# NEURAL NET CLASSIFIERS FOR ROBUST SPEECH RECOGNITION UNDER NOISY ENVIRONMENTS

**K.K. Paliwal**
Computer Systems and Communications Group
Tata Institute of Fundamental Research
Homi Bhabha Road, Bombay-400005. India
(Current address: Acoustics Research Dept.,
AT&T Bell Labs., Murray Hill. NJ 07974. USA)

**ABSTRACT** — Recently, the multi-layer perceptron (MLP) type of neural net classifiers have been used extensively for speech recognition. However, their performance has not been studied so far for noisy speech. In this paper, the MLP classifier is studied for the recognition of noisy speech and its performance is compared with the conventional pattern classifiers (such as the maximum likelihood (ML) classifier and the k-nearest neighbor (kNN) classifier). The linear prediction (LP) parameters derived through 10-th order LP analysis are used here as the recognition parameters. Different LP parametric representations are compared as to their recognition performance with the MLP classifier and the cepstral coefficient representation is found to be the best parametric representation. Using 10 cepstral coefficients as recognition parameters, performance of the MLP classifier is found to be significantly better than that of the ML and the kNN classifiers for noisy speech. Use of 15 cepstral coefficients (obtained by extrapolating the 10 cepstral coefficients) improves the recognition performance of the MLP classifier for noisy speech further.

## 1. INTRODUCTION

Significant progress has been made in the area of speech recognition over the last few years. It is now possible to recognize the spoken words with almost 100% accuracy from a small vocabulary (size less than 100 words) [1]. But, this performance can be achieved only when the recognizer is used in noise-free (clean) environments. However, in practice, the recognizer has to cope with speech that is corrupted due to the presence of background noise. If the recognizer can be trained under the background noise condition which it has to encounter during its testing phase of operation, it can give satisfactory recognition performance [2]. However, in practice, it is not always possible to have matched noise conditions during training and testing phases due to various reasons. For example, it may be due to the variability in the background noise conditions. As a result, the recognizer has to operate under the background noise condition which was not present during its training phase. This mismatch in background noise conditions can cause severe degradation in the recognition performance. For example, it has been reported that the recognizer that performs almost perfectly for clean speech (recorded under laboratory conditions) can recognize spoken words with only 30% accuracy at 5 dB signal-to-noise ratio (SNR) [3].

The problem arising due to mismatched noise conditions is very difficult to handle. Some studies have been reported in the literature to deal with this problem [3-6]. These studies either use robust estimation methods to obtain better estimates of the parameters from noisy speech [3,4] or modify the distance measure used in the pattern classifier to compensate for the additive white noise distortion [5,6]. In the present paper, we study the problem of speech recognition under the mismatched noise conditions using the neural network for pattern classification.

Since the introduction of the back propagation algorithm for training the multi-layer perceptron (MLP) classifiers by Rumelhart et al. [7], the neural networks have become very popular in the area of speech recognition [8]. Besides their computational advantages, the MLP classifiers offer two main advantages over the conventional linear pattern classifiers. First, the MLP classifiers do not assume any distribution for the probability density functions. Second, these classifiers are capable of providing nonlinear mappings [9]. Because of these advantages, the MLP classifiers result in better speech recognition performance than the conventional pattern classifiers [10]. The MLP classifiers have been studied extensively for the recognition of clean speech [9], but their performance has not been evaluated so far for the recognition of noisy speech. The aim of the present paper is to study the MLP classifier for the recognition of noisy speech and compare its performance with the conventional pattern classifiers.

In order to study the performance of the MLP classifier for the recognition of noisy speech, we use here the single-frame vowel recognition task as a test bed. Vowel recognition experiments are conducted here in both speaker dependent and speaker independent modes. Speech signal is represented here in terms linear prediction (LP) parameters. Since the LP analysis can result in a number of different parametric representations each of which providing equivalent information about the speech spectral envelope, it becomes important which LP parametric representation should be used with the MLP classifier for speech recognition. In the present paper, we compare different LP parametric representations and show that the MLP classifier performs best with the cepstral coefficient representation. Using the cepstral parameters, the MLP classifier is studied at different SNR conditions and its performance is compared with the two conventional pattern classifiers [11]: 1) the maximum likelihood (ML) classifier and 2) the k-nearest neighbor (kNN) classifier. The ML classifier is a parametric classifier. It assumes a parametric form for the class-conditional probability density functions (or, the likelihood functions). The parameters of these likelihood functions are estimated from the data in training set. The ML classifier classifies the test pattern in favor of a class having maximum likelihood. In the present paper, the likelihood functions in the ML classifier are assumed to be multivariate Gaussian. The kNN classifier is a nonparametric classifier providing suboptimal performance with respect to the Bayes classifier. But, with an unlimited amount of training patterns it has a probability of error which is always less than twice the Bayes probability of error [11]. This classifier classifies the test pattern in favor of a class most heavily represented among its k nearest neighbors. In the present paper, the value of k used with the kNN classifier is set to 3. But, we have also studied other values of k and found similar results.

The paper is organized as follows. Section 2 describes the speech data base used in the present study. Different LP parametric representations are studied in Section 3 and their comparative recognition performance is reported for both the speaker-dependent and the speaker-independent modes using the MLP, the ML and the kNN classifiers. Section 4 describes experiments to evaluate the MLP classifier with respect to the ML and the kNN classifiers for noisy speech at different SNR conditions. Conclusions are reported in Section 5.

## 2. DATA ACQUISITION AND PREPROCESSING

The speech data base used in the present study consists of 900 utterances, having 30 repetitions of different /b/-vowel-/b/ syllables, spoken by three speakers (two male and one female). Recordings of these utterances is done in an ordinary office room. The speech signal is low-pass filtered at 4 kHz and digitized at 10 kHz sampling rate. The steady-state part of the vowel segment is manually located for each of the 900 utterances and a 20 ms segment is excised from its center. A

CH2847-2/90/0000-0429 $1.00 © 1990 IEEE

10-th order LP analysis is performed for each 20 ms segment using the autocorrelation method (with 20 ms Hamming window and without preemphasis).

In order to perform speaker-dependent vowel recognition experiments, the first 15 repetitions of all the vowels for each speaker are used for training and the remaining 15 repetitions for testing. For speaker-independent vowel recognition experiments, the first 15 repetitions from each of the three speakers are pooled together to have 45 repetitions for training. The remaining 45 repetitions from the three speakers are used for testing.

In the present study, the speech signal is assumed to be corrupted by the addition of white Gaussian noise. This noise is generated on computer through a pseudo-random number generator. It is added to each frame of vowel sound to make its SNR equal to the desired value.

## 3. SELECTION OF THE BEST LP PARAMETRIC REPRESENTATION

The LP analysis of speech can provide a number of parametric representations such as the cepstral coefficient representation, the LP coefficient representation, the reflection coefficient representation, etc. These parametric representations are related to each other through nonlinear transformations which are reversible in nature. Though each of these representations contain equivalent information about the LP spectral envelope, these parametric representations can lead to different recognition performances when used with a conventional pattern classifier [12]. Since we intend to use LP parameters in the present study with the MLP classifier, it is important to know whether different LP parametric representations lead to different recognition performances with the MLP classifier. If so, which is the best parametric representation for the MLP classifier.

Since the MLP classifier is capable of providing nonlinear mappings, it has been argued in the literature [13] that selection of a particular parametric representation may not be very crucial for the MLP classifier as it may lead to similar performance with different representations. However, we show in this section that this is not the case. Different parametric representations differ in a significant manner when used with the MLP classifier. In order to show it, we consider first only the following two LP parametric representations: 1) the cepstral coefficient representation and 2) the LP coefficient representation. We train the MLP classifier for each of these two representations in speaker-dependent mode for the first speaker.

We use here an MLP classifier with three layers (input layer, output layer and one hidden layer). The hidden layer has 16 nodes. Since there are 10 vowel classes, the output layer has 10 nodes. The number of nodes in the input layer is the same as the number of LP parameters used for recognition (10 in the present case). The back propagation algorithm [7] is used here without the momentum term for training the MLP classifier. This algorithm is executed here in block-mode; i.e., in each iteration, all the data in the training set is used in one shot to update the connection weights. In the back propagation algorithm, the step size (or, the adaptation constant) is an important parameter. It decides about the convergence of the algorithm and controls its convergence rate. The algorithm can be made to converge for small values of step size, but its convergence will be slow. For higher values of step size, the algorithm may converge faster. But, in this case, its convergence is not always guaranteed. However, there is no way to know the value of step size a priori. It has to be found out through experimentation.

Here, we train the MLP classifier using two different values (1 and 18) for the step size and study the performance of the cepstral and the LP coefficient representations for both the training and the test data sets. Results are shown in Figures 1 and 2. It can be seen from these figures that the cepstral and the LP coefficient representations differ from each other in two different ways: 1) The convergence is faster for the cepstral coefficient representation than for the LP coefficient representation, and 2) The cepstral coefficient representation leads to significantly better recognition performance on test data than the LP coefficient representation, though both the representations result in the same 100% recognition accuracy on the training data set. Results for the speaker-independent vowel recognition experiment are shown in Figures 3 and 4. Similar observations can be made from these figures.

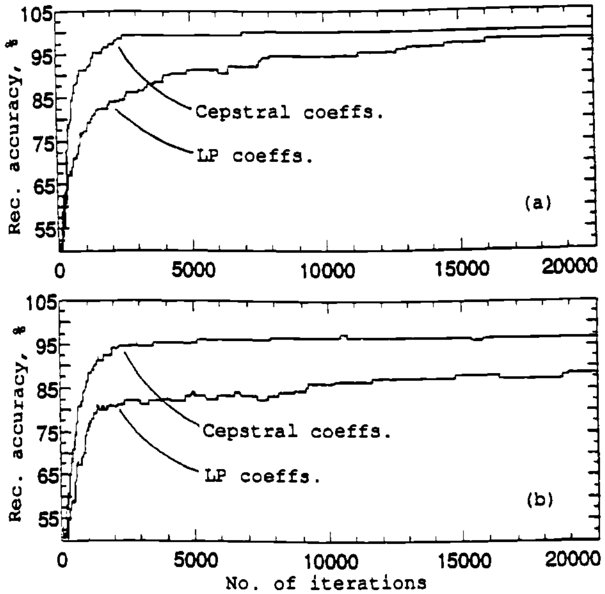

Fig. 1: Recognition performance of the MLP classifier on (a) training data set and (b) test data set in the speaker-dependent mode with step size=1.

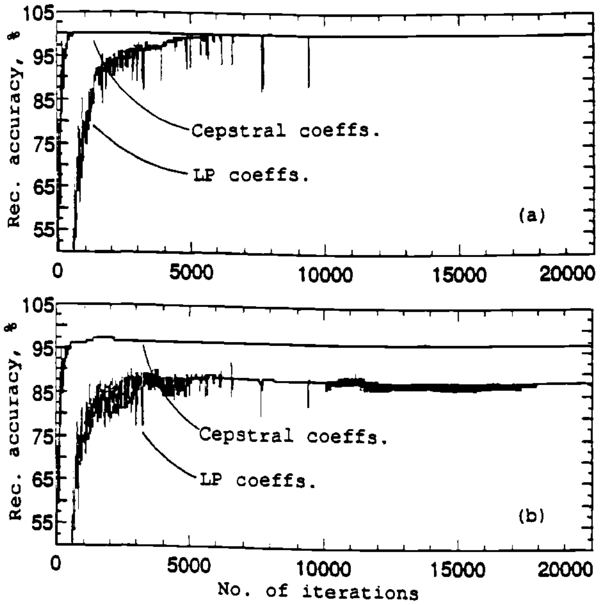

Fig. 2: Recognition performance of the MLP classifier on (a) training data set and (b) test data set in the speaker-dependent mode with step size=18.

Thus, we have seen that the proper choice of the LP parametric representation is important for the MLP classifier as the better representation leads to faster convergence and better generalizing capability on the test data set (different from the training data set). In order to find the best LP parametric representation, we study nine different representations formed from the following LP parameters: 1) LP coefficients, 2) cepstral coefficients, 3) reflection coefficients (RCs), 4) log area ratios, 5) inverse sine of RCs, 6) area coefficients, 7) impulse response of the LP synthesis filter, 8) autocorrelation coefficients, and 9) line spectral-pair frequencies (LSFs). These nine parametric representations are evaluated on the test data set with the MLP, the ML and the kNN classifiers and the results are shown in Table 1 for the speaker-dependent case and in Table 2 for the speaker-independent case. Since

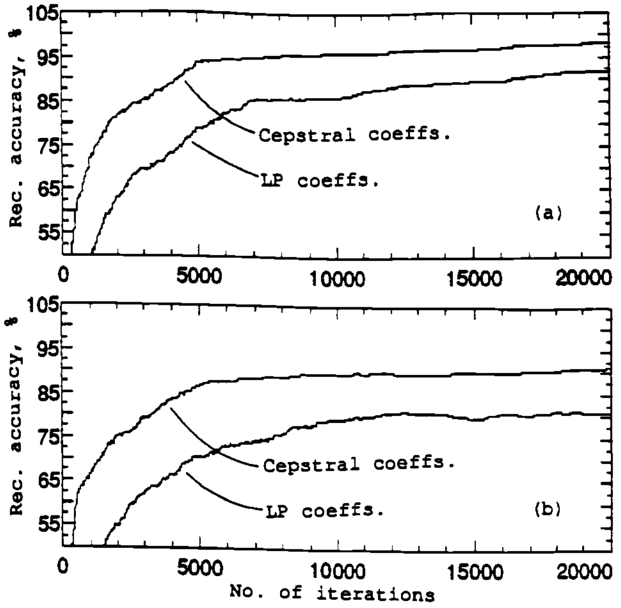

Fig. 3: Recognition performance of the MLP classifier on (a) training data set and (b) test data set in the speaker-independent mode with step size=1.

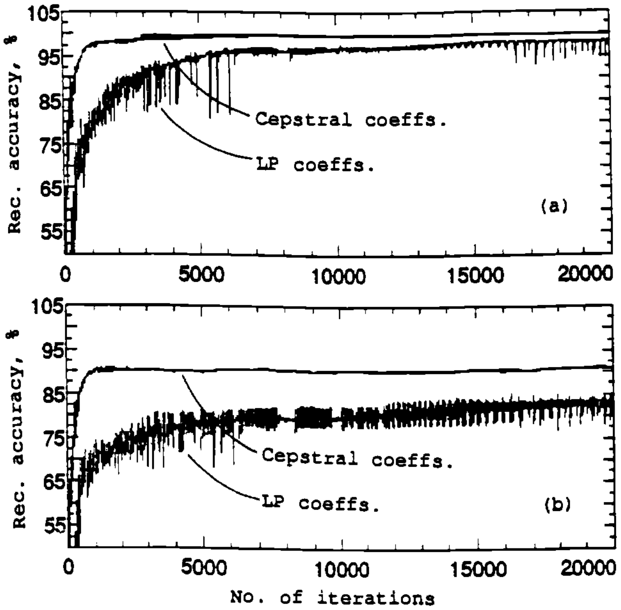

Fig. 4: Recognition performance of the MLP classifier on (a) training data set and (b) test data set in the speaker-independent mode with step size=18.

Table 1: Recognition performance of the MLP, the ML and the kNN classifiers for clean speech with different LP parametric representations in the speaker-dependent mode.

| Parameters | Recognition accuracy (in %) with | | |
|---|---|---|---|
| | MLP classifier | ML classifier | kNN classifier |
| LP coeffs. | 82.9 | 67.8 | 71.5 |
| Cepstral coeffs. | 91.1 | 88.5 | 89.6 |
| Reflection coeffs. | 84.6 | 83.6 | 85.6 |
| Log area ratios | 88.2 | 83.1 | 88.2 |
| Inverse sine RCs | 87.3 | 83.6 | 87.8 |
| Area coeffs. | 62.7 | 73.6 | 59.1 |
| Impulse Response | 91.3 | 79.2 | 85.3 |
| Auto. coeffs. | 82.9 | 78.0 | 76.9 |
| LSFs | 90.5 | 92.2 | 86.5 |

Table 2: Recognition performance of the MLP, the ML and the kNN classifiers for clean speech with different LP parametric representations in the speaker-independent mode.

| Parameters | Recognition accuracy (in %) with | | |
|---|---|---|---|
| | MLP classifier | ML classifier | kNN classifier |
| LP coeffs. | 80.9 | 88.0 | 69.6 |
| Cepstral coeffs. | 91.3 | 92.7 | 88.2 |
| Reflection coeffs. | 77.3 | 82.9 | 85.1 |
| Log area ratios | 84.0 | 85.3 | 87.6 |
| Inverse sine RCs | 80.7 | 84.4 | 87.6 |
| Area coeffs. | 67.1 | 74.0 | 58.0 |
| Impulse Response | 89.8 | 89.2 | 85.3 |
| Auto. coeffs. | 76.0 | 64.0 | 74.9 |
| LSFs | 85.6 | 90.2 | 86.0 |

Table 3: Recognition performance of the MLP, the ML and the kNN classifiers for noisy speech (SNR=25 dB) with different LP parametric representations in the speaker-dependent mode.

| Parameters | Recognition accuracy (in %) with | | |
|---|---|---|---|
| | MLP classifier | ML classifier | kNN classifier |
| LP coeffs. | 51.3 | 35.3 | 39.8 |
| Cepstral coeffs. | 79.1 | 69.3 | 70.9 |
| Reflection coeffs. | 57.6 | 61.1 | 61.8 |
| Log area ratios | 68.7 | 59.3 | 73.1 |
| Inverse sine RCs | 64.2 | 60.7 | 66.7 |
| Area coeffs. | 60.0 | 55.6 | 52.0 |
| Impulse Response | 83.1 | 73.8 | 79.8 |
| Auto. coeffs. | 82.4 | 77.3 | 76.7 |
| LSFs | 75.3 | 77.4 | 64.4 |

Table 4: Recognition performance of the MLP, the ML and the kNN classifiers for noisy speech (SNR=25 dB) with different LP parametric representations in the speaker-independent mode.

| Parameters | Recognition accuracy (in %) with | | |
|---|---|---|---|
| | MLP classifier | ML classifier | kNN classifier |
| LP coeffs. | 60.9 | 77.1 | 42.2 |
| Cepstral coeffs. | 84.0 | 74.2 | 69.8 |
| Reflection coeffs. | 57.8 | 64.0 | 62.7 |
| Log area ratios | 64.4 | 68.9 | 68.2 |
| Inverse sine RCs | 60.4 | 65.6 | 64.7 |
| Area coeffs. | 58.4 | 60.0 | 55.3 |
| Impulse Response | 81.8 | 61.8 | 81.3 |
| Auto. coeffs. | 75.8 | 64.7 | 74.9 |
| LSFs | 80.9 | 74.9 | 65.6 |

the amount of the training data is relatively small for the speaker-dependent case, we use for this case the diagonal covariance matrix with the ML classifier; while the full covariance matrix is used with the ML classifier for the speaker-independent case. It can be seen from these tables that the MLP classifier results in the best recognition performance for both the speaker-dependent and the speaker-independent modes. Also, the conventional classifiers (ML and kNN) perform best with the cepstral coefficient representation, confirming our results reported in an earlier paper [12].

These nine LP parametric representations are also evaluated for noisy speech (SNR=25 dB) and their performances with the three classifiers (MLP, ML and kNN) are listed in Table 3 for the speaker-dependent mode and in Table 4 for the speaker-independent mode. We can see from these tables that the cepstral coefficient representation gives, in general, best results for the noisy speech. However, comparison of Table 3 with Table 1 (and Table 4 with Table 2) reveals that the autocorrelation coefficient and the impulse response representations are two other representations quite robust to noisy speech.

## 4. RECOGNITION RESULTS FOR NOISY SPEECH

In this section, the MLP classifier is studied for the recognition of noisy speech at different SNR conditions. The classifier is trained using the clean speech and tested on the noisy speech (i.e., under mismatched noise conditions). Performance of the MLP classifier is compared with

Table 5: Recognition performance of the MLP, the ML and the kNN classifiers for noisy speech in the speaker-dependent mode using 10 cepstral coefficients.

| SNR | Recognition accuracy (in %) with | | |
|---|---|---|---|
| | MLP classifier | ML classifier | kNN classifier |
| Clean | 91.1 | 88.5 | 89.6 |
| 35 | 89.6 | 84.9 | 85.6 |
| 30 | 84.0 | 79.1 | 80.4 |
| 25 | 79.1 | 69.3 | 70.9 |
| 20 | 71.8 | 63.3 | 62.2 |
| 15 | 63.8 | 56.9 | 52.2 |
| 10 | 54.2 | 47.8 | 43.1 |

Table 6: Recognition performance of the MLP, the ML and the kNN classifiers for noisy speech in the speaker-independent mode using 10 cepstral coefficients.

| SNR | Recognition accuracy (in %) with | | |
|---|---|---|---|
| | MLP classifier | ML classifier | kNN classifier |
| Clean | 91.3 | 92.7 | 88.2 |
| 35 | 90.2 | 91.1 | 84.2 |
| 30 | 88.2 | 85.6 | 78.0 |
| 25 | 84.0 | 74.2 | 69.8 |
| 20 | 75.6 | 64.7 | 60.7 |
| 15 | 58.9 | 46.7 | 49.1 |

Table 7: Recognition performance of the MLP classifier for noisy speech in the speaker-independent mode using 10 and 15 cepstral coefficients.

| SNR | Recognition accuracy (in %) using | |
|---|---|---|
| | 10 cepstral coefficients | 15 cepstral coefficients |
| Clean | 91.3 | 92.0 |
| 35 | 90.2 | 92.0 |
| 30 | 88.2 | 90.4 |
| 25 | 84.0 | 87.1 |
| 20 | 75.6 | 78.9 |
| 15 | 58.9 | 69.6 |

that of the ML and the kNN classifiers. Since the cepstral coefficient representation is shown to be the best representation in the preceding section, it is used here as parametric representation with all the three classifiers.

All the three classifiers (MLP, ML and kNN) are studied in both the speaker-dependent and the speaker-independent modes. Ten cepstral coefficients derived through 10-th order LP analysis are used as recognition parameters. Recognition results for all the three classifiers at different SNR conditions are shown in Table 5 for the speaker-dependent mode and in Table 6 for the speaker-independent mode. It can be seen from these tables that the MLP classifier gives significantly better recognition accuracy than the ML and the kNN classifiers for noisy speech at all the SNR conditions. Degradation in recognition accuracy due to decrease in SNR occurs at a slower rate in the MLP classifier than in the other two classifiers.

So far, we have used 10 cepstral coefficients derived through 10-th order LP analysis as recognition parameters. With conventional classifiers, some authors [14,15] have found it useful to use more than M cepstral coefficients with M-th order LP analysis for recognition (though these additional cepstral coefficients do not add any new information about the LP spectral envelope). In order to see whether this is useful for the MLP classifier, we use here 15 cepstral coefficients as recognition parameters. These cepstral coefficients are obtained by extrapolating the 10 cepstral coefficients through a recursive relation (Equ. 2.14 in [5]). We study the performance of the MLP classifier for noisy speech under different SNR conditions. Results are shown in Table 7. It can be seen from this table that the recognition performance is better at all SNR conditions with 15 cepstral coefficients than with 10 cepstral coefficients.

It has been reported in the literature [14-17] that the use of a lifter with the cepstral coefficients improves the recognition performance of the conventional classifiers. In order to see whether this is true for the MLP classifier, we study here the linear lifter with the 10 cepstral coefficients [16] and the raised sinusoidal lifter with the 15 cepstral coefficients [14]. Our results show that none of these lifters is helpful in improving the recognition performance of the MLP classifier.

## 5. CONCLUSIONS

In this paper, the MLP classifier is studied for the recognition of noisy speech and its performance is compared with the conventional pattern classifiers (ML and kNN). The LP parameters derived through 10-th order LP analysis are used here as the recognition parameters. Different LP parametric representations are compared as to their recognition performance with the MLP classifier and the cepstral coefficient representation is found to be the best parametric representation. Using 10 cepstral coefficients as recognition parameters, performance of the MLP classifier is found to be significantly better than that of the ML and the kNN classifiers for noisy speech at all the SNR conditions, showing its robustness for the recognition of noisy speech. Use of 15 cepstral coefficients (obtained by extrapolating the 10 cepstral coefficients) improves the recognition performance of the MLP classifier for noisy speech further. It is hoped that when the MLP classifier is applied with a more robust spectral parametric representation [4], we may get further improvement in recognition performance for noisy speech.

## REFERENCES

[1] L.R. Rabiner and J.G. Wilpon, "Some performance benchmarks for isolated word recognition", Computer Speech and Language, Vol. 2, pp 343-357, 1987.

[2] B.A. Dautrich, L.R. Rabiner and T.B. Martin, "On the effects of varying filter bank parameters on isolated word recognition", IEEE Trans. ASSP-31, pp. 793-806, 1983.

[3] Y. Ephraim, J.G. Wilpon and L.R. Rabiner, "A linear predictive front-end processor for speech recognition in noisy environments", Proc. ICASSP, pp. 1324-1327, 1987.

[4] D. Mansour and B.H. Juang, "The short-time modified coherence representation and noisy speech recognition", IEEE Trans. ASSP-37, pp. 795-804, 1989.

[5] D. Mansour and B.H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition", IEEE Trans. ASSP-37, pp. 1659-1671, 1989.

[6] F.K. Soong and M.M. Sondhi, "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise", IEEE Trans. ASSP-36, pp. 41-48, 1988.

[7] D.E. Rumelhart et al., "Learning representations by back propagating errors", Nature, Vol. 323, pp. 533-536, 1986.

[8] R.P. Lippmann, "Review of neural networks for speech recognition", Neural Computation, Vol. 1, pp. 1-38, 1989.

[9] R.P. Lippmann, "An introduction to computing with neural nets", IEEE ASSP Magazine, Vol. 4, pp. 4-22, 1987.

[10] W.Y. Huang and R.P. Lippmann, "Comparisons between conventional and neural net classifiers", Proc. ICNN, 1987.

[11] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1976.

[12] K.K. Paliwal and P.V.S. Rao, "Evaluation of various linear prediction parametric representations in vowel recognition", Signal Processing, Vol. 4, pp. 323-327, 1982.

[13] T. Kohonen, "The neural phonetic typewriter", Computer, pp. 11-22, Mar. 1988.

[14] B.H. Juang, L.R. Rabiner and J.G. Wilpon, "On the use of bandpass liftering in speech recognition", IEEE Trans. ASSP-35, pp. 947-954, 1987.

[15] F. Itakura and T. Umezaki, "Distance measures for speech recognition based on the smoothed group delay spectrum", Proc. ICASSP, pp. 1257-1260, 1987.

[16] K.K. Paliwal, "On the performance of the quefrency weighted cepstral coefficients in vowel recognition", Speech Communication, Vol. 1, pp. 151-154, 1982.

[17] B.A. Hanson and H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise", IEEE Trans. ASSP-35, pp. 968-973, 1987.