

SPEECH FUNDAMENTAL PERIOD ESTIMATION USING A TRAINABLE PATTERN CLASSIFIER

I. S. Howard and M. A. Huckvale

Department of Phonetics and Linguistics,
University College London, Gower Street, London WC1E 6BT.

INTRODUCTION

This paper describes a pattern recognition algorithm for the location of the points of vocal fold closure in a noisy speech signal. The algorithm uses a multi-layer perceptron (MLP) classifier with inputs from a window on the speech signal, and an output signifying the presence of a vocal fold closure at the centre of the window. The location of the vocal fold closures, or the fundamental period epoch markers are given the name T_x , and our perceptron algorithm the name MLP-Tx. In this paper, we describe the MLP-Tx algorithm in relation to other methods for fundamental frequency estimation, and compare its performance with another algorithm for T_x estimation from speech. The MLP-Tx algorithm is shown to have good performance in noise and potential application in a signal-processing hearing-aid.

APPROACHES TO FUNDAMENTAL FREQUENCY ESTIMATION

Speech fundamental frequency estimation algorithms can be broadly classed as frequency domain, time domain, hybrids of the two domains, and devices that make use of direct evidence of vocal fold activity. A comprehensive review is given in [1]. Frequency domain algorithms operate on spectral representations of the signal, and in general have a degree of *smoothing* in their fundamental frequency estimates. Time-domain algorithms, on the other hand, estimate fundamental periods directly, and do not perform frequency smoothing. The MLP-Tx algorithm described in this paper is of this type. Time-domain methods have less inherent delay and retain irregularities in vocal fold vibration, which have been shown to be beneficial in the application of fundamental frequency algorithms in hearing aids [2].

In this paper we compare the new algorithm with (i) the output from a Laryngograph [3], (ii) the output from a time-domain algorithm developed for hearing aids called the "peak-picker" [4], and (iii) the frequency-domain algorithm known as SIFT [5].

DESCRIPTION OF THE MLP ALGORITHM

The MLP provides a means of performing complex pattern recognition tasks. It is capable of implementing non-linear transformations found from example in an iterative training procedure [6]. Such a classifier has an advantage over a classical algorithm like the Bayes classifier for normal patterns [7] because it does not assume the form of the distributions of the pattern classes *a priori*. Previous experience with speech data suggests that such assumptions may not always be valid [8].

In the MLP-Tx system, the input to the classifier consists of a 20ms window on a high-resolution wideband filter-bank output. There are two layers of "hidden" units and a single output unit that is trained to switch on at the instants of vocal fold closure.

SPEECH FUNDAMENTAL PERIOD ESTIMATION USING A TRAINABLE PATTERN CLASSIFIER

The filterbank was designed to ensure that the individual vocal fold closures could be resolved from its output. It comprised nine second order IIR Butterworth filters with -3dB points of 50-300Hz, 300-600Hz, 600-900Hz, 900-1200Hz, 1200-1600Hz, 1600-2000Hz, 2000-2400Hz, 2400-2800Hz and 2800-3300Hz. The outputs from the filters were half-wave rectified and then smoothed by means of a second order low-pass Butterworth filter with a cut-off frequency of 1KHz. Half-wave rectification was employed as opposed to full wave because the latter has the effect of doubling the periodicity of the filter outputs. The smoothed outputs were then down-sampled (from a 10KHz input sampling rate) to 2KHz. The input to the classifier was 41 frames long comprising 369 units, there were two hidden layers of 10 units and one output unit. (Note: this configuration was obtained by trial and error and is probably not yet optimum). Adjacent layers were fully interconnected.

The output signal used for training (and as a reference in evaluation) was obtained automatically by means of a relatively simple algorithm that makes use of the output of a Laryngograph.

TRAINING OF THE ALGORITHM

Five male speakers were recorded reading the "Rainbow" passage [9] in an anechoic room using a high-quality microphone and a Laryngograph. The material was recorded twice for each speaker; the first being used for training, the second for testing. The signals were low-pass filtered at 5kHz and sampled at 10kHz using a 12-bit A/D converter. The data was then contaminated with noise at levels of 0dB and 20dB SNR. The noise signal was recorded in the College refectory at lunchtime ("canteen" noise), and includes impulsive noise and background conversations. The SNR was defined in terms of the maximum power found in any 500ms window. Two separate identical networks were trained for operation in the two different noise conditions. The training of the MLP networks was performed by 10 passes over the input data with learning parameters $\alpha=0.9$ and $\eta=0.05$. This was equivalent to about 3 million pattern vector presentations. The MLP algorithm was written in 'C' and ran under Unix on a Masscomp MC5600 series computer.

Figure 1 shows a close up of part of the training material. Item A is the contaminated speech, item B the Laryngograph (Lx) waveform. The input data to the network generated by the filterbank is shown as a grey-level display C. It can be seen that temporal variation concerning the excitation is retained. Item D is the reference Tx markers generated from B. The Tx markers are used to train the output of the MLP network.

RESULTS

Figure 2 shows a typical output of the MLP network for the two noise conditions. It can be seen that the MLP achieves a remarkable similarity to the target Tx on which it was trained. It is to be noted that the performance is not degraded by much even in the 0dB SNR case. The epoch marker locations due to the peak-picker are also shown. It can be seen that it experiences considerable difficulty in the 0dB SNR case.

Figure 3 shows some typical output data represented in terms of a set of frequency contours (Fx). For the time-domain algorithms the conversion is made by calculating the reciprocal of the time period. Such an Fx contour is therefore free from any smoothing, and vocal fold closure irregularity is preserved. It can be seen that the MLP algorithm operates very well, the jitter in frequency being due to the coarse time quantization of the Tx. The SIFT algorithm gives reasonable contours in the louder regions of the

SPEECH FUNDAMENTAL PERIOD ESTIMATION USING A TRAINABLE PATTERN CLASSIFIER

signal. The performance of the peak-picker is clearly unsatisfactory with such noisy speech.

PERFORMANCE EVALUATION

The quantitative performance of the algorithm is given in terms of two measures: the *Receiver Operating Characteristic* (ROC) and the *Relative Jitter Distribution* [10]. The ROC is a plot of correct classifications against the number of false alarms as a varying threshold value is applied to the classifier output. This gives a clear visual indication of the ability of the algorithm to detect the vocal fold closures. The jitter histogram gives an indication of the positioning of the located vocal fold closures with respect to the reference locations. It gives an indication of how precisely in time the vocal fold closures are detected.

Figure 4 (a,b,c,d) shows the jitter distributions for the peak-picker and the MLP-Tx algorithm for both 20dB and 0dB SNR speech of one speaker. Notice that the effect of noise on the MLP-Tx algorithm is quite small, whereas it has much more effect on the peak-picker. This is consistent with the observations of the Fx plots. Notice also that the MLP-Tx algorithm operating on 0dB SNR speech is better than the peak-picker operating on 20dB SNR speech.

Figure 5 shows the ROC's for the MLP-Tx algorithm and the peak-picker in the 20dB SNR condition. Notice that the ROC for MLP-Tx represents performance much closer to ideal performance than does that of the peak-picker. This indicates that MLP-Tx is better at detecting Tx markers than the peak-picker.

CONCLUSIONS

The MLP-Tx system performs creditably on the noisy speech used in the experiment and better than the alternative time-domain system. The performance of the system on the test data was promising for applications of such an algorithm in hearing-aids of the type worked on at UCL [11] although in the current implementation there is a delay of 10ms between input and output, without taking into account any delay introduced by processing. The algorithm can also only locate Tx values to within 0.5ms (i.e. 5% quantisation noise at 100Hz)

A particular point of interest of the system was how it has generalised the solution to the problem from the training material. Firstly, performance on the training data and the test data were practically equivalent. Secondly, the algorithm located vocal fold closures correctly even when the Lx signal had not detected them, at for example the end of utterances [12].

In terms of further developments of the algorithm:

- (a) Clearly the device must also be trained and tested with female speech.
- (b) Currently the "receptive field" for the Tx feature detector is rectangular. A better arrangement might be to have higher resolution around the centre of the window and gradually less at the sides. In this way it is hoped that contextual information may be made available to the recognition process without dramatically increasing the computational load.
- (c) The use of a non-symmetrical window on the speech that uses a lot of past evidence, but little future evidence, should be investigated. It is hoped that such a scheme will retain the high

SPEECH FUNDAMENTAL PERIOD ESTIMATION USING A TRAINABLE PATTERN CLASSIFIER

- performance of the current system, but with a shorter delay between input and output.
- (d) The present filterbank is roughly based on a wideband spectrogram. Consequently the filter bandwidth remains fixed with increasing centre frequency. It may prove advantageous to incorporate a constant Q filterbank, with logarithmically spaced filters. One possibility would be to use an auditory-model filterbank.

ACKNOWLEDGEMENTS

This work was supported by MRC studentship RS-85-2 and by a SERC fellowship. The software version of the peak-picker was implemented by D. M. Howard.

REFERENCES

- [1] W HESS, *Pitch determination of speech signals*, Springer-Verlag, Berlin, (1983).
- [2] E ABBERTON, A J FOURCIN, S R ROSEN, J R WALLIKER, D M HOWARD, B C J MOORE, E E DOUEK, & S FRAMPTON, 'Speech perceptual and productive rehabilitation in electro-cochlear stimulation', In R A Schindler & M Merzenich (eds), *Cochlear Implants*, New York: Raven Press, 527-537, (1985).
- [3] A J FOURCIN & E ABBERTON, 'First applications of a new laryngograph', *Med. and Biol. Illust.*, 172-182, (1971).
- [4] D M HOWARD, 'Digital peak-picking fundamental frequency estimation', *Speech hearing and language; Work in progress*, 2, London: UCL, (1986).
- [5] L R RABINER & R W SCHAFFER, *Digital processing of speech signals*, Prentice-Hall, (1978).
- [6] D E RUMELHART & J L McCLELLAND, *Parallel distributed processing*, MIT press, (1985).
- [7] J T TOU & R C GONZALEZ, *Pattern recognition principles*, Addison-Wesley, (1976).
- [8] I S HOWARD & M A HUCKVALE, 'The application of adaptive constraint satisfaction networks to acoustic phonetic attribute determination', *Proc. Euro. Confr. Sp. Tech.*, (1987).
- [9] P MERMELSTEIN, 'On detecting nasals in continuous speech', *JASA*, 61 p581, (1977).
- [10] I S HOWARD & D M HOWARD, 'Quantitative comparisons between time domain speech fundamental frequency estimation algorithms', *Proc. IOA*, Vol 8, 323-330, (1986).
- [11] A J FOURCIN, E DOUEK, B C J MOORE, S R ROSEN, J R WALLIKER, D M HOWARD, E R M ABBERTON, & S FRAMPTON, 'Speech perception with promontory stimulation', *An. New York Acad. Sci.*, 405, 280-294, (1983).
- [12] D M HOWARD & G LINDSEY, 'Conditioned variability in voicing offsets', *IEEE Trans. on ASSP*, Vol. 36, No. 3, (1988).

SPEECH FUNDAMENTAL PERIOD ESTIMATION USING A TRAINABLE PATTERN CLASSIFIER

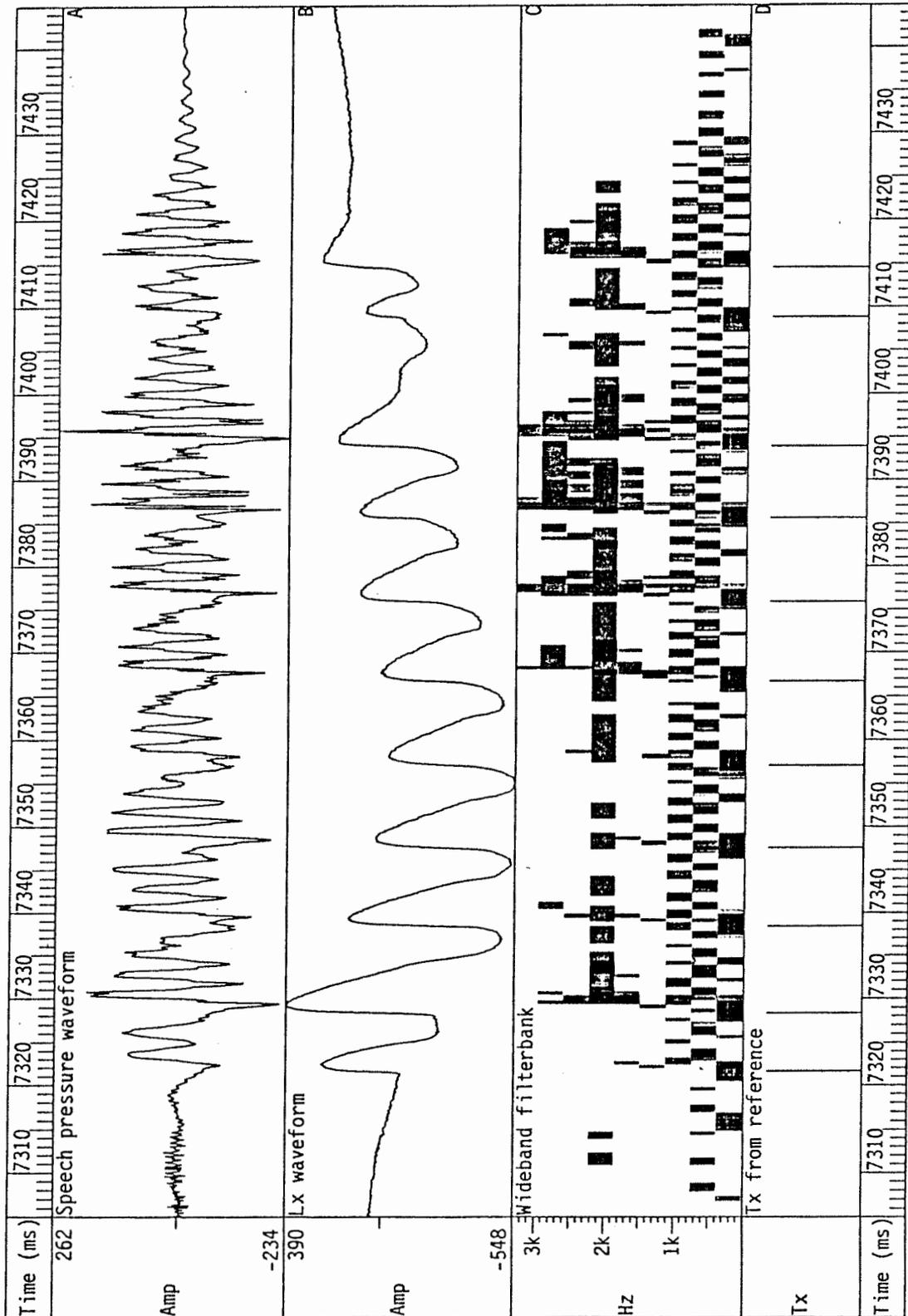


Figure 1.

SPEECH FUNDAMENTAL PERIOD ESTIMATION USING A TRAINABLE PATTERN CLASSIFIER

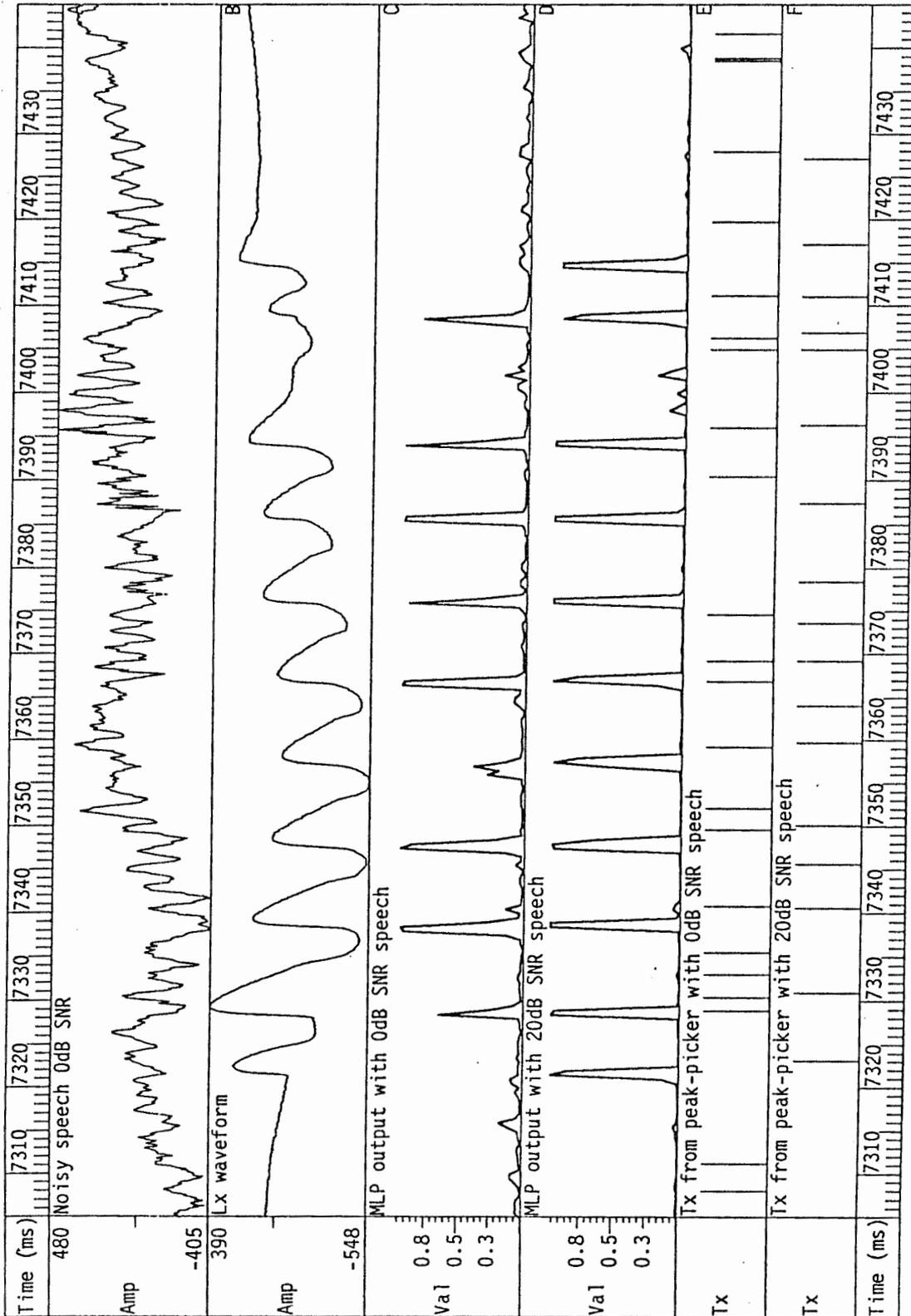


Figure 2.

SPEECH FUNDAMENTAL PERIOD ESTIMATION USING A TRAINABLE PATTERN CLASSIFIER

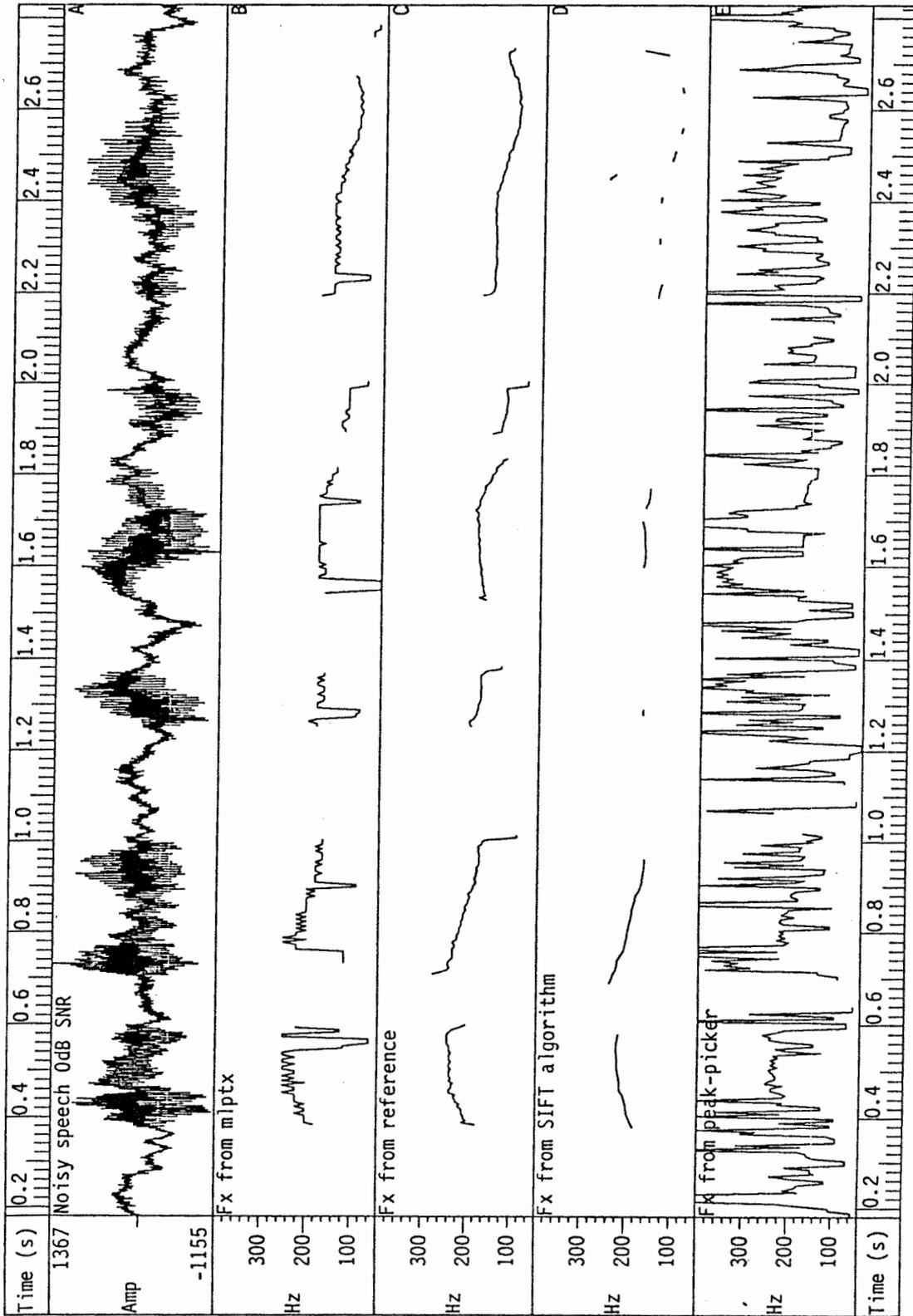
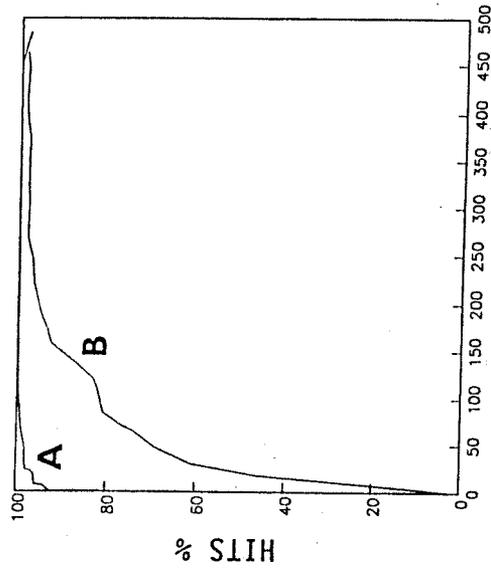
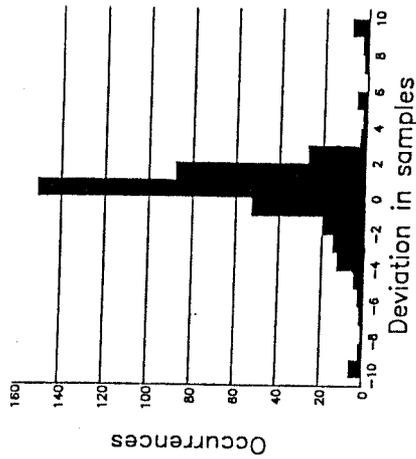


Figure 3.

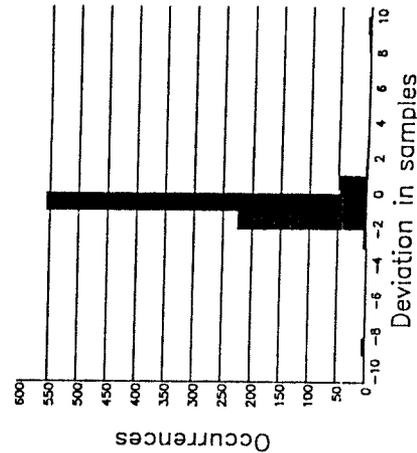


FALSE ALARMS

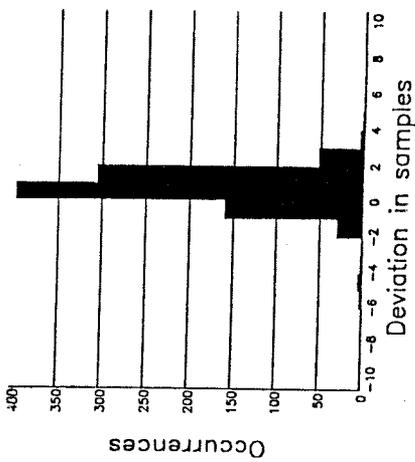
Figure 5.
Curve A is ROC for MLP-Tx.
Curve B is ROC for Peak-picker.
In both cases the SNR was 20dB.



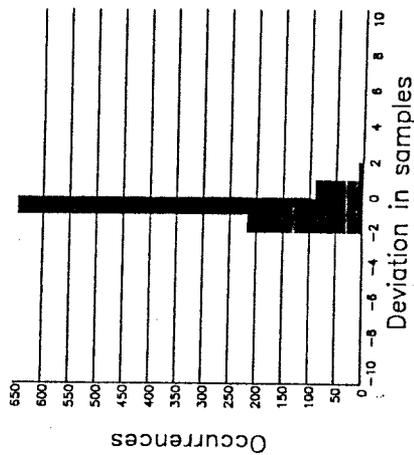
b) Peak-picker, 0dB SNR.



d) MLP-Tx, 0dB SNR.



a) Peak-picker, 20dB SNR.



c) MLP-Tx, 20dB SNR.

Figure 4. Jitter histograms for the peak-picker and MLP-Tx.
One sample represents 0.5 ms.