would double their periodicities (Hess, 1983). The smoothing was carried out to prevent aliasing of the signal in the subsequent downsampling to 2KHz, which was performed to achieve data-reduction. This sampling rate was a compromise between a usable time-resolution and the amount of computation required by the subsequent pattern classifier stage. It is to be noted than several established fundamental frequency estimation algorithms also perform downsampling to 2KHz before their basic extraction stages, again to reduce the computational load (for example SIFT, Markel, 1972).

2.3.5 Data for the MLP-Tx experiment

The data use to train and test the MLP-Tx algorithm was the same as that was for the earlier voicing estimation task. The rainbow passages training data contained approximately $3x10^5$ 0.5ms frames of data, as did the testing data passages.

2.3.6 Adding noise to the speech signal

To give a more realistic task for the MLP-Tx algorithm, copies of the original Rainbow passage data was contaminated with additive canteen noise at levels of 0dB and 20dB SNR, providing two sets of data. The noise signal was recorded in the UCL refectory at lunchtime, and included impulsive noise and background conversations. The SNR was defined in terms of the maximum power found in any 500ms window. The spectrum for the canteen noise appears in figure 4. It can be seen that a lot of noise power lies in the region below 1KHz in which speech contains most power. Two experiments were carried out using this data: Testing the performance of the MLP-Tx algorithm on the 5 male speakers at both 20dB SNR and at 0dB SNR.

2.3.7 Labelling training and testing data with excitation markers

The output pattern classes for the data were labelled automatically using the previously described algorithm operating on the output of the laryngograph. In this case, an output class was defined every 0.5ms. Whenever a period epoch marker was present within a frame, the target pattern for that frame was set to a

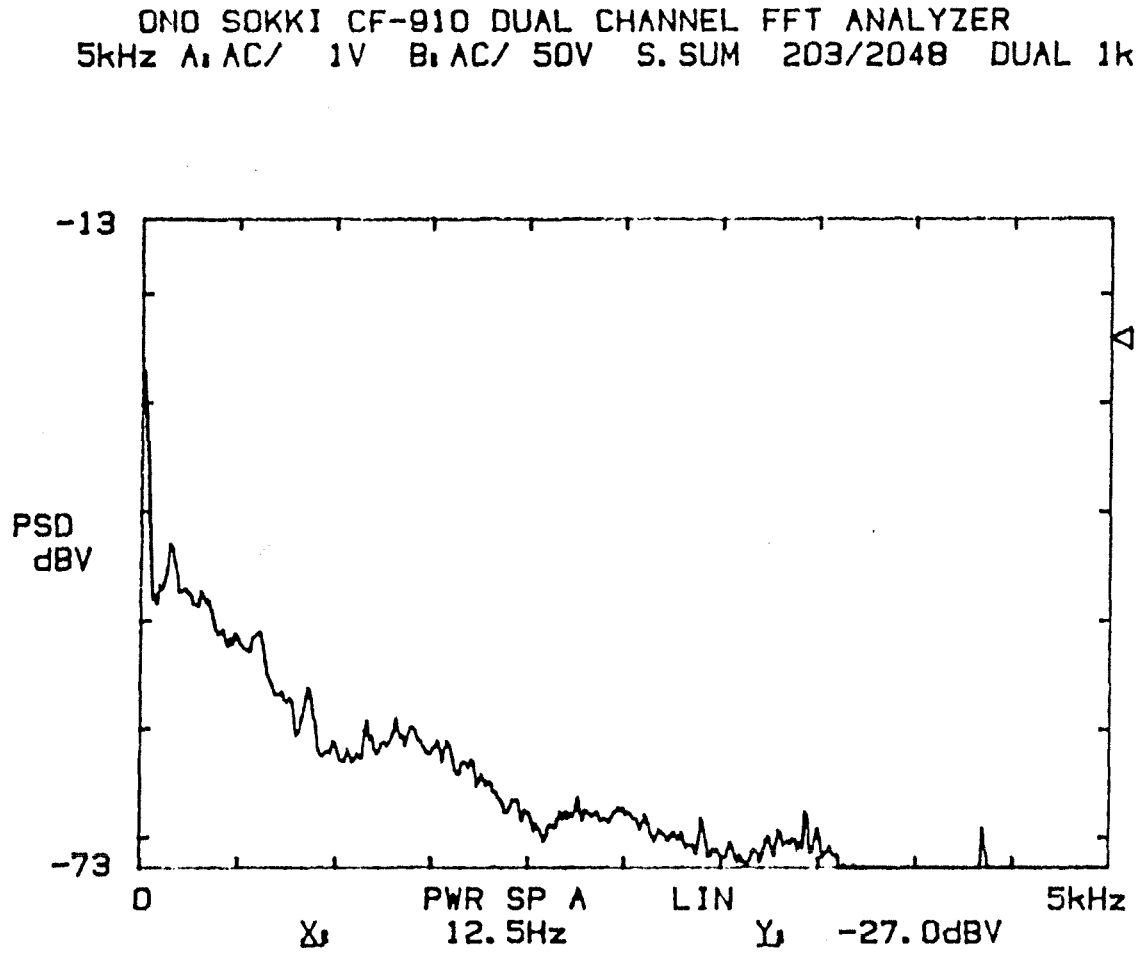


FIGURE 4
Power spectrum for canteen noise

value of 1.0. Otherwise the target patterns were set to a value of 0.0.

Figure 5 shows a close up of part of a small piece of speech and associated signals. Trace A is the speech pressure waveform, trace B the laryngograph waveform. The output from the filterbank is shown as a grey-level display in item C. It can be seen that temporal variation concerning the excitation is retained. Trace D shows the reference period markers generated from B.

2.3.8 Pattern vector generation

In order for the pattern classifier to have a sufficient information to perform its task of detecting the excitation points in the input speech signal, it is necessary for it to have evidence of the input over a window of greater width than just a single 0.5ms frame. For the first experiment on a moderate sized male database, a symmetrical input window of 20.5ms was used to generate the input to the pattern classifier.

2.3.9 Training the networks

Two networks with the same structure were trained for operation in the two different noise conditions. Training was performed using 10 passes over the input data with learning parameters alpha=0.9 and eta=0.05. The weight changes were made after each pattern presentation, and all in all about 3 million pattern presentations were made. The MLP algorithm was written in 'C' and ran under Unix on a Masscomp MC5600 series computer. The training was very slow and took several weeks. The error signal generated during training was used to gauge the completion of training (this was the normalized mean-square error between the targets and the MLP output averaged over all the training data). After running a set of similar experiments, it was found that good results could be achieved using a MLP network with an input window of 41 frames. In addition two layers of hidden units gave more canonical outputs than could be achieved using only one hidden layer. Finally a network was arrived at that had 369 inputs, two hidden layers of ten units and 1 output. Adjacent layers were fully interconnected. The schematic diagram for this configuration of the MLP-Tx algorithm is shown in figure 6.

The network trained on the 20dB SNR speech was used to generate output on the 20dB SNR test speech, and the network trained on the 0dB SNR speech was used to generate output on the 0dB SNR test speech. The location of the period markers were determined from the MLP outputs by simply locating the local generalized maximum peaks in the output signal that exceeded a threshold on 0.5 (the midway point between the 0.0 and 1.0 values that can be generated by the MLP) over a time window of ±1ms. The latter procedure avoids the generation of spurious pulses close to the main marker by setting the maximum detectable period to 2ms (which corresponds to a maximum fundamental frequency of 500Hz).

2.3.10 Qualitative evaluation of results

The results from these preliminary experiments are first given in terms of visual examination of the output from the MLP networks and of frequency contours

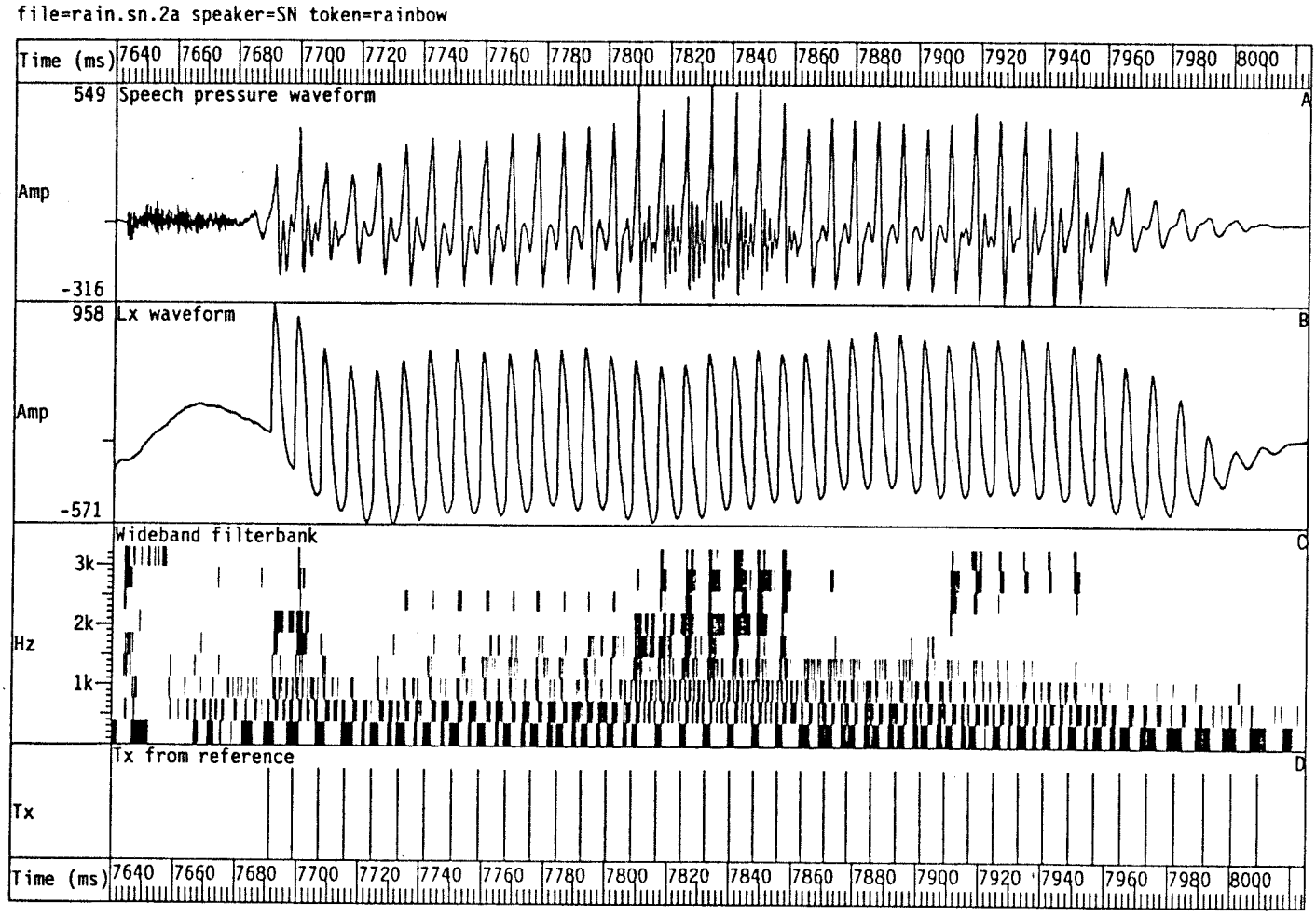


FIGURE 5

Plot showing speech pressure waveform (trace A) and corresponding laryngograph signal (trace B). The output from the 9-channel wideband filterbank is shown in trace C. The fundamental period estimates obtained from the laryngograph are shown in trace D.

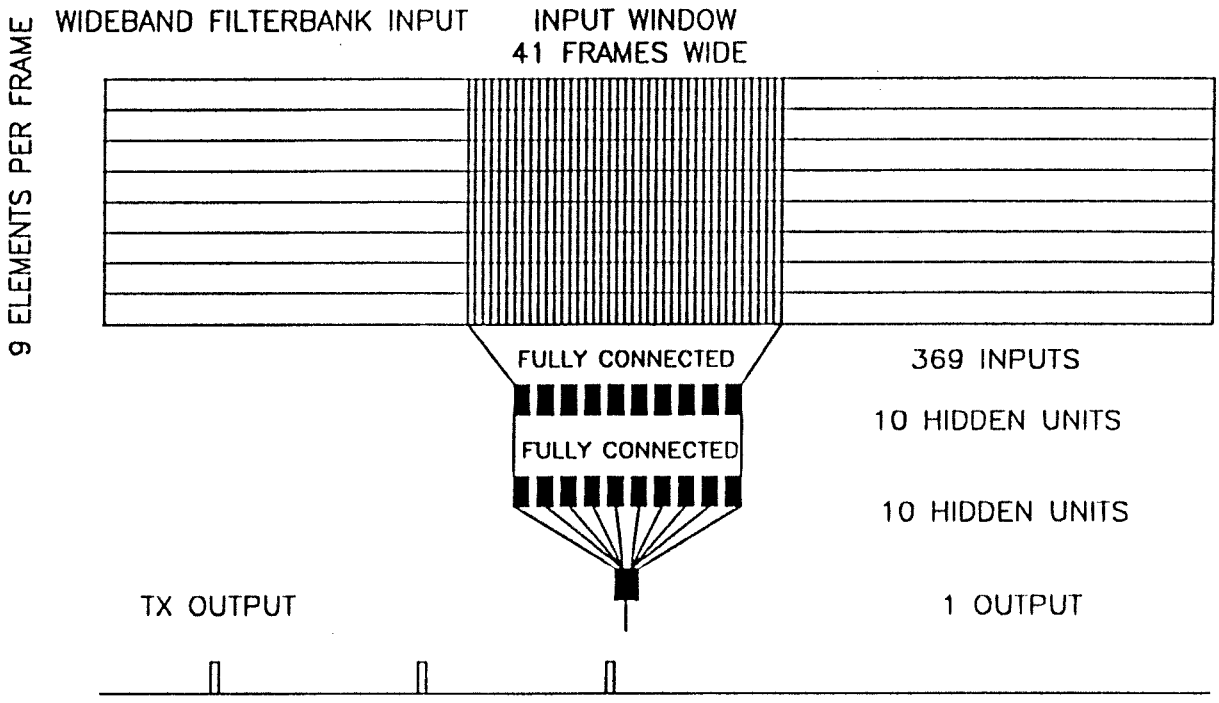


FIGURE 6
Schematic diagram of the initial MLP-Tx algorithm used for the first experiments employing a moderately sized database.

generated from their corresponding period markers. Two quantitative comparisons were then carried out.

Trace C in figure 7 shows the output of the MLP-Tx algorithm operating on a piece of test speech at a 20dB SNR, which is shown in trace A. Trace D shows the period markers from the reference period marker algorithm using the laryngograph waveform, shown in trace B. It can be seen that there is good correspondence between them and the output from the MLP-Tx algorithm. Trace E shows the period markers that can be obtained from the MLP-Tx output by detecting its peaks. Trace F shows the output of the peak-picker algorithm for the purposes of comparison, because this is the device the MLP-Tx algorithm is intended to replace; it is described in section 3.

Trace C in figure 8 shows the frequency contour generated from the output of the MLP-Tx algorithm operating on the 20dB SNR speech shown in trace A. For the purpose of comparison, the frequency contour for the reference laryngograph algorithm and the peak-picker algorithm are shown in traces B and D respectively. The MLP-Tx frequency contour follows the reference quite well. The effects of the reduced sampling rate of 2KHz can be seen in the coarser quantization of this contour than those due to the reference or the peak-picker.

Figure 9 shows the output from the MLP-Tx algorithm operating on the 0dB SNR speech signal. The first window shows the noisy speech pressure waveform. The second window shows its wideband (300Hz bandwidth) spectrogram. The third window shows the corresponding laryngograph waveform. The MLP-Tx output is given in the lowest window. It can be seen that performance is relatively unaffected by the additive canteen noise.

Trace C in figure 10 shows the frequency contour generated from the output of the MLP-Tx algorithm operating on the 0dB SNR speech shown in trace A. The frequency contour for the reference laryngograph algorithm and the peak-picker algorithm are shown in traces B and D respectively. The MLP-Tx frequency contour again follows the reference quite well, whereas the peak-picker shows poor performance ont only in period estimation but also in voicing determination.

2.3.11 Quantitative evaluation of results

Two quantitative comparisons were carried out on the output fundamental periods generated by the MLP-Tx algorithms and the peak-picker. These methods are described by Howard & Howard (1986). The problems involved with this type of comparisons is illustrated in figure 11 and figure 12, which show the strategies adopted to implement the comparisons. These are based upon an alignment algorithm that uses dynamic programming (Cooper & Cooper, 1983).

The first comparison involved calculating the receiver operating characteristic (ROC) (Levine & Schefner, 1981) for MLP-Tx algorithm and for the peak-picker. This is a plot of the number of false alarms (incorrectly generated period markers) against the number of hits (correctly generated period markers) generated by a given detector as its detection criterion is swept between lax (that is hits are never

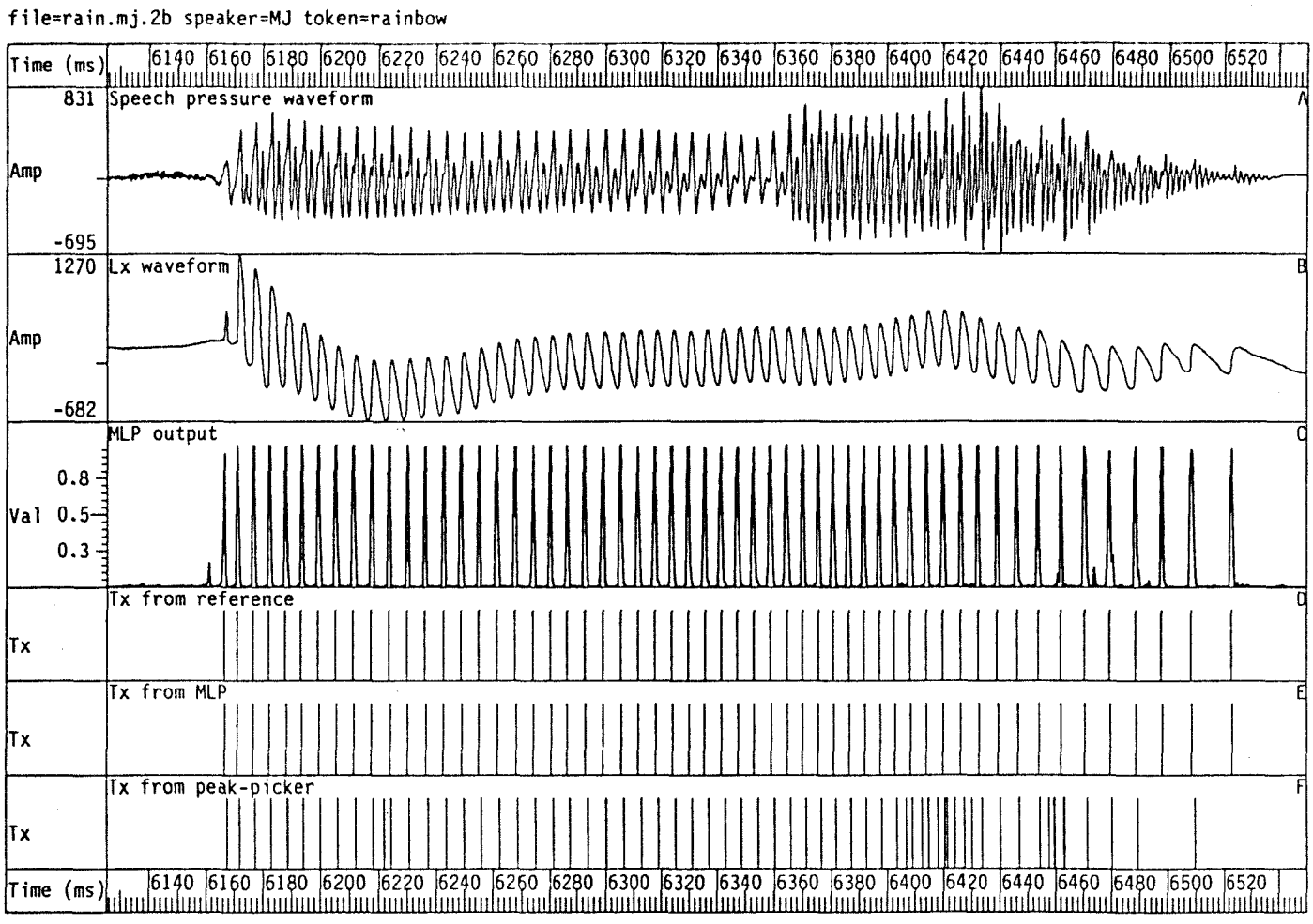


FIGURE 7

Plot showing the output from the MLP-Tx algorithm.

Trace A shows the speech pressure waveform and its corresponding laryngograph waveform is shown in trace B. Trace C shows the MLP output. The period markers from the reference, MLP-Tx algorithm and peak-picker are shown in traces D, E and F respectively. It can be seen that there is good agreement between the MLP-Tx algorithm and the reference.

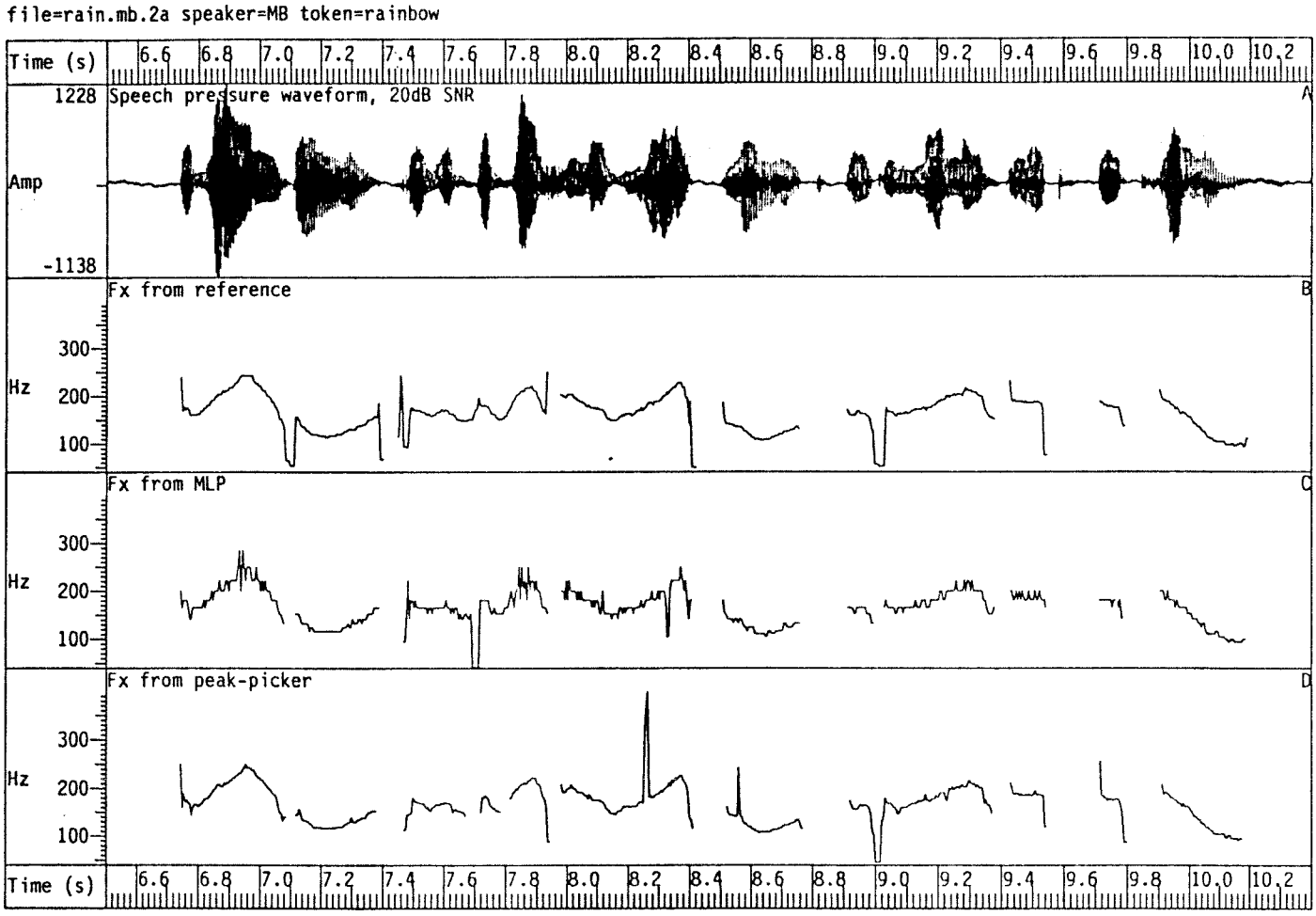


FIGURE 8

Plot showing the output from the MLP-Tx algorithm represented as a frequency contour in trace C. Trace A shows the speech pressure waveform at a 20dB SNR. The reference frequency contour is shown in trace B, and the frequency contour from a peak-picker is shown in trace D. The quantization error due to the 2KHz output frame rate is clearly visible in the output due to the MLP-Tx algorithm.

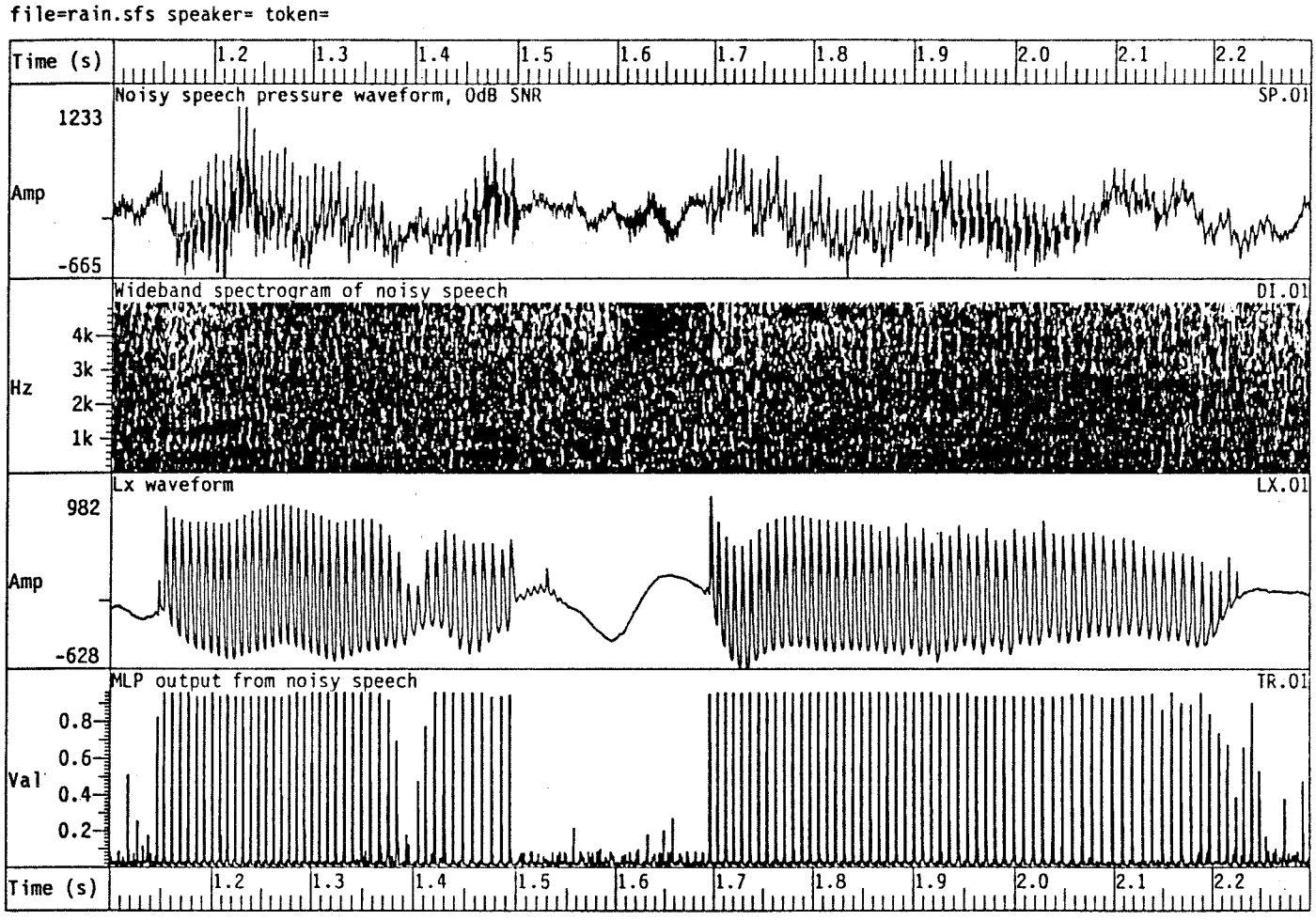


FIGURE 9
Plot showing (in th bottom window) the output from the MLP-Tx algorithm operating on speech with added canteen noise at a 0dB SNR. The second trace shows a wideband spectrogram of the input speech, with the laryngograph waveform shown below it.

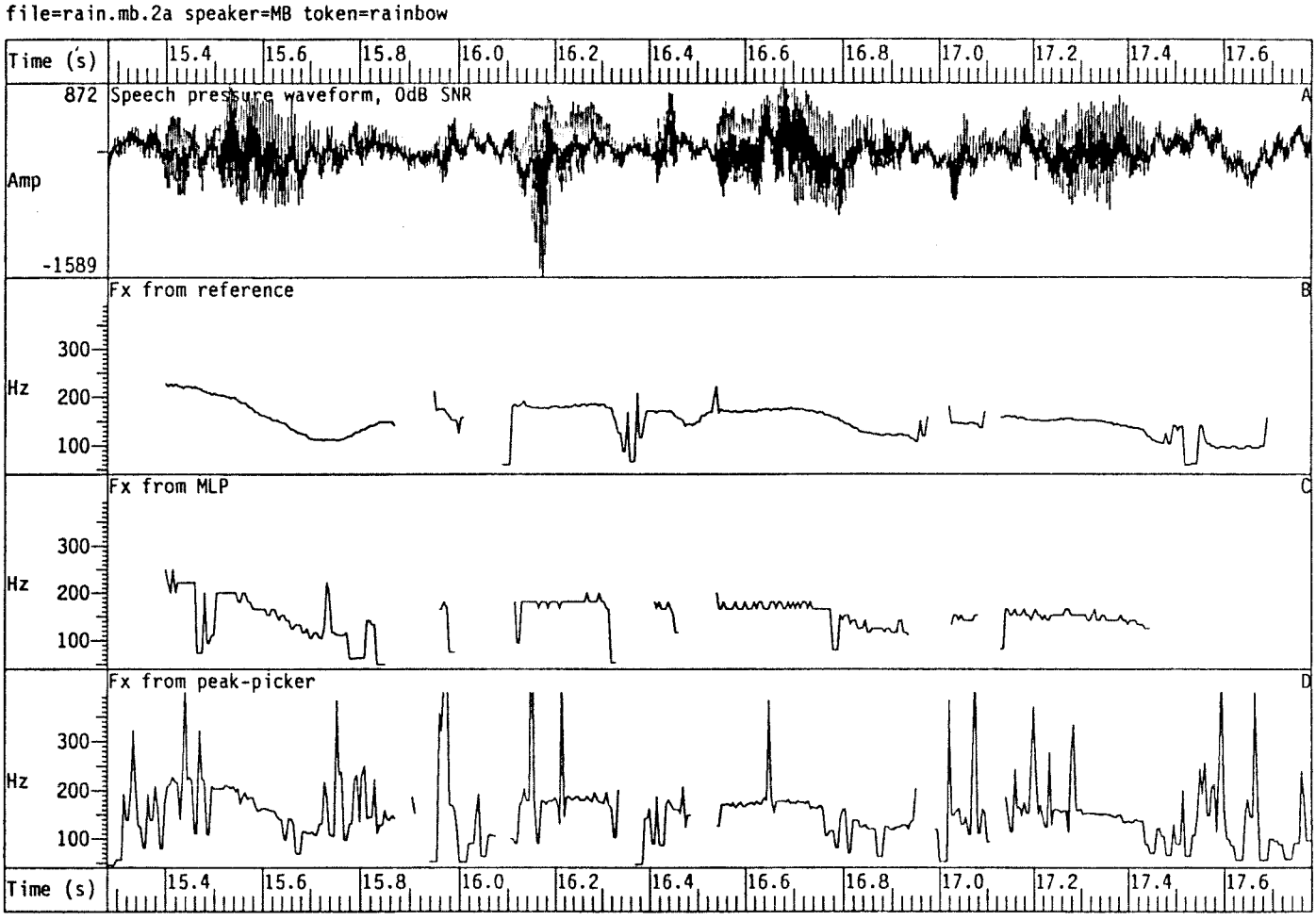


FIGURE 10

Plot showing the performance of the MLP-Tx algorithm operating in the presence of "canteen noise" at a 0dB SNR. The noisy speech pressure waveform is shown in trace A. The reference frequency contour is shown in trace B, the MLP-Tx frequency contour is shown in trace C and the frequency contour from a peak-picker is shown in trace D. It can be seen that the performance of the peak-picker is more affected by the noise than is the MLP-Tx algorithm.

REQUIRE ALIGNMENT BETWEEN REFERENCE AND TEST TX

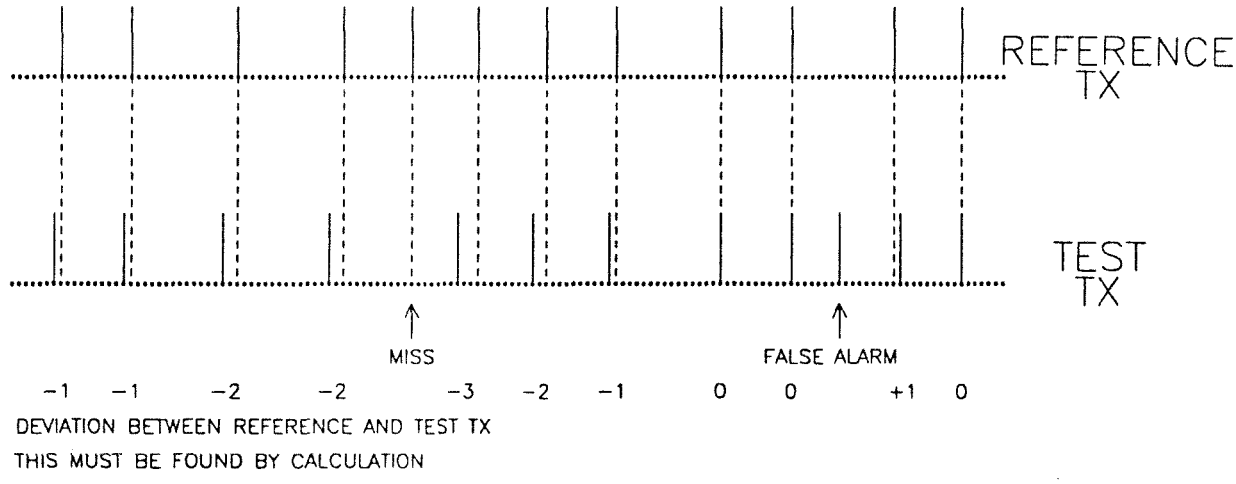


FIGURE 11

Illustration of a sequence of reference and test period markers, with the differences shown. To perform comparisons, it is necessary to first identify the correspondence between the test markers and the reference markers. After this has been accomplished using a dynamic programming algorithm, useful information relating to the jitter between the test and reference markers, the absence of test markers (misses) and in inclusion of unwanted test markers (false alarms) can be found. The jitter deviation between the test marker and reference marker gives an indication of the accuracy of the algorithm under evaluation.

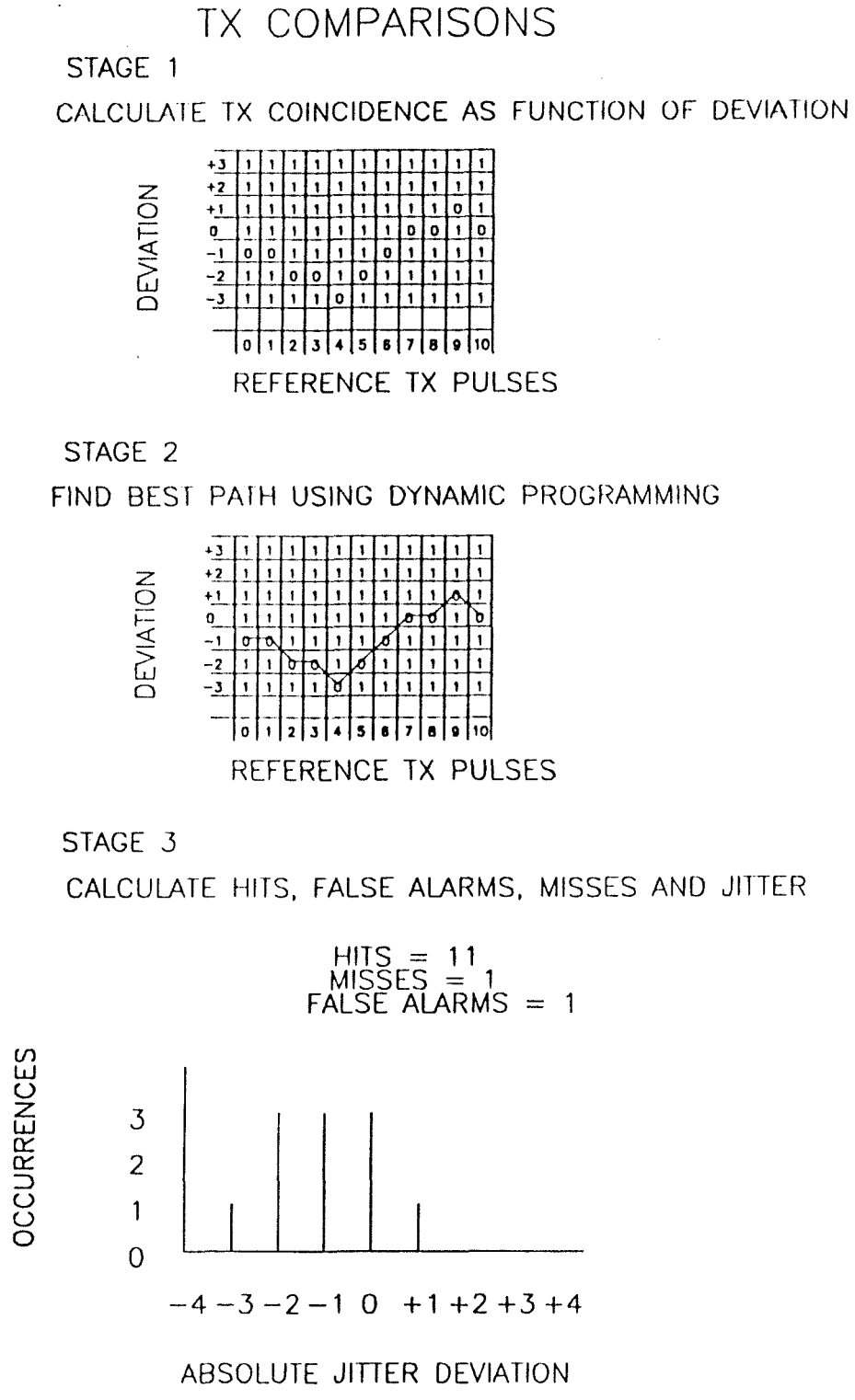


FIGURE 12
Illustration of the operation of the fundamental period comparisons. Stage 1 involved determining the local coincidences of a period marker against test period markers. The results from this is shown in the first matrix. The next stage involved finding the best path through the deviation matrix, so that it is possible to assign corresponding reference markers and test markers to each other. After this has been performed, the number of hits is simply given by counting the reference markers that have corresponding markers, and the false alarms are the test markers without a corresponding reference marker. The local time differences (the jitter) between the test and reference markers is the warp vale associated with each hit. These operations are illustrated in stage 3 of the figure.

missed, but false alarms are detected) and harsh (that is, false alarms are never generated, but some hits are not detected). This corresponds to changing the threshold between 0.0 and 1.0 in the case of the MLP-Tx algorithm. Figure 13 shows the ROCs for the MLP-Tx algorithm (marker A) and for the peak-picker (marked B) respectively, both operating on the 20dB SNR speech. The higher curve for the MLP-Tx algorithm indicates that it is a better detector of the period marker than the peak-picker (these ROCs were only calculated for a small portion of the test data, because of a limitation in the earlier analysis programs. However, the form of the ROCs obtained for different portions of the data all showed the same trends as indicated in figure 13).

The second comparison involves calculating the "jitter" in the placement of the period markers by the test algorithm relative to those generated by the reference laryngograph algorithm. The results are presented as histograms of the jitter for all the test period markers that have corresponding reference markers. Figure 14 shows the jitter histograms for all the Rainbow test data at both 20dB and 0dB SNRs. Plots A) and C) show the results for the peak picker and MLP-Tx algorithm respectively, operating on the 20dB SNR speech. Similarly, plots B) and D) show the results in the case of 0dB SNR speech. It can be seen that the distribution due to the MLP-Tx algorithm is narrower than for the peak-picker, and it is no wider in 0dB SNR case than in the 20dB SNR case. The distribution due to the peak-picker is wider than that for the MLP-Tx algorithm in both cases, and shows degradation between the 20dB SNR and the 0dB SNR conditions.

The results from the ROC indicate that on the test database used, the MLP-Tx algorithm performs firstly as a better detector of the excitation marker events, although this measure does not take into account their exact location. The jitter histograms indicate that it was a more accurate (on average) detector of the excitation points in the speech than the peak-picker. Notice that this result was achieved even though there was a limit on the time resolution from the MLP-Tx algorithm of 0.5ms, whereas the peak-picker operated to a 0.1ms resolution determined by the 10KHz sampling rate of the speech waveform.

3. FURTHER EXPERIMENTS IN SPEECH FUNDAMENTAL PERIOD ESTIMATION USING PATTERN CLASSIFICATION

3.1 Limitations of the preliminary experiment

3.1.1 Preliminary experiment

The initial results showed that the MLP-Tx system performed creditably on the noisy speech used in the experiment and better than the single alternative time-domain system. However more experiments were needed in order to prove the generality of the approach, because there were several limitations to the preliminary experiments. Some of the main limitations were as follows:

3.1.2 Limitations in the testing data

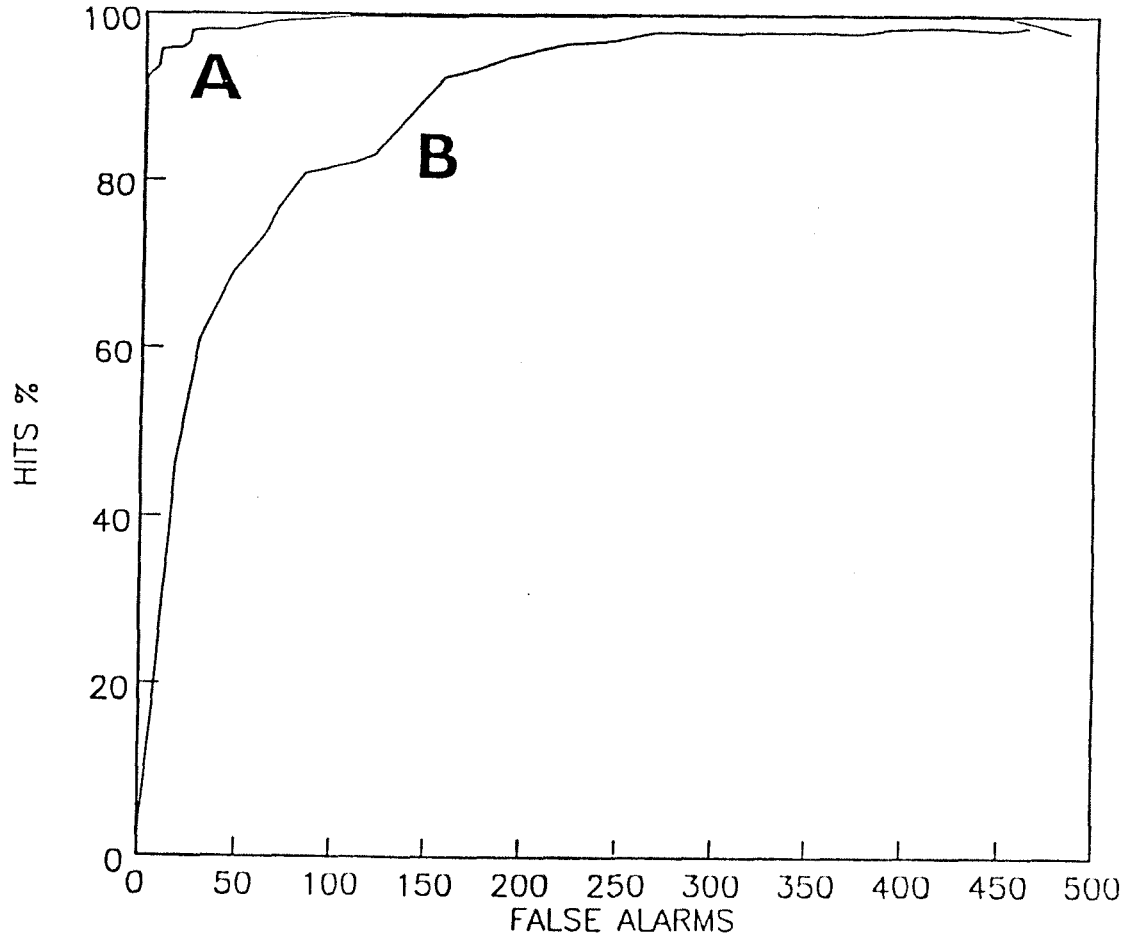
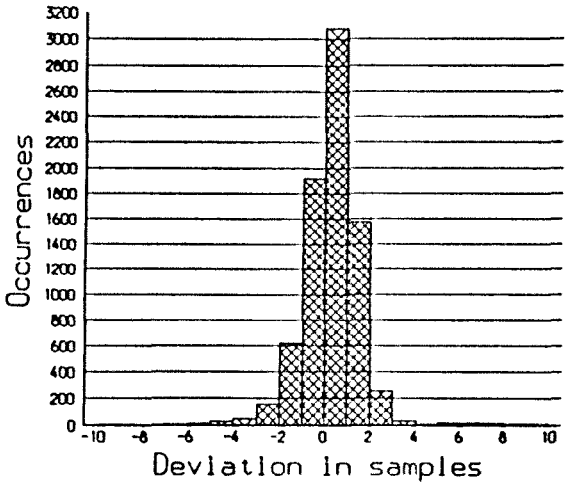
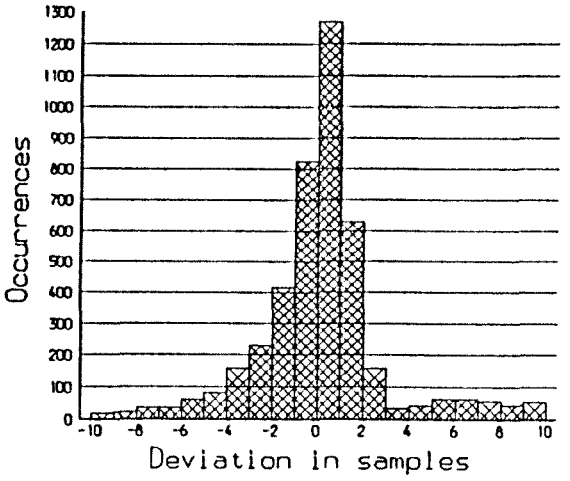


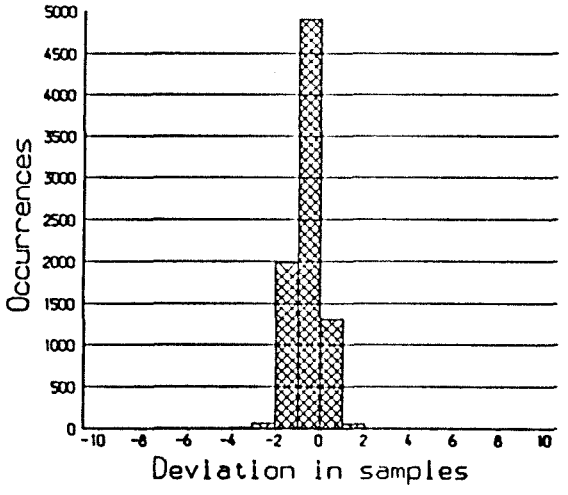
FIGURE 13
Receiver operating characteristic for the MLP-Tx algorithm (curve A) and the peak-picker (curve B) operating on speech at a 20dB SNR. These curves indicate that the MLP-Tx algorithm performs as a better detector on the test data than does the peak-picker.



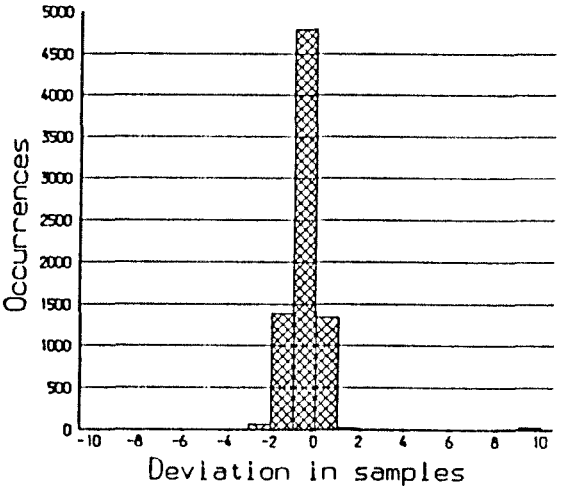
A) Peak-picker, 20dB SNR.



B) Peak-picker, 0dB SNR.



C) MLP-TX, 20dB SNR.



D) MLP-TX, 0dB SNR.

FIGURE 14

Jitter histograms for the MLP-Tx algorithm and the peak-picker. Ideal performance would be represented by a single bar at zero deviation. Graphs A and C are for the MLP-Tx algorithm, and graphs B and D are for the peak-picker, in 20dB and 0dB SNR conditions respectively. Notice that the MLP-Tx algorithm is less affected by noise than the peak-picker. All deviations are shown in 0.5ms samples.

In the preliminary experiment, the training and testing data was only composed of adult male speech. It was clearly necessary to test the algorithm on the speech from women. In addition, the same speakers were used for testing and for training. To avoid any possible advantage because the MLP-Tx algorithm adapted to the training speakers, different training and testing speakers should be used in future experiments. The speech in the original experiments was recorded anechoically and was therefore free from the effects that reverberation introduce. In addition, the speech was recorded at a constant fixed distance of 15cm from the speaker's lips. In real use (in the EPI hearing aid) the MLP-Tx algorithm would have to function in reverberant conditions at a range of distances from the speaker.

3.1.3 Lack of optimization of MLP-Tx parameters

The MLP-Tx algorithm had not been rigorously investigated especially with respect to the pre-processing employed. The filterbank used was a first attempt at a pre-processing system, and needed greater analysis and investigation.

3.1.4 Limited output period marker time resolution

The output frame duration of 0.5ms precluded the use of the algorithm in many applications, simple because the quantization error associated with this frame duration is too large. It must be noted that is **was suitable** for use in signal processing hearing aid, because such patients have poor frequency difference limens.

3.1.5 Organization of this section

This section first describes the issues concerning the requirements of database. This involves the specification of the speech material and its recording conditions, as well as the important issue of the alignment of the speech and the laryngograph signals. Next the pre-processing stage in the MLP-Tx algorithm is considered in more detail and several schemes are described. The training of the MLP is then considered, and two schemes to reduce training times are described. Finally quantitative frequency contour comparisons are described and used to compare different configurations of the MLP-Tx algorithm against two established techniques.

3.2 Requirements for a new database

3.2.1 Required range of speech and speakers.

In order to both train and then rigorously evaluate the MLP-Tx fundamental period estimation algorithm, it was necessary to have a set of speech data that were representative of the kind of data that will be typical in the real use of such an algorithm. Consequently a wide range of voice excitations were required from both men and women.

3.2.2 Choice of reading passages

It is an important to take into account the amount of time and effort involved in obtaining the database, since there was not unlimited time or resources available for this task. By using reading passages, the recording sessions could be kept relatively short and undemanding on the speakers. Continuous discourse would have been a more realistic situation, but with this approach it would be more

difficult to get a good coverage of the phonetic range of the speaker. To achieve a good coverage of different speech sounds, speakers were asked to read two phonetically balanced passages, the Rainbow passage (Mermelstein, 1977) and Arthur the Rat (Abercomie, 1964). The passages were divided into paragraphs that lasted about 15 seconds. This was done so that there would be natural break-points in the recordings that would aid loading them into separate data files on the computer system.

3.2.3 Selection of recording environments

In addition to having a database that was representative of variations in speech between different speakers, it was also important to represent a wide range of environmental conditions and for the recordings to contain natural reverberation and background noise. To achieve this aim, the speech was recorded in five typical rooms, which were chosen to provide a good range of reverberant conditions. There was also be some background noise present in these recordings. The distance between the speaker and microphone was also varied between about 30cm and 200cm, since these reflect the typical range of operation for the SIVO hearing aid during use.

3.2.4 Time delay between speech and laryngograph signals

Because the laryngograph records the activity of the vocal folds by direct electrical measurement and the speech signal was recorded after it has propagated though the air over a distance between 30cm and 2m, there was a significant time delay between the speech signal recorded at the microphone and the corresponding laryngograph signal. For a sampling rate of 8KHz, taking the speed of sound as 340ms-1, a time delay between the laryngograph signal and the speech signal equal to a single sample period corresponds to a distance of:

340/8000 m = 42.5mm.

3.2.5 Effect of head movements

It is quite possible that the speaker could move by a distance of 42.5mm in the direction of the recording microphone in the course of reading the passages. If this is the case, the delay would become a function of time which would make alignment (to within a sample at 8KHz) of the speech and laryngograph signals difficult or impossible, because the alignment would then not just be due to a linear time-shift.

3.2.6 Original experiment

In the preliminary MLP-Tx experiment, using a frame rate of 2KHz, one frame corresponded to a distance of

340/2000 = 170mm.

It is clear that the movement of the head is less critical at this frame rate. Indeed, problems relating to the alignment of the speech never arose in the preliminary experiments because the recordings were carried out with the speaker in a chair with a head rest and using a fixed 15cm microphone distance from the speaker's

lips. In this case, the 170mm was sufficient to take account of the differences in the lengths of the different speaker's vocal tracts.

3.2.7 Recording speech and laryngograph data with a fixed time delay between the two signals

To train the MLP-Tx algorithm, it was essential that all the speech and laryngograph signals are consistently time aligned. That is to say, the peak differential of the laryngograph waveform had to always correspond to the same point in a speech pressure waveform, independent of speaker or recording distance. If this was not the case, the training data would not uniquely define the relationship between the excitation point in the speech pressure waveform and the peak differential in each laryngograph cycle. Consequently, the training data would have been contradictory and the MLP-Tx algorithm would not train properly.

Therefore for the training data, it was necessary to ensure that there was a constant delay (or at least constant enough for changes to be insignificant compared to a sampling period at the 8KHz rate) between the speech pressure waveform and the laryngograph signal. To ensure that this requirement was met, for the training data the speech and laryngograph signals were recorded with a microphone that was attached to a fishing rod that was fitted into a helmet that the speaker wore. In this way, any head movements had no effect on the time delay between the speech and laryngograph signal, because the microphone was always a constant distance from the speaker.

3.2.8 Selection of number of speakers

Finding willing subject for the purpose of the generation of a database was a difficult and time consuming operation. Altogether around 80 speakers were recorded, which provided a large pool of testing and training data, and also permitted several poor recordings to be discarded (sometimes it was difficult to get good laryngograph signals, and this was only full evident after the data had been acquired onto the computer system for proper analysis).

3.2.9 Training, preliminary testing and final testing data sets.

Three separate data sets were needed to both train and test the MLP-Tx algorithm. Firstly, it was necessary to use different speakers for the training and testing of the MLP-Tx algorithm. Secondly, because the algorithm was optimised in the course of its development by evaluation on test data, it was necessary to ensure that there was final previously unseen set of testing data (no used for optimization) against which the optimised MLP-Tx algorithm could be compared against established techniques. This avoided any bias towards the MLP-Tx algorithm it may have had due to adaption of its performance to the preliminary testing data. Only results for the final test data given in this paper. See Howard (1991) for other results.

3.2.10 Training data

The training data set was composed of 4 men and 4 women speakers, each of which read the Rainbow Passage and the Arthur the Rat passage. The reason that only 8 speakers, with a large amount of data per speaker were used is because it

was very important to be able to guarantee that the speech and laryngograph signals were time aligned, and this could only be checked using the consistency of alignments between different files for the same speaker recorded at the same distance. If many separate speakers had been used, with only a small amount of data per speaker, it would not have been possible to check alignment in this way.

3.2.11 Final testing data set

The final testing data set consisted of 20 men and 20 women speakers. Only one paragraph per speaker was used, and these were rotated to achieve maximum coverage of the passages.

3.3 Input signal conditioning and recording of the database

3.3.1 Recording the test databases

The microphone used for the testing recordings was a B&K 4134 condenser omni-directional pressure microphone (standard type) fitted in a B&K sound pressure meter. The microphone was mounted on a tripod and could be raised and lowered to between 1-2m from the ground to give a range of different microphone heights. The output was calibrated to 94dbA at 1KHz using a B&K calibrator. The output from a laryngograph was also recorded on the other channel. Recordings were made using a Sony DAT recorder, with 16-bit resolution at a 48KHz sampling rate. The levels for the recordings were left fixed after initial setting up from the calibrator, thus giving an absolute calibration level.

Training data was recorded using a small high quality Knowles BL-1785 piezo-electric microphone at a constant distance from the speaker, for reasons previously discussed. The frequency response of the microphone was flat within +-1dB over the range 40Hz-1KHz, and to within +-3dB over the range of 1KHz to 8KHz.

The output signal from the given microphone was fed via a pre-amplifier into a 3rd order Bessel HPF, to remove LF noise. This was necessary because there was significant signal power present below 50Hz which would otherwise lead to problems with dynamic range at the A/Ds. To achieve minimum phase distortion, a Bessel filter was used instead of a Butterworth filter.

3.3.2 Choice of sampling rate for digital acquisition

The data was acquired directly onto the MASSCOMP computer using a 12bit A/D converters operating at 8KHz in conjunction with 4-pole Butterworth low-pass anti-aliasing filtering at 3.5KHz. A sampling frequency of 8KHz constituted the lowest practical rate at which the intelligibility of the speech can be preserved. In addition it is about the lowest acceptable time resolution of the period markers that are of general. More importantly, it is also the sampling frequency adopted by telephone companies. As a consequence of this, A/D and D/A converters that operate at this frequency are easily available and significantly cheaper than those that operate at (for example) 10KHz. For practical implementations, such as in the EPI hearing aid, an 8KHz sampling rate is an practical and economically sensible choice. For example, one such device that performs the required function is the 16-bit sigma delta linear Codec chip AD28MSP02, available from Analogue

devices. The data was acquired in sections of about 15 seconds length, which was possible because pauses had been left between groups of sentences in the passages. The level was set up to make use of the full 12bit range of the A/D converters. The passages were places into a set of SFS files (Huckvale, 1988) to facilitate further signal processing and manipulation operations.

3.3.3 Automatic alignment of the speech and laryngograph signals

An automatic and highly reliable method was devised to align the speech and laryngograph signals in the training data that did not require any distance measurement to be made between the speaker and recording microphone. The first stage involved in this alignment was the estimation the period markers derived from the laryngograph signal. The these markers were then aligned to correspond to the speech pressure waveform using a two stage process.

3.3.4 Initial bootstrap alignment

In the first phase of the alignment procedure, the peaks of one speech file, corresponding to one particular distance, were found using the peak-picker algorithm (Howard & Fourcin, 1983). A linear alignment program then calculated the cross-correlation of coincidences of all the period markers from the reference and peak-picker algorithms for a range of positive and negative offsets. The correlation peak was found automatically and its location corresponded to the time-shift between the peak differential in the laryngograph cycles and the peaks in the speech. This was then used to time-align the reference period markers. This provided one file of appropriately aligned training data for the MLP-Tx algorithm, and a direct speech MLP-Tx algorithm (description given later) was initially trained upon this. Period markers generated by the partially trained MLP-Tx algorithm were used to align the speech and the laryngograph on ALL the training speech data. This ensures that all the training data is aligned self-consistently to within 1 sample at the 8KHz sampling rate. This procedure has been found very successful.

3.3.5 Checking speech polarity

A vital issue concerning the alignment of the speech and the laryngograph signals is that of speech polarity. It was very important that the speech polarity is self-consistent for all the recordings, because otherwise the alignment could not be performed. Observation of the speech pressure waveform alone is not sufficient to guarantee speech polarity. In addition, it is also not possible to use the quality of the frequency estimates from the MLP-Tx algorithm to reliably estimate speech polarity, although it does often exhibit a preference for speech of the same polarity for which it was trained. However particularly in the initial boot-strap alignment stage, when the MLP-Tx algorithm was not fully trained, it was not always easy to determine polarity on the basis of the frequency contours. This is illustrated in figure 15. One manifestation that speech inversion does have is in the location of the period markers. This is illustrated in figure 16. There is a significant shift is marker location depending upon speech polarity. This phenomenon can be used to give a very clear indication of speech polarity if the cross-correlation alignment procedure is applied to MLP-Tx period markers found using both polarities. The cross-correlation for the correct polarity showed a much more distinct correlation

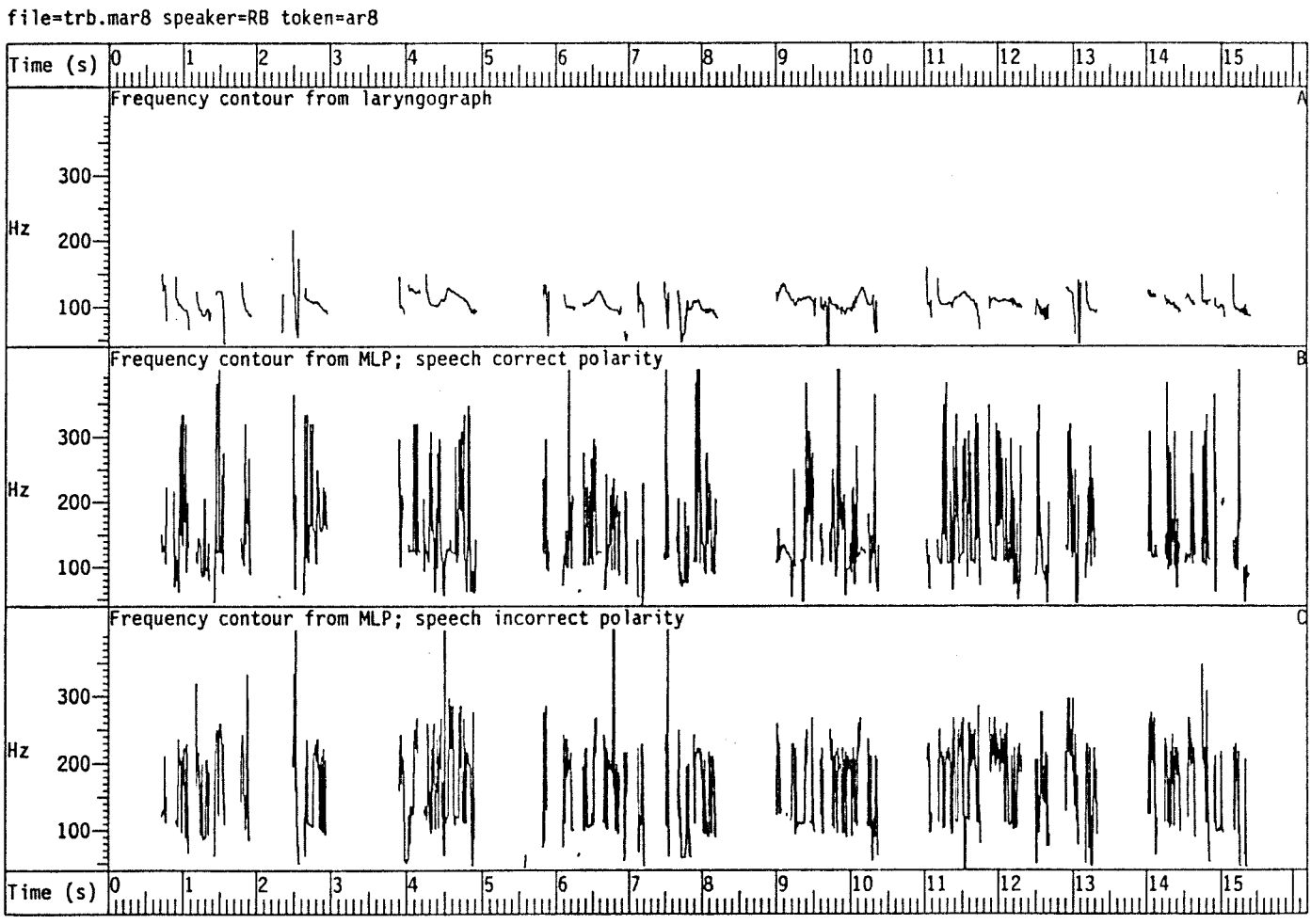


FIGURE 15

Plot illustrating the effect of speech inversion on the frequency contours that can be obtained from a partially trained MLP-Tx algorithm, as a means of speech polarity determination. The reference laryngograph contour is shown in trace A. Traces B and C show the frequency contours from the MLP-Tx algorithm with correct and incorrect polarity speech respectively. In this example, the performance is poor in both cases, and the correct polarity give barely any observable improvement in contour form. It would not be possible to determine polarity on the basis of these contours.

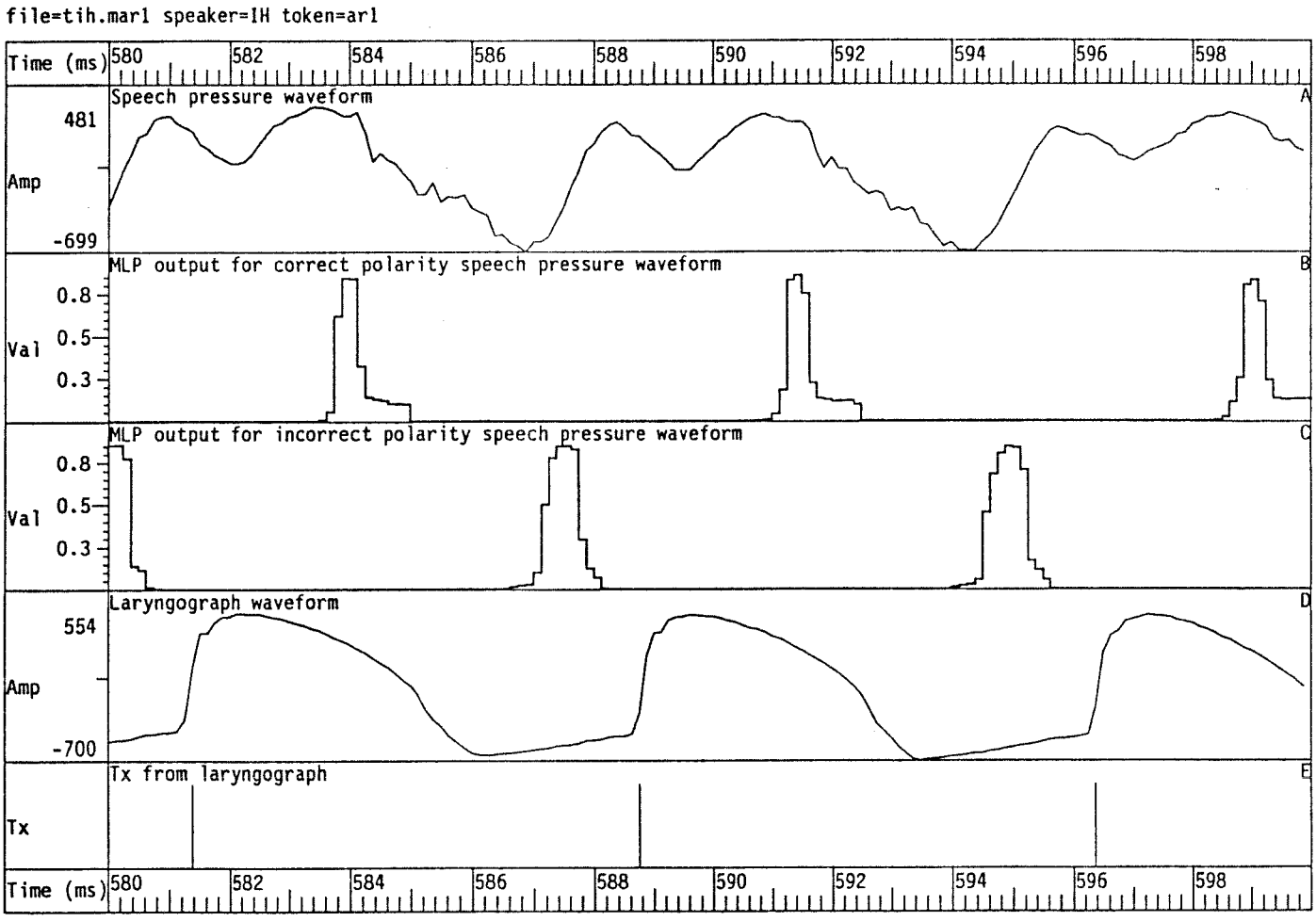


FIGURE 16

Plot showing the different occurrence times of the period marker estimates depending upon the speech polarity. Trace A shown the speech pressure waveform. Traces B and C show an output from a partially trained MLP operating on correct and incorrect polarity speech respectively. It can be seen that the two sets of locations differ by up to half a period in this case. Trace D shows the unaligned laryngograph waveform (that is, it is shown before alignment with the speech pressure waveform using the period marker cross-correlation procedure). Trace E shows the period markers obtained from the laryngograph signal.

peak than for the incorrect polarity. This is illustrated in figure 17. Notice that the difference shown in this figure is considerable, even though the correponding frequency contours showed little difference. The correctly time aligned speech, laryngograph and MLP-Tx output signals are shown in figure 18.

3.4 Using different pre-processing schemes

3.4.1 The task of the pre-processing stage

The input vector to a pattern classifier should be chosen such that it contains the information necessary to permit the desired discrimination to be carried out. In addition, the data should be represented in such a way that aspects of the signal of importance in discrimination are emphasised as much as possible, whilst at the same time information that is not required for the discrimination should be suppressed. Three different pre-processing strategies were investigated. The speech was either use directly after a linear scaling, pre-processing by means of a wideband filterbank (similar to before) or by means of an auditory filterbank.

One problem with the original design for MLP-Tx was that the location of vocal fold closures in time were too imprecise for many applications. There is naturally a compromise between the amount of processing required and the time resolution. Adopting a brute force approach, increasing the resolution by a factor N results in an input vector with N times as many elements, for a given time window width. In addition, it increases the number of frames in a given unit of time of the input data by a factor N. Consequently there is an increase of computation in the classifier according to a factor of N^2. The computation is also proportional to the number of output channels generated by the pre-processing scheme. Using direct speech input only generates a single output channel. In this case, using the fully sampling rate of 8KHz without decimation was not too computationally expensive, because there is only one input channel. Using a high input sampling rate posses much more of a problem in computational terms when the filterbanks are used for pre-processing, because they give rise to multiple output channels. For this reason, the full 8KHz sampling rate was only investigated on the direct waveform pre-processing configuration. In all cases, an input window of 20.5ms was used.

3.4.2 Direct operation on the sampled speech pressure waveform

To process the input speech samples directly, the speech was first multiplied by a small number (0.001) to scale the values of the speech samples to within the ±1.0 range. This is necessary because the MLP system used trained best when the range of inputs was of the same order as the output range of the sigmoid non-linearity.

3.4.3 Filterbank to approximate wide band spectrogram

The filterbank comprised six second order IIR band-pass Butterworth filters with -3dB points of 50-300Hz, 300-600Hz, 600-900Hz, 900-1200Hz, 1200-2000Hz, 2000-3000Hz. The outputs were half-wave rectifier, low-pass filtered at 1KHz, down-sampled to 2KHz and then linearly scaled to the range of ±1.0. This system was a cut-down version of the original filterbank designed to permitted its real-

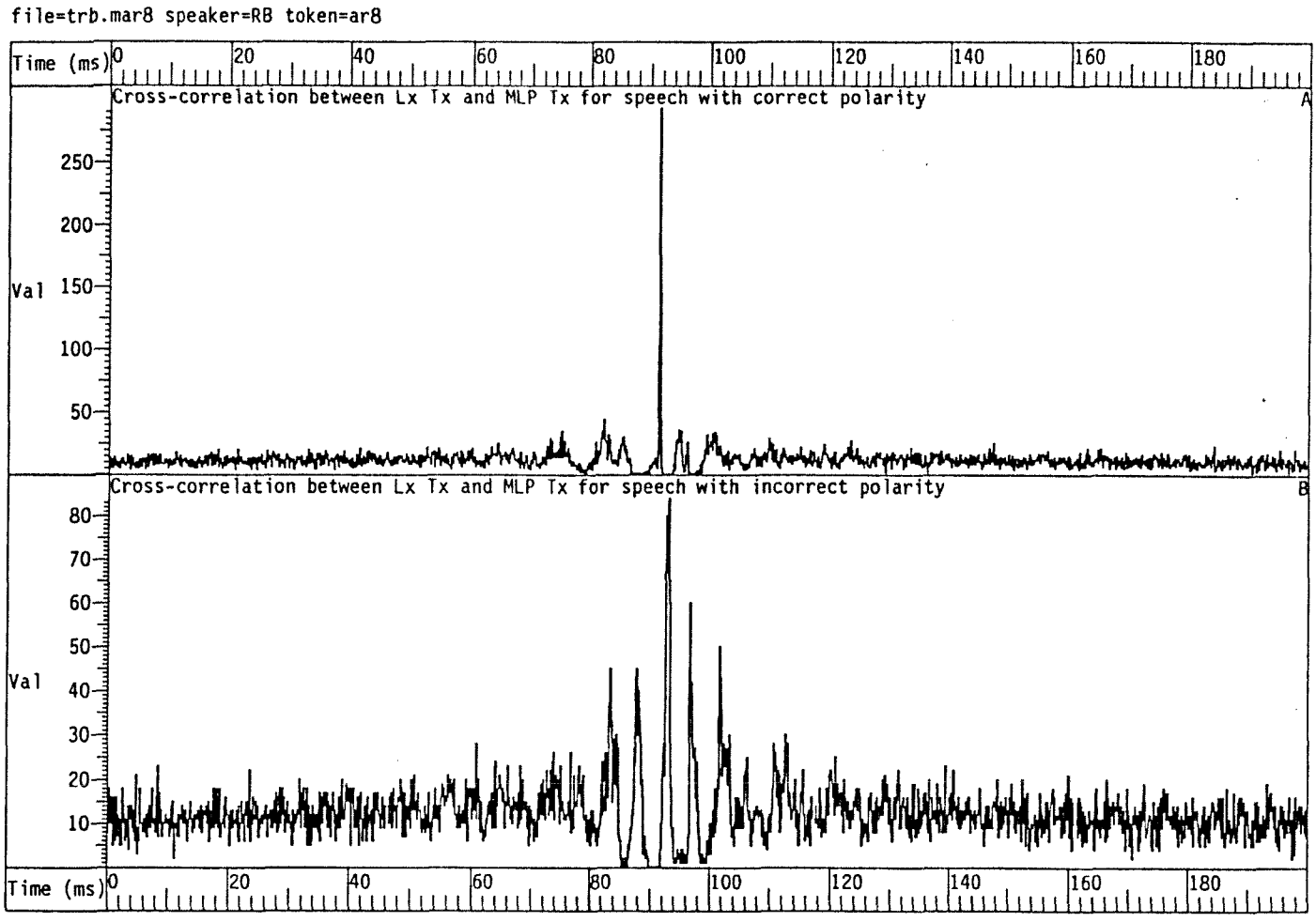


FIGURE 17

Illustration of the use of the cross-correlation between the reference period markers and the MLP-Tx period markers as a means of polarity determination. This plot is for the same speech passage as shown in figure 15. In the top of the figure, trace A shows the cross-correlation for the correct speech polarity, whereas trace B shows the cross-correlation for the incorrect polarity. The traces are shown enlarged around the peaks of the correlation in the lower part of the figure. It can be seen that this measure gives a clear indication of speech polarity, and provides a reliable method of its estimation whereas observation of the frequency contours did not. It simultaneously provides the time delay between the speech pressure waveform and the laryngograph signal that is need to align them.

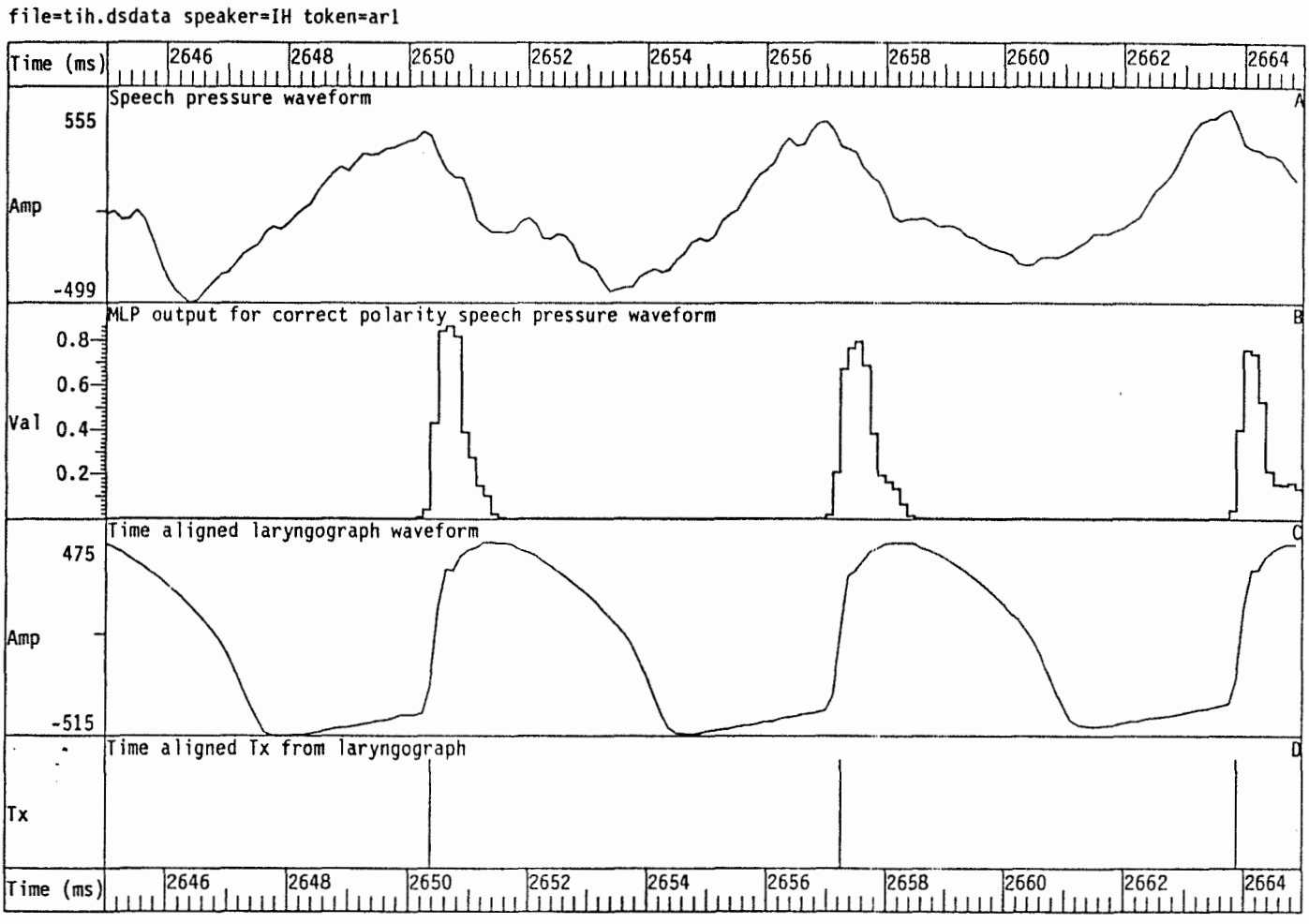


FIGURE 18

Illustration of the alignment achieved using the cross-correlation of the MLP-Tx markers and laryngograph period markers. Trace A shown the speech pressure waveform, and trace B shows the output from a partially trained MLP-Tx algorithm operating on it. The laryngograph waveform and associated period markers aligned by it are shown in traces C and D respectively.

time operation on a DSP system (Howard & Walliker, 1989; Walliker & Howard, 1990).

3.4.4 Pre-processing using an 'Auditory filterbank'

It is well known that the auditory system performs filtering of the input sounds incident on the ears. It was considered prudent to investigate an input filterbank using filters with some of the properties of those in the auditory system, because one can be sure that they do not discard important information relating to the speech excitation. For the purposes of this work, a simplified auditory model was used that consisted of a bank of gamma-tone filters, with a 1 ERB (equivalent rectangular bandwidth) spacing between their centre frequencies. This is the minimum filter density that maintains the information present in the input signal. This resulted in 12 filter channels to cover the required frequency range of 50Hz to 1KHz. The output from the filterbank channels were then half-wave rectified, low-pass filtered at 1KHz, down-sampled to 2KHz and then linearly scaled to a ± 1.0 range. This filterbank is described by Holdsworth, Nimmo-smith, Patterson & Rice (1988).

3.5 Training the MLP classifier

3.5.1 Long training times

The original MLP-Tx algorithm took a long time to train, because the MLP algorithm needed many iterations over the data-set, each of which required a lot of processing. Two techniques were used to speed up the training.

3.5.2 Adaption of the learning rate and the momentum term

One such technique is due to Chan & Fallside (1987) and it operates by dynamically adjusting the momentum term and learning rate parameters.

3.5.3 The number of patterns used to estimate weight changes

In their work, Chan & Fallside used the adaption scheme in conjunction with an updating of the weights over the **entire** training data set whenever practical, or over representative sub-sets (batches) of the data for those circumstances wherever such a scheme was not practical. The advantage of using the latter procedure is that it is possible to make MLP weight changes over a relatively small set of patterns, but ones which give a good reflection of the possible range of patterns in the data set. This is better that making the update after each pattern, because the latter is not guaranteed to give a good gradient descent, and the direction of the weight changes tends to fluctuate widely between successive updates, which makes it impossible to use adaptive learning rate and momentum term learning schemes. In practice one would not wish to update the weights only once per pass of all the data, since this would result in very slow learning. This is because it is only possible to alter the weights by a relatively small amount per update, and since the time taken to determine each update would be relatively large in this case, to perform enough updates to find a suitable solution would take a long time.

3.5.4 Sorting the pattern vectors

To implement the batch learning, the data pattern vectors used to train the MLPs were sorted into representative groups such that each group had at least one of each kind of pattern class in it.

3.6 Selective emphasis training of the MLP

3.6.1 Selective emphasis training of the MLP

Another method use to speed up training was to use selective emphasis of the training data. This method works by changing the relative emphasis of different pattern vectors, depending upon various factors during training, and was developed during the course of this work. It operates by scaling the weight changes that result from a given pattern by a factor that depends upon the estimated importance of that pattern. The importance of the pattern vector is estimated with regard to several considerations.

3.6.2 Emphasise incorrectly recognized patterns

It has been found valuable to concentrate the training on the patterns that are falsely recognized, and not overwhelm the MLP with less important weight changes from the data that is dealt with acceptably. Using this scheme, the network is only trained on those patterns it has difficulty with. This can be achieved by making the emphasis dependent on the output from the MLP as well as the target pattern class. Thus a pattern that results in an output above a preset threshold is made to give rise to weight changes which are scaled differently than if the output was below the same threshold. In practice three thresholds were employed, one for high output target patterns classes, one for low output target patterns and another for uncertain output target pattern classes. It is possible to arrange the emphasis such that patterns that give rise to outputs which are close enough to the targets are ignored (thus speeding up program operation). This makes it possible to reduce the contribution of certain regions in the training data to zero.

3.6.3 De-emphasis of the importance of boundaries

In addition to emphasizing patterns that are wrongly recognized, it was found beneficial to take less notice of patterns if their precise labelling was not important or even uncertain. This was the case in the close vicinity of a period marker. The input patterns adjacent to the one corresponding to the excitation marker will be similar. Consequently, it is difficult to train the MLP to generate one class (1.0 in this case) at this pattern, and the other class (0.0 in this case) immediately around it. However, providing the MLP can be trained to generate an output that has a monotonic rising and falling transitions around the period marker frame, the fact that adjacent output frames from the MLP are non-zero will not be important.

The use of an uncertain region around the period marker was found to be very important and beneficial. The different zones are shown in figure 19. It can be seen that zone0 corresponds to an unvoiced signal, zone1 corresponds to the pre-period marker uncertain zone, zone2 corresponds to a period marker, zone 3 corresponds to the post-period marker uncertain region and zone 4 corresponds to the region in between period markers within a voiced segment of speech.

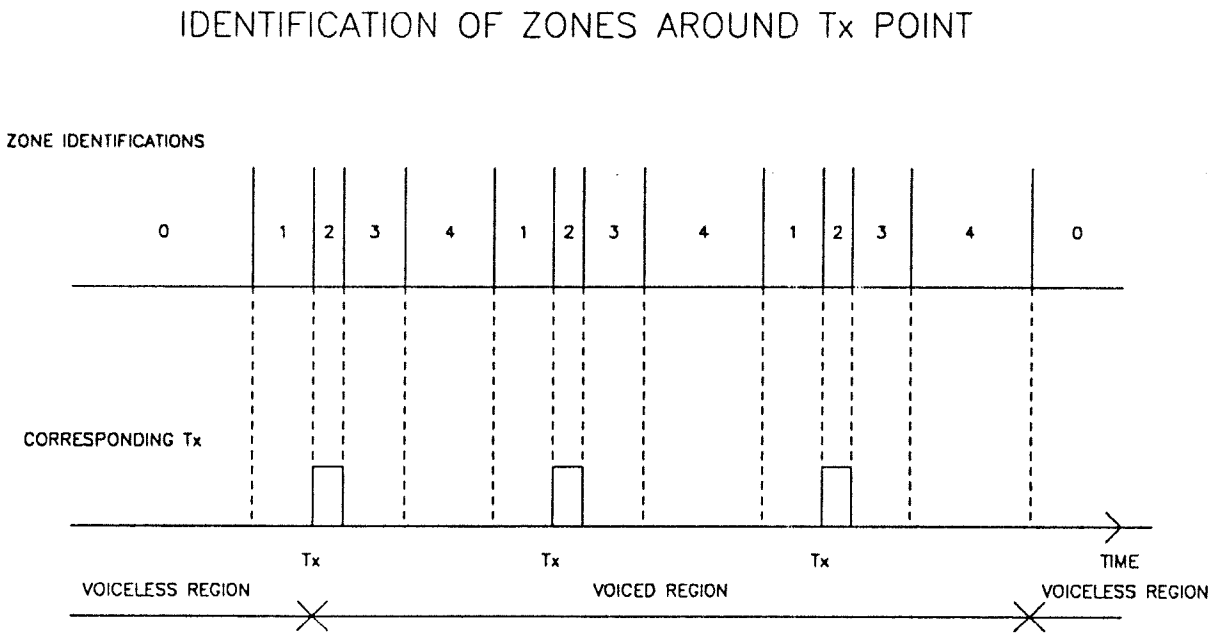


FIGURE 19
Definition of the regions around period excitation time markers. The labelling of different zones of the training data makes it possible to treat the zones differently during training. In particular, the importance of patterns that occurs in zones 1 and 3 (before and after a period excitation marker) can be de-emphasised.

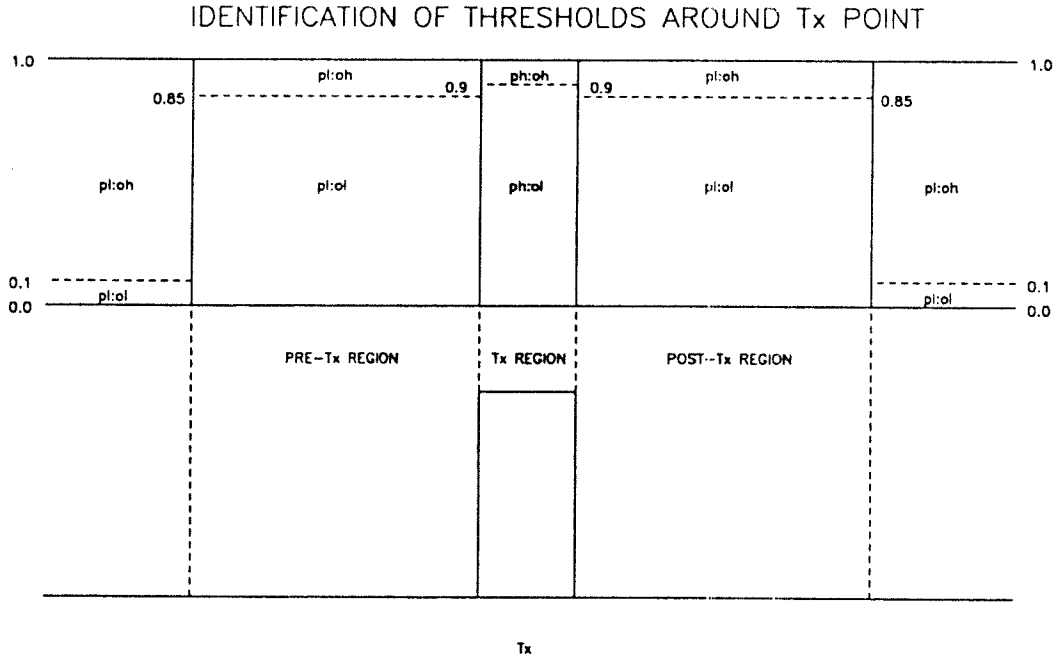


FIGURE 20
Identification of thresholds around a period excitation marker, which are used to determine whether or not an output is close enough to its target. These thresholds are used during selective emphasis training of the MLP. For the period excitation marker zone, a threshold of 0.9 is employed. For the zones adjacent to the period excitation marker zone, a threshold of 0.85 is employed. Elsewhere, a threshold of 0.1 is used.

Thresholds of 0.1 for the low zone, 0.85 for the uncertain zone and 0.9 for high zones were employed. Notice that normally the class of the uncertain zones would be set low (0.1). Using a high threshold here (means that these regions are ignored, **unless** the output from the MLP is above a value of 0.85, in which case it may compete with the **true** period marker location as the local maximum. In this case, the uncertain region is used, and trained to be of low class. These thresholds are illustrated in figure 20.

3.6.4 Faster training with selective emphasis

The selective emphasis training substantially speeded up the training of MLP networks. Using an update per pattern vector presentation, there is a speed-up in passing through the training data in excess of ten times. With larger updates, there is a speed-up in the passage through the data by about three times. This results from the fact that for those patterns that are correctly recognised, no back-propagation of error or adaption of the weights needs to be carried out.

3.6.5 Similar techniques to selective emphasis

The modified relaxation adaption algorithms due to Mays (1963) exhibit similarities to the selected emphasis method that has just been described, although the selective emphasis method was discovered independently..

3.6.6 Generalization of the training data to testing data

Baum & Haussler (1989) derived a relationship between the number of weights W in a network, the accuracy ε of the classification and the required number of training examples m. Provided that the training examples are taken from the same distribution as the testing examples, then for fully-connected feed-forward nets using any learning algorithm, they arrived at the relationship that using any fewer than $\Omega(W/\varepsilon)$ training patterns would result in the failure to correctly classify, for at least some of the time, more than a $1-\varepsilon$ fraction of the future test examples, where Ω is some constant. If we ignore this constant and set $\Omega=1$ to get a rough estimate, then for an accuracy of 90% (that is $\varepsilon = 0.1$) this implies we need at least 10 times as many training examples as there are weights in the network. This figure agrees reasonable well this the rule-of-thumb adopted by Widrow (1987). The number of excitation markers in the training data used in this work far exceed the number of weights in the MLP networks, since the largest network used has about 1620 weights and there were well over 50000 period markers in the women training data set.

3.7 Training the MLP-Tx algorithm using the different configurations

3.7.1 Training different MLP-Tx configurations

Three different experiments were carried out, using each of the pre-processing schemes described. In each case, the algorithms were only trained and tested on speech from women. Results on men and women can be found in Howard, (1991). In each case, the MLP networks were trained on all the women training data. Three passes through the data were made, all using selective emphasis. The first pass was made using weight update per pattern presentation, with no adaption of the learning parameters. The second pass was made using batch learning with

100 patterns contribution to weight updates, and employing learning parameter adaption. The final pass was the same, but using weight updates per 1000 patterns. This strategy has been found effective because the initial small updates result in fast training, whereas the later larger updates provide a final improvement in the quality of the training. The MLP networks used were as follows: For the direct speech experiment, the network had 161 inputs, 10 hidden units and 1 output unit. For the wideband filterbank, the network had 246 inputs, two layer of hidden units each containing 6 hidden units, and 1 output. For the auditory filterbank, the input had 533 inputs, two layer of hidden units each containing 6 hidden units, and 1 output. All of these network configurations were arrived at by consideration to the performance of the MLP-Tx algorithms on the preliminary training data (the results of which cannot be given here because of space limitations).

3.7.2 Generating frequency contours from the MLP-Tx algorithms

The raw MLP output were then processed to generate period markers. The task of the post-processor is to take the sampled output waveform from the MLP network that is generated as a function of time, and determine from it discrete events that correspond to the period excitation markers. This was done as before by means of a comparator circuit, with forward and backward inhibition. The period marker outputs from the respective MLP-Tx algorithms were generated on the women testing data, and the outputs converted to frequency contours samples at 100Hz by taking the reciprocal of the resulting period values. This format was required to permit comparison between the MLP-Tx algorithm and the established techniques. In all cases, the laryngograph based algorithms (described previously) were used to provide the reference frequency contours.

3.8 Fundamental frequency algorithm Comparison techniques

3.8.1 Quantitative frequency contour comparisons

Comparisons of the frequency contours generated from the MLP-Tx algorithms and the two established techniques were carried out.
It is important that the two frequency contours are aligned, and this was performed by calculating all the error metrics for a range of possible offsets, and then selecting the best fitting alignment.

3.8.2 Assessing improvements to algorithms

Evaluation techniques provide the means to test out modifications to the algorithm. If the performance of an algorithm can be estimated, the effect that alterations of parameters have on its performance can be monitored. Although the assessment of fundamental frequency and fundamental period estimation algorithms is an important issue, little work has been reported on quantitative comparisons in the literature. The best known study to compare speech fundamental frequency algorithms is due to Rabiner et al. (1976).

3.8.3 Implementation of frequency contour comparisons

A set of frequency contour comparison metrics were implemented by the author. The details of the comparisons implemented are now explained. (See Figure 21)

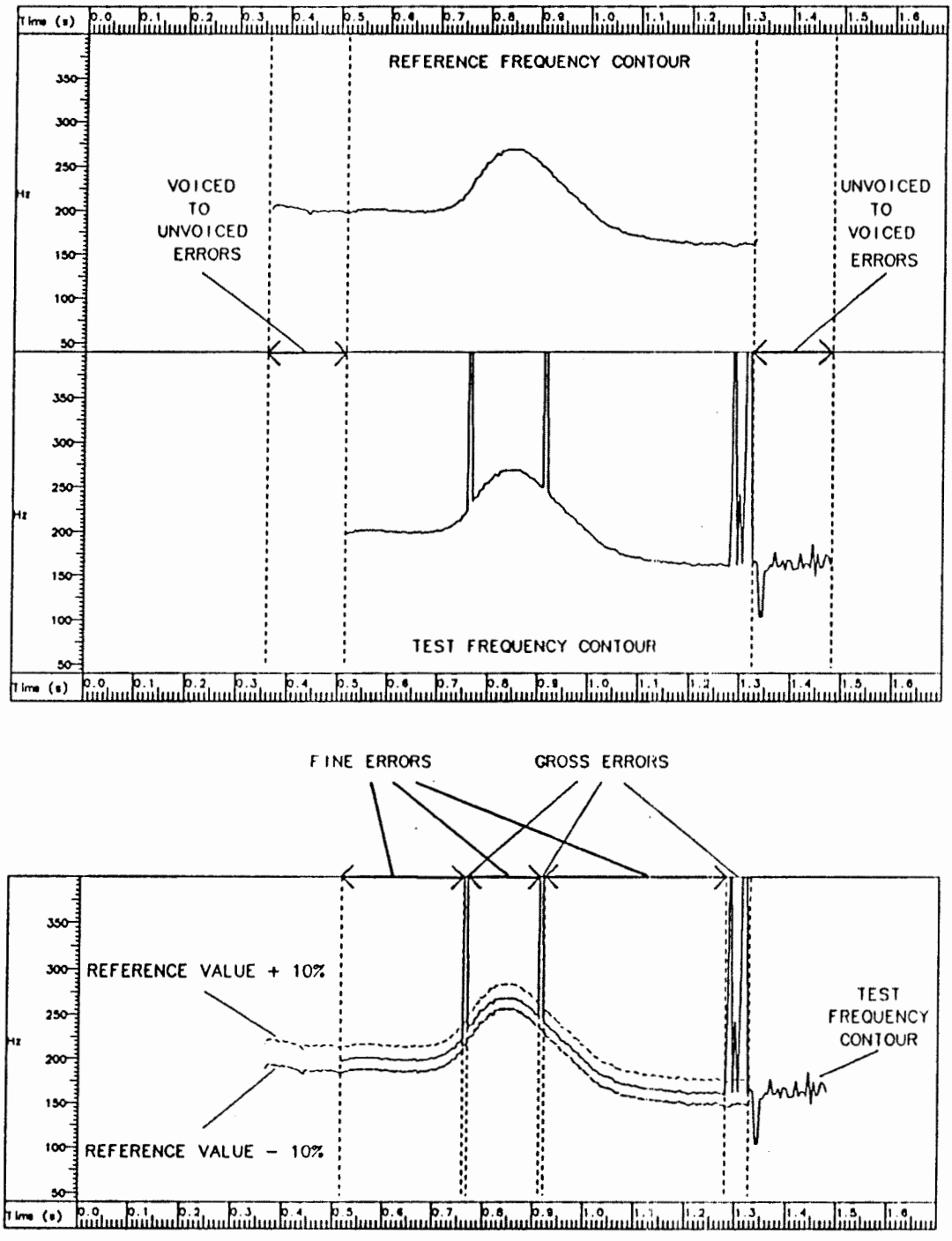


FIGURE 21

Upper box shows comparison between two frequency contours showing examples of voiced-to-unvoiced errors and unvoiced-to-voiced errors. Lower box shows comparison between a test frequency contour a reference contour to illustrate gross errors. Whenever the frequency of the test contour deviates by more than 10% of the frequency value of the reference contour (a limit shown by the two sets of dotted contours) there is a gross error. In this examples, the gross errors exceed the reference value, and are consequently chirp errors.

3.8.4 Check frame rates

The first stage in the comparison is to check that the two frequency contours are specified at the same frame-rates. Naturally on cannot make comparisons if the two sets of frequency values are not defined at the same points in time. A frame rate of 100Hz was used for all the comparisons in this paper.

3.8.5 Estimate time difference between test and reference contours

Secondly, the time-delay between the reference and the test frequency contours must be known. It makes no sense to compare frames if they correspond to different times of the input speech. If the time delay is known a priori, it can simply be entered directly to align the two contours. If not, then it must be calculated. To perform this task, the standard deviation between the two frequency contours is calculated over a preset interval (±500ms was used). The minimum point in this time function is then located, which usually corresponds to the best time alignment of the two contours. Notice that the minimum is typically well defined. This procedure can reliably align the test and reference contours, unless the test contour contains a large number of gross errors, or the search range used is too large. To be sure that the time alignment was always correct, the time alignment for all the data files known to have the same delay were examined. The modal value of the delay was then used to re-align the test and reference data in those few cases in which the algorithm had failed.

3.8.6 Calculation of errors in voicing determination

The number of voiced to unvoiced errors were calculated as follows. Since a voiced to unvoiced error occurs whenever the test frequency contour indicated no voicing and the reference indicated voicing, the maximum number of such errors is equal to the number of voiced frames in the reference frequency contour. Therefore the voiced to unvoiced errors were then expressed as a percentage of the voiced frames in the reference frequency contour.

The number of unvoiced to voiced errors were also calculated. Since an unvoiced to voiced error occurs whenever the test frequency contour indicated voicing and the reference indicated no voicing, the maximum number of such errors is equal to the number of unvoiced frames in the reference frequency contour. Therefore the unvoiced to voiced errors were then expressed as a percentage of the unvoiced frames in the reference frequency contour.

3.8.7 Calculation of gross errors

A gross error occurs whenever both the reference and test frequency contours indicated voicing, and the test frequency contour is deviated by more that ±10% of the frequency value from the reference frequency contour. These errors were then expressed as a percentage of the number of frames in the test and reference frequency contours that were both voiced at the same time. It is a trivial task then to classify the gross errors into those that corresponded to a deviation of more than 10% of the reference value and those that corresponded to a deviation of less than 10% of the reference value. This then gives an indication to whether the error are 'chirps', which occur when there is a local false rise in frequency, or 'drops', which occur when there is a local false drop in frequency.

3.8.8 Calculation of fine error statistics

A fine error occurs whenever both the reference and test frequency contours indicated voicing, and the test frequency contour is less that ±10% of the frequency value of the reference frequency contour. Theses error were then expressed as a percentage of the number of frames in the test and reference frequency contours that were both voiced at the same time. The mean and standard deviation of the fine errors were then calculated.

3.8.9 Calculation of contour statistics with different labels

The reference frequency contour used for the comparisons could also be annotated with labels to indicate the different sound types, or its local reliability. These annotation labels could then be used by the comparison program so that the metrics described above could be calculated for each label class. In this way it is possible to determine the performance of the test frequency contour with respect to different types of input sounds. More importantly, this facility was used with labels that indicated that the reference frequency contour was reliably defined, so that meaningful statistics could be generated.

3.8.9 Problems arising with comparisons

To make valuable comparisons between different algorithms, it is important to bear in mind that the metrics that we have discussed so far constitute a set of interdependent measurements relating to the performance of a algorithm. For example, in the case of the frequency contour comparisons, to give an overall rating of a particular algorithm involves consideration to its voicing determination performance as well as the gross errors and fine errors it generates. If the criterion of detection of voicing for an algorithm is altered, the relationship between the hits and false alarms changes. In addition, the number of gross errors and the statistics for the fine errors may change. Therefore, for a given algorithm, a different set of numbers (for voicing errors, unvoiced errors, gross errors, etc) may be generated depending upon the setting of the voicing detection threshold. In the evaluation of such an algorithm, we are interested in its inherent performance, not its performance for such an arbitrary threshold value. The solution to this problem used here was to set the threshold for all algorithms (whenever possible) to give similar voicing errors.

3.9 Comparing the different MLP-Tx configurations against established techniques

3.9.1 Standard fundamental frequency analysis techniques for comparison

Comparisons of various configurations of three configurations of the MLP-Tx algorithm and two established algorithms and were made against the reference laryngograph analysis system. The established techniques chosen for the purpose of comparison were cepstral analysis and a peak-picker. There is now a brief description of these techniques.

3.9.3 Cepstrum Processing

The cepstral technique (Noll 1964,1967) that is used for speech fundamental frequency determination is a special case of what is known as homomorphic

filtering (Oppenheim & Schafer, 1975). The idea behind the cepstrum is to separate out the effect of the excitation source and the response of the vocal tract from the speech waveform.

The excitation signal manifests itself in the log power spectrum of the speech as a high frequency cosine-like ripple, whereas the vocal tract response gives rise to a low frequency ripple (Noll, 1967). By calculating the inverse Fourier transform of the log power spectrum, one then gets back to a time-domain (although the signal is now the cepstrum of the input signal) which in which the temporal effects of the vocal tract and excitation are separate. Thus the cepstrum exhibits a strong peak at a quefrency (which is the term used to denote time in the cepstral domain) equal to the fundamental period duration T_0 of the input signal. At the time it was first published, the cepstral technique constituted a breakthrough, since it was much more reliable than many other approaches. However, the cepstral technique requires that there be many adjacent harmonics in the input signal; otherwise there will not be a series of ripple in the log power spectrum of the input signal and the cepstral technique will fail. On the other hand, the cepstral technique is quite able to deal with a strong formant structure in the input signal. The implementation used in this paper was due to ILS and operated with a 32ms time window on the input speech signal.

3.9.4 Peak Picker

The peak-picker is a simple time-domain device of this type (Howard & Fourcin, 1983). It was based upon earlier work by Gruenz & Schott (1949). A version of this algorithm is used in this paper for comparisons, and it is a software implementation of a small battery powered device developed as part of the External Pattern Input (EPI) group cochlear implant prothesis, at University College London (Howard, 1986). The peak-picker operates by detecting the principle peaks in the speech pressure waveform that occur at the beginning of each period. Because it operates on a period-by-period basis and is thus able to retain irregularity in the laryngeal excitation. In addition, the input to output delay is relatively small with this algorithm, making it well suited for real-time applications in pattern processing hearing aids. The problem with this algorithm is that its performance in noise and harsh environmental conditions are inadequate for many applications.

3.9.5 Discussion of results

Figure 22 shows the gross error for the six different algorithms. It can be see that the cepstral analysis gives the fewest number of gross error, and the MLP-Tx algorithm using direct speech operation gives the next least. The MLP-Tx algorithm using the auditory filterbank is better than with the wideband filterbank. The peak-picker gave the worst performance.

Figure 23 shows the chirp errors for the six different algorithms.
The cepstrum gives the lowest chirp errors, and the direct speech MLP-Tx algorithm gives second best performance. The auditory filterbank MLP-Tx algorithm is again better than the wideband filterbank MLP-Tx algorithm.

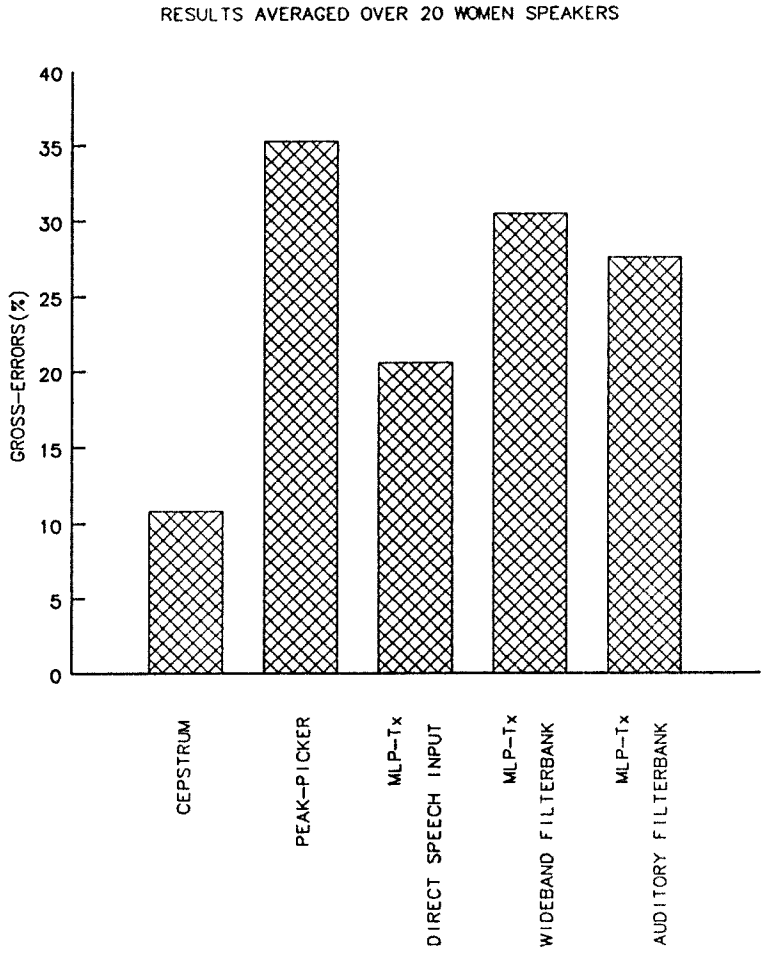


Figure 22
Bar-graph showing the gross errors generated by the six algorithms under evaluation. The comparisons were all made against an interactive reference algorithm that makes use of the output from a laryngograph. The results shown are the average performance over 20 different women speakers, with about 15 seconds of speech per speaker.

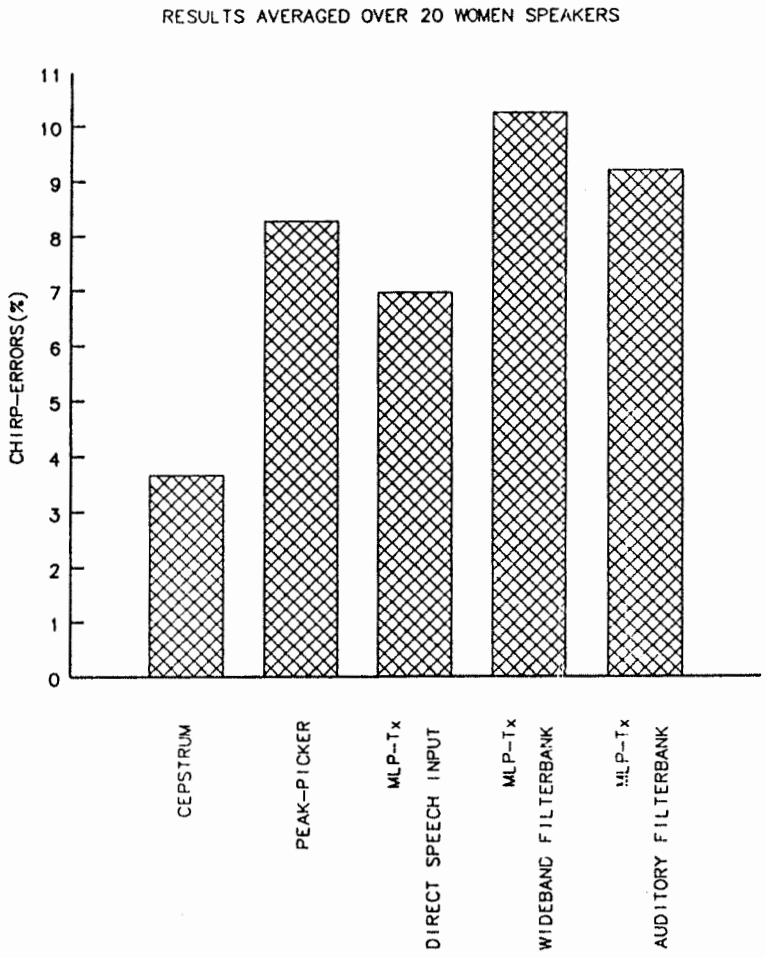


Figure 23
Bar-graph showing the chirp errors generated by the six algorithms under evaluation. The comparisons were all made against an interactive reference algorithm that makes use of the output from a laryngograph. The results shown are the average performance over 20 different women speakers, with about 15 seconds of speech per speaker.

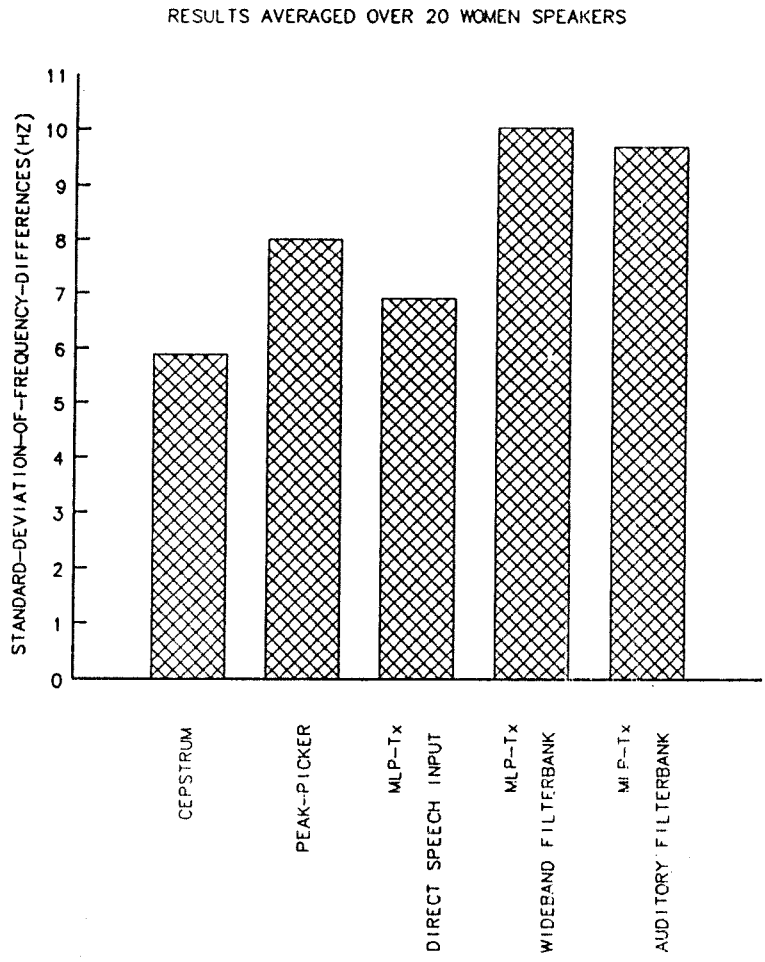


Figure 25
Bar-graph showing the standard deviation of fine frequency differences errors generated by the six algorithms under evaluation. The comparisons were all made against an interactive reference algorithm that makes use of the output from a laryngograph. The results shown are the average performance over 20 different women speakers, with about 15 seconds of speech per speaker.

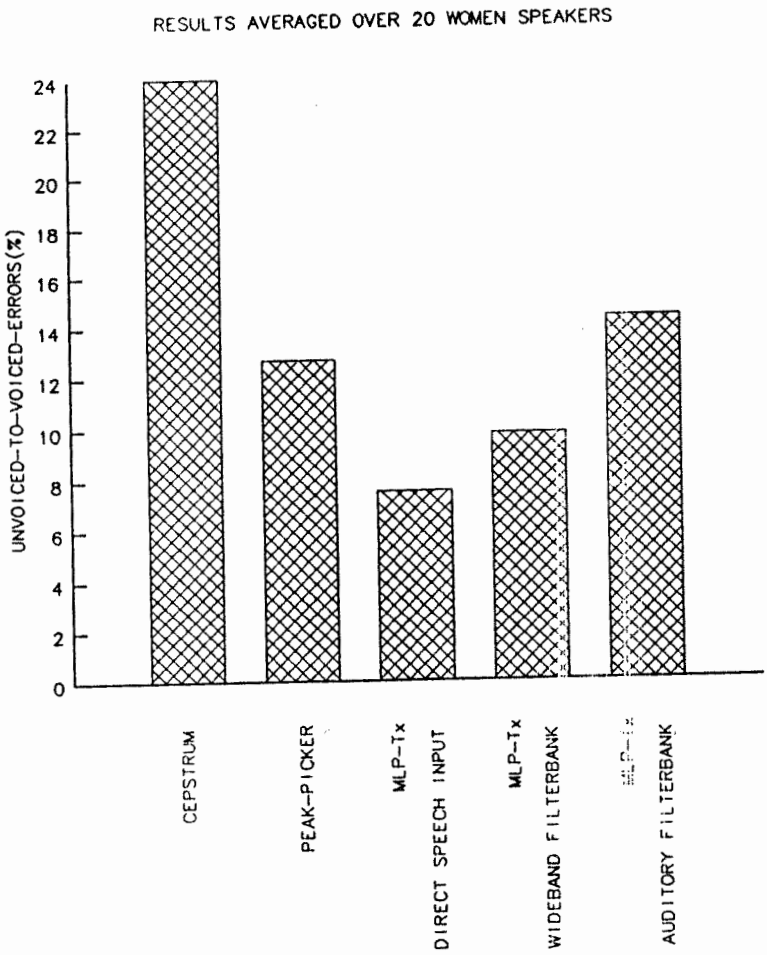


Figure 27
Bar-graph showing the unvoiced to voiced errors generated by the six algorithms under evaluation. The comparisons were all made against an interactive reference algorithm that makes use of the output from a laryngograph. The results shown are the average performance over 20 different women speakers, with about 15 seconds of speech per speaker.

Palmer helped with some of the recordings. I am grateful to Adrian Fourcin, Andrew Faulkner and Stuart Rosen for useful suggestions they made on parts of the manuscript.

References

E ABBERTON, (1974), Listener identification of speakers from larynx frequency, Proc. 8th Int. Congr. On Acoustics, London: Chapman & Hall.

E ABBERTON, (1976), A laryngographic study of voice quality, PhD thesis, University of London.

E R M ABBERTON, A J FOURCIN, S R ROSEN, J R WALLIKER, D D ABERCOMIE, (1964), English Phonetic Texts, London, Faber & Faber, (The story of Arthur the Rat, after H SWEET, (1985) A Primer of Spoken English. Oxford: Clarendon Press).

M HOWARD, B C J MOORE, E E DOUEK, & S FRAMPTON, (1985), Speech perceptual and productive rehabilitation in electro-cochlear stimulation, In R A Schindler & M Merzenich (eds), Cochlear Implants, New York: Raven Press, pp527-537.

B S ATAL, (1972), Automatic speaker recognition based on pitch contours, J. Acoust. Soc. Amer., 52, 6, pp1687-1697.

B S ATAL, (1974), Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, JASA, 55, pp.1304-1312.

B S ATAL & L R RABINER, (1976), A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition, IEEE Trans. ASSP-24, No.3, pp.201-212.

L ATLAS, R COLE, Y MUTHUSAMY, A LIPPMANN, J CONNOR, D PARK, M EL-SHARKAWI & R J MARKS, (1990), A performance comparison of trained multi-layer perceptrons and trained classification trees, Proceedings of the IEEE, Volume 78, No. 9, September 1990.

E B BAUM & D HAUSSLER, (1989), Neural computation 1, pp151-160, MIT.

G J BORDEN & K S HARRIS, (1980), Speech science primer, Baltimore, Williams and wilkins.

H BOURLARD & C J WELLEKENS, (1987), Multi-layer perceptrons and automatic speech recognition, IEEE First annual international conference on neural networks, San Diego, California.

M CHAN & F FALLSIDE, (1987), An adaptive training algorithm for back propagation networks, Cambridge University Engineering Department, CUED F-INFENG\TR2.

L COOPER & M COOPER, (1983), Introduction to dynamic programming, Pergamon press.

H DUDLEY, (1939), The Vocoder, Bell Sys. Tech. J., Vol. 45, pp1493-1509.

G FANT, (1970), Acoustic theory of speech production,Mouton, The Hague.

A FAULKNER, V BALL & A J FOURCIN, (1990), Compound speech pattern information as an aid to lipreading, Speech, Hearing and Language: Work in progress, University College London, Department of Phonetics and Linguistics, 4, pp.63-80.

J L FLANAGAN, (1972), Speech analysis, synthesis and perception, New York, Springer-Verlag.

A J FOURCIN, (1974), Laryngographic examination of vocal fold vibration, In: Ventilatory and phonatory control systems, Ed B. Wyke, London, OUP pp315-333.
A J FOURCIN & E ABBERTON, (1971), First applications of a new laryngograph, Med. and Biol. Illust., Vol. 21, pp172-182.
A J FOURCIN, E DOUEK, B C J MOORE, S R ROSEN, J R WALLIKER, D M HOWARD, E R M ABBERTON, & S FRAMPTON, (1983), Speech perception with promontory stimulation, An. New York Acad. Sci., 405, pp280-294.
Soc. Amer., 34, pp916-921.
O O GRUENZ & L O SCHOTT, (1949), Extraction and portrayal of pitch of speech sounds, J. Acout. Soc. Amer., 21, pp487-495.
W HESS, (1983), Pitch determination of speech signals, Springer-Verlag, Berlin.
W HESS & H INDEFRY, (1984), Accurate pitch determination of speech signals by means of a laryngograph, Proc. ICASSP-84, 1-4.
J HOLDSWORTH, I NIMMO-SMITH, R PATTERSON & P RICE, (1988), Implementing a GammaTone Filterbank, Annex C of the SVOS Final Report, MRC APU, Cambridge.
J N HOLMES, (1980), The JSRU 19-channel vocoder, IEE Proc., vol 127, part F, No. 1.
D M HOWARD, (1986), Digital peak-picking fundamental frequency estimation, Speech hearing and language; Work in progress, 2, London: UCL.
D M HOWARD & A J FOURCIN, (1983), Instantaneous voice period measurement for cochlear stimulation, Elect. Letters, Vol. 19, pp776-778.
D M HOWARD & G LINDSEY, (1988), Conditioned variability in voicing offsets, IEEE Trans. on ASSP, Vol. 36, No. 3.
I S HOWARD, (1991), Speech fundamental period estimation using pattern classification, PhD Thesis, To be submitted, University of London.
I S HOWARD & D M HOWARD, (1986), Quantitative comparisons between time domain speech fundamental frequency estimation algorithms, Proc. IOA, Vol 8, pp323-330.
I S HOWARD & M A HUCKVALE, (1987), The application of adaptive constraint satisfaction networks to acoustic phonetic attribute determination,Proc. Euro. Confr. Sp. Tech.
I S HOWARD & M A HUCKVALE, (1988), Speech fundamental period estimation using a trainable pattern classifier, FASE88.
I S HOWARD & J R WALLIKER, (1989), The implementation of a portable real-time multi-layer perceptron speech fundamental period estimator, Proc. Eurospeech, Paris.
M A HUCKVALE, (1988), Speech filing system, Part 1, SFS for users, Version 1.1, October 1988; Part 2, SFS for programmers, November 1988, Phonetics & Linguistics, University College London.
W Y HUANG & R LIPPMANN, (1987), Comparisons between neural net and conventional classifiers, ICNN, San Diego, CA, 21-24 June 1987.
M J HUNT & C E HARVENBERG, (1986), Generation of controlled speech stimuli by pitch-synchronous LPC analysis of natural utterances, Proc. Int. Cong. Acoust., Vol. 1, Paper A4-2, Toronto.
M J HUNT, D A ZWIERZYNSKI & R C CARR, (1989), Issues in high quality LPC analysis and Synthesis, Proc. Eurospeech, Paris.
M LEVINE & J SHEFNER, (1981), Fundamentals of sensation and perception,

Addison-Wesley, Reading, MA.
W LAWRENCE, (1953), The synthesis of speech from signals which have a low information rate, Communication Theory, Ed. W Jackson, Butterworths, London, England, pp.460-469.
R LIPPMANN, (1987), An introduction to computing with neural nets, IEEE ASSP Magazine April 1987.
J D MARKEL, (1972), The SIFT algorithm for fundamental frequency estimation, IEEE Trans. AU-20, pp367-377.
M McGRATH & Q SUMMERFIELD, (1985), Intermodal timing relations and audio-visual speech recognition by normal-hearing adults, JASA, 77, pp678-685.
C H MAYS, (1963), Adaptive threshold logic, PhD Thesis, Technical report 1557-1, Stanford Electron. Labs, Stanford, CA, April 1963.
P MERMELSTEIN, (1977), On detecting nasals in continuous speech, JASA, Vol. 61 pp581.
A M NOLL, (1964), Short time spectrum and cepstrum techniques for vocal pitch detection, J. Acoust. Soc. Amer., 36, pp296-302.
A M NOLL, (1967), Cepstrum Pitch determination, J. Acoust. Soc. Amer., 41, pp293-309.
A V OPPENHEIM & R W SCHAFER, (1975), Digital signal processing, Prentice-hall.
D J B PEARCE & L C WHITAKER, (1986), Reference formant analysis, Int. confr. on Speech Input/Output; Techniques and applications, London.
S M PEELING & J S BRIDLE, (1986), Experiments with a learning network for a simple phonetic recognition task, Proc. I.O.A. Confr., Windemere.
L R RABINER, (1977), One the use of autocorrelation analysis for pitch detection, IEEE Trans. ASSP-25, 23-33.
L R RABINER, M H CHENG, A E ROSENBERG & C A McGONEGAL, (1976), A comparative study of several pitch detection algorithms, IEEE Trans. ASSP-24, 5, pp399-413.
S ROSEN & A J FOURCIN, (1983), When less is more, Work in progress, Departments of Phonetic and Linguistics, University College, London.
S ROSEN, J R WALLIKER, A FOURCIN & V BALL, (1988), "A microprocessor based acoustic hearing aid for the profoundly impaired listener", J. Rehab. Res. Dev., Vol. 24, pp. 239-260.
A E ROSENBERG & M R SAMBUR, (1975), New techniques for automatic speaker verification, IEEE Trans. ASSP-23, 2, pp169-176.
D E RUMELHART, G E HINTON & R J WILLIAMS, (1986), Learning internal representations by error propagation, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1, D E Rumelhart & J L McClelland (Eds.), Cambridge, MA: MIT Press, pp318-362.
M M SONDHI, (1968), New methods of pitch extraction, IEEE Trans. AU-16, pp262-266.
J R WALLIKER, E E DOUEK, S FRAMPTON, E A ABBERTON, A J FOURCIN, D M HOWARD, S NEVARD, S ROSEN & B C J MOORE, (1985), Physical and surgical aspects of external single channel electrical stimulation of the totally deaf, Schindler R.A. & Merzenich, M.M. (Eds), Cochlear Implants, Raven Press, New York.
J R WALLIKER & I S HOWARD, (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.
J R WALLIKER, S ROSEN & A FOURCIN, (1986), Speech pattern prostheses for the profoundly and totally deaf, IEE Confr. Pub. No. 258, Int. Conf. Speech Input/Output, London, England, pp194-199.
B WIDROW, (1987), Adaline and Madaline - 1963 Plenary Speech, Volume 1: Proc. IEEE, 1st Int. Confr. on Neural Networks, San Diego, CA, pp143-158.
B WIDROW & M A LEHR, (1990), 30 years of adaptive neural networks: Perceptron, Madaline, and Backpropagation, Proceeding IEEE, September 1990.