Microphone Arrays and Neural Networks for Robust Speech Recognition

C. Che+, Q. Lin+, J. Pearson, B. de Vries*, and J. Flanagan+*

+CAIP Center, Rutgers University, Piscataway, NJ 08855-1390
and
*David Sarnoff Research Center, Princeton, NJ 08543-5300

ABSTRACT

This paper explores use of synergistically-integrated systems of microphone arrays and neural networks for robust speech recognition in variable acoustic environments, where the user must not be encumbered by microphone equipment. Existing speech recognizers work best for "high-quality close-talking speech." Performance of these recognizers is typically degraded by environmental interference and mismatch in training conditions and testing conditions. It is found that use of microphone arrays and neural network processors can elevate the recognition performance of existing speech recognizers in an adverse acoustic environment, thus avoiding the need to retrain the recognizer, a complex and tedious task. We also present results showing that a system of microphone arrays and neural networks can achieve a higher word recognition accuracy in an unmatched training/testing condition than that obtained with a *retrained* speech recognizer using array speech for both training and testing, i.e., a *matched* training/testing condition.

1. INTRODUCTION

Hidden Markov Models (HMM's) have to date been accepted as an effective classification method for large vocabulary continuous speech recognition. Existing HMM-based recognition systems, e.g., SPHINX and DECIPHER, work best for "high-quality close-talking speech." They require consistency in sound capturing equipment and in acoustic environments between training and testing sessions. When testing conditions differ from training conditions, performance of these recognizers is typically degraded if they are not retrained to cope with new environmental effects.

Retraining of HMM-based recognizers is complex and time-consuming. It requires recollection of a large amount of speech data under corresponding conditions and reestimation of HMM's parameters. Particularly great time and effort are needed to retrain a recognizer which operates in a speaker-independent mode, which is the mode of greatest general interest.

Room reverberation and ambient noise also degrade performance of speech recognizers. The degradation becomes more prominent as the microphone is positioned more distant from the speaker, for instance, in a teleconferencing application. Previous work has demonstrated that beamforming/matched-filter microphone arrays can provide higher signal-to-noise ratios than can conventional microphones used at distances (see, e.g., [1, 2]). Consequently, there is increasing interest in microphone arrays for hands-free operation of speech processing systems [3]-[7].

In this report, a system of microphone arrays and neural networks is described which expands the power and advantages of existing ARPA speech recognizers to practical acoustic environments where users need not be encumbered by hand-held or body-worn microphone systems. (Examples include Combat Information Centers, large group conferences, and mobile hands-busy eyes-busy maintenance tasks.) Another advantage of the system is that the speech recognizer need not be retrained for each particular application environment. Through neural network computing, the system learns and compensates for environmental interference. The neural network transforms speech-feature data (such as cepstrum coefficients) obtained from a distant-talking microphone array to those corresponding to a high-quality, close-talking microphone. The performance of the speech recognizer can thereby be elevated in the hostile acoustic environment without retraining of the recognizer.

The remainder of the paper is organized as follows. First, a new speech corpus with simultaneous recording from different microphones is described in Section 2. Next, the system of microphone arrays and neural networks is discussed in Section 3. The system is evaluated using both the SPHINX speech recognizer and a Dynamic-Time-Warping (DTW) based recognizer. The results are presented in Sections 4 and 5, respectively. In Section 6, performance comparisons are made of different network architectures to identify an optimal design for room-acoustic equalization. Finally, we summarize the study and discuss our future work in Section 7.

2. SPEECH CORPUS

A speech database has been recently created at the CAIP Center for evaluation of the integrated system of microphone arrays, neural networks, and ARPA speech recognizers. The database consists of 50 male and 30 female speakers. Each speaker speaks 20 isolated command-words, 10 digits, and 10 continuous sentences of the Resource Management task. Of the continuous sentences, two are the same for all speakers and the remaining 8 sentences are chosen at random. Two recording sessions are made for each speaker. One session is for simultaneous recording from a head-mounted close-talking microphone (HMD 224) and from a 1-D beamforming line array microphone (see Section 3.1). The other is for simultaneous recording of the head-mounted close-talking microphone and a desk-mounted microphone (PCC 160). The recording is done with an Ariel ProPort with a sampling fre-

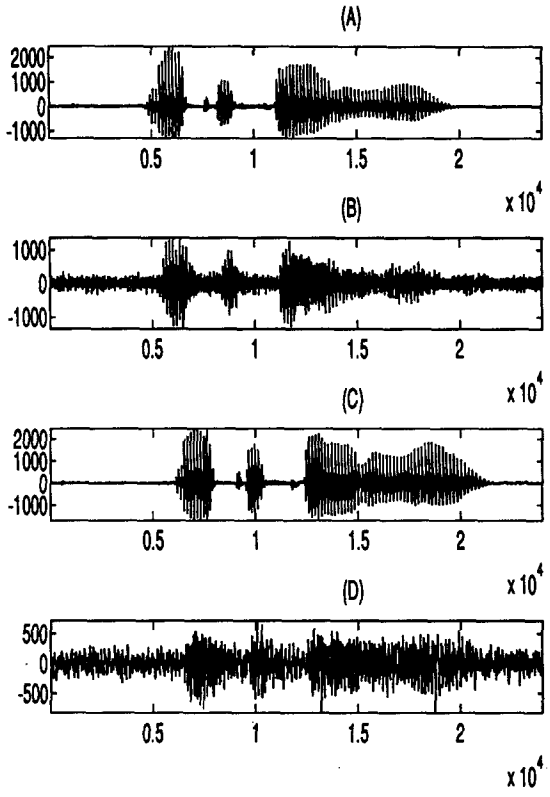

Figure 1: Speech waveforms from the head-mounted microphone (A and C), from the 1-D line array microphone (B), and from the desk-mounted microphone (D). (A) and (B) are simultaneously recorded in a session and (C) and (D) in a following session. The utterance is: "Microphone array," spoken by a male speaker (ABF).

quency of 16 kHz and 16-bit linear quantization. The recording environment is a hard-walled laboratory room of $6 \times 6 \times 2.7$ meters, having a reverberation time of approximately 0.5 second. Both the desk-mounted microphone and the line array microphone are placed 3 meters from the subjects. Ambient noise in the laboratory room is from several workstations, fans, and a large-size video display equipment for teleconferencing. The measured 'A' scale sound pressure level is 50 dB. Indicative of the quality differences in outputs from various sound pickup systems, signal waveforms are given in Figure 1. Because of wave propagation from the speaker to distant microphones, a delay of approximately 9 msec is noticed in outputs of the line array and the desk-mounted microphone. Wave propagation between the subject's lips to the head-mounted close-talking microphone is negligible. The reader is referred to [8] for more details.

3. SYSTEM OF MICROPHONE ARRAYS AND NEURAL NETWORKS

Figure 2 schematically shows the overall system design for robust speech recognition in variable acoustic environments, incorporating microphone arrays, neural networks, and ARPA speech recognizers.

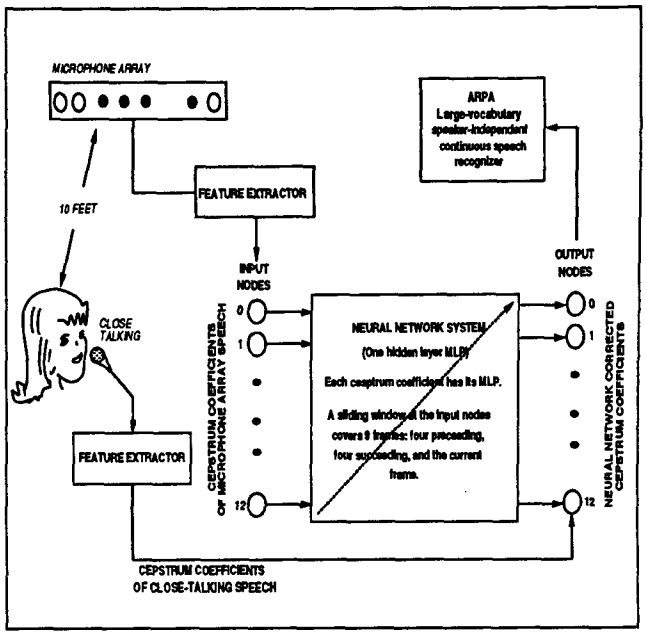

Figure 2: Block diagram of the robust speech recognition system. The neural network processor is trained using simultaneously recorded speech. The trained neural network processor is then used to transform spectral features of array input to those appropriate to close-talking. The transformed spectral features are inputs to the speech recognition system. No retraining or modification of the speech recognizer is necessary. The training of the neural net typically requires about 10 seconds of signal.

3.1. Beamforming Microphone Arrays

As the distance between microphones and talker increases, the effects of room reverberation and ambient noise become more prominent. Previous studies have shown that beamforming/matched-filter array microphones are effective in counteracting environmental interference. Microphone arrays can improve sound quality of the captured signal, and avoid hand-held, body-worn, or tethered equipment that might encumber the talker and restrict movement.

The microphone array we use here is a one-dimensional beamforming line array. It uses direct-path arrivals to produce a single-beam delay-and-sum beamformer [1, 2]. (The talker typically faces the center of the line array.) The array consists of 33 omni-directional sensors, which are nonuniformly positioned (nested over three octaves). From Figure 1 it is seen that the waveform of the array resembles that of the close-talking microphone more than the desk-mounted microphone.

3.2. Neural Network Processors

One of the neural network processors we have explored, is based on multi-layer perceptrons (MLP). The MLP has 3 lay-

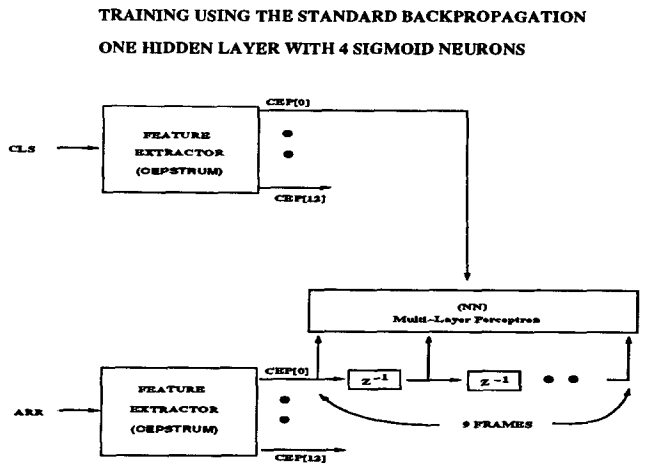

Figure 3: A feedforward delay network for mapping the cepstral coefficients of array speech to those of close-talking speech.

ers. The input layer has 9 nodes, covering the current speech frame and four preceding and four following frames. as indicated in Figure 3. There are 4 sigmoid nodes in the hidden layer and 1 linear node in the output layer. 13 such MLP's are included, with one for each of the 13 cepstrum coefficients used in the SPHINX speech recognizer [14]. (Refer also to Figure 2.) The neural network is trained using a modified backpropagation method when microphone-array speech and close-talking speech are both available (see Figure 3).

It is found that 10-seconds of continuous speech material are sufficient to train the neural networks and allow them to "learn" the acoustic environment. In the present study, the neural nets are trained in a speaker-dependent mode; That is, 13 different neural networks (one for each cepstrum coefficient) are dedicated to each subject[1]. The trained networks are then utilized to transform cepstrum coefficients of array speech to those of close-talking speech, which are then used as inputs to the SPHINX speech recognizer.

4. EVALUATION RESULTS WITH SPHINX RECOGNIZER

As a baseline evaluation, recognition performance is measured on the command-word subset of the CAIP database. Performance is assessed for matched and unmatched testing/training conditions and include both the *pretrained* and retrained SPHINX system.

The results for the pretrained SPHINX are given in Table 1. It includes four processing conditions: (i) close-talking; (ii) line array; (iii) line array with mean subtraction (MNSUB) [15]; and, (iv) line array with the neural network processor (NN).

Table 2 gives the results for the retrained SPHINX under five processing conditions: (i) close-talking; (ii) line array; (iii) desk-mounted microphone; (iv) line array with mean subtraction (MNSUB); and, (v) line array with the neural network processor (NN). The SPHINX speech recognizer is retrained using the CAIP speech corpus to eliminate system conditions in collection of the Resource Management task (on which the original SPHINX system has been trained) and the CAIP speech database.

[1] The learning rate is 0.01 and the momentum term is 0.5. The training terminates at 1000 epochs.

Testing Microphone	Word Accuracy
Close-Talking	88%
Line-Array	16%
Line-Array +MNSUB	24%
Line-Array + NN	82%

Table 1: Baseline evaluation of recognition performance (% correct), using the *pretrained* SPHINX speech recognizer.

As shown in Tables 1 and 2, the array-neural net system is capable of elevating word accuracy of the speech recognizer. For the retrained SPHINX, the microphone array and neural network system improves word accuracy from 21% to 85% for distant talking under reverberant conditions. On the other hand, the mean subtraction method under these conditions improves the performance only marginally.

It is also seen from Table 2 that if the SPHINX system has been retrained with array speech at a distance of 3 meters, the performance is as high as 82%. The figure, obtained under a matched training/testing condition, is, however, lower than that obtained under an unmatched training/testing condition with microphone array and neural network. Similar results have been achieved for speaker identification [9, 10].

5. EVALUATION RESULTS WITH DTW RECOGNIZER

To more effectively and efficiently assess the capability of microphone arrays and neural network equalizers, a DTW-based speech recognizer is implemented [12]. The back end of DTW classification is simple, and hence, the results do not tend to be influenced by the complex back end of an HMM-based recognizer, including language models and word-pair grammars.

Testing	Training Close-Talking	Training Line-Array
Close-Talking	95%	-
Line-Array	21%	82%
Desk-mounted	13%	-
Line-Array + MNSUB	27%	-
Line-Array + NN	85%	-

Table 2: Baseline evaluation of recognition performance (% correct), using a *retrained* SPHINX recognizer based on the CAIP speech database.

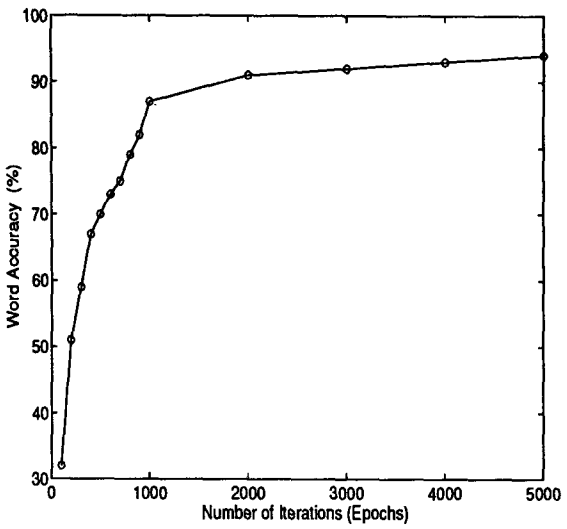

Figure 4: Word recognition accuracy on the testing set as a function of the number of iterations when training the neural network processor.

The DTW recognizer is applied to recognition of the command words. End-points of close-talking speech are automatically determined by the two-level approach [11][2]. Attempts have also been made to automatically detect end-points of array speech [13], but in the present paper, the starting/ending points are inferred from the simultaneously recorded close-talking speech, with an additional delay resulting from wave propagation. The DTW recognizer is speaker dependent, and is trained using close-talking speech. The measured features are 12th-order LPC-derived cepstral coefficients over a frame of 16 msec. The frame is Hamming-windowed and the consecutive windows overlap by 8 msec. The DTW recognizer is tested on array speech (with the originally computed and neural-network corrected cepstral coefficients) and on the other set of the close-talking recording. The Euclidean distance is utilized as the distortion measure.

The recognition results, pooled over 10 male speakers, are presented in Table 3. The configuration of MLP used in this DTW based evaluation differs from that in Section 4. A single MLP with no window-sliding is now used to collectively transform all of 12 cepstral coefficients from array speech to close-talking. The MLP has 40 hidden nodes and 12 output nodes. The network is again speaker-dependently trained with standard backpropagation algorithms. The learning rate is set to 0.1 and the momentum term to 0.5. The backpropagation procedure terminates after 5000 iterations (epochs).

It can be seen that the results in Table 3 are similar to those in Tables 2 and 1. The use of microphone arrays and neural networks elevates the DTW word accuracy from 34% to 94% under reverberant conditions. The elevated accuracy is close to that obtained for close-talking speech (98%).

[2]The automatic results conform well with manual editing.

Testing Microphone	Word Accuracy
Close-Talking	98%
Line-Array	34%
Line-Array + NN	94%

Table 3: Baseline evaluation of recognition performance using DTW classification algorithms.

Figure 4 illustrates the relationship between the number of training iterations of the neural networks and the word recognition accuracies. It is seen that as the iteration number increases from 100 to 1000, the recognition accuracy quickly rises from 32% to 87%. It can also be seen that after 5000 iterations the network is not overtrained, since recognition accuracy on *the testing set* is still improving.

6. PERFORMANCE COMPARISON OF DIFFERENT NETWORK ARCHITECTURES

We also perform comparative experiments with respect to different network architectures. It has been suggested in the communications literature that recurrent non-linear neural networks may outperform feedforward networks as equalizers. Since our problem can be interpreted as a room acoustics equalization task, we decide to evaluate the performance of recurrent nets. For the experiments reported here, we only train on data from the 3rd cepstral coefficient (out of 13 bands). The input to the neural net is the cepstral data from the microphone array; the target cepstral coefficient is taken from the close-talking microphone. The squared error between the target data and the neural net output is used as the cost function. The neural nets are trained by gradient descent. The following three different architectures have been evaluated: (i) a linear feedforward net (adaline) [16], (ii) a non-linear feedforward net, (iii) and a non-linear recurrent network. The input layer of all nets consisted of a tapped delay line. The network configurations are depicted in Figures 5 and 6.

Experimental results are summarized in Table 4, where the entry "nflops/epoch" stands for the number of (floating point) operations required during training per epoch. The entry "#parameters" holds the number of adaptive weights in the network.

It is clear that, for this dataset, the non-linear networks perform better than the linear nets, but at the expense of considerably more computations during adaptation. This is not a problem if we assume that the transfer function from speaker to microphone is constant, but in a changing environment (moving speaker, doors opening, changing background noise) this is a problem, as the neural net needs to track the change in real-time. It should be noted that the used cost function, the squared error, is in all likelihood not a monotonic function of the recognizer performance. Currently experiments are underway that evaluate the performance of various network architectures in terms of word recognition accuracy.

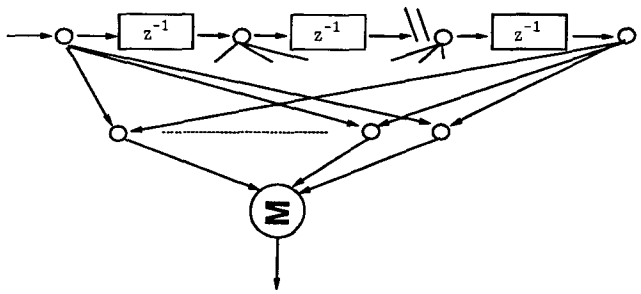

Figure 5: The feedforward net. The hidden units are nonlinear (*tanh*).

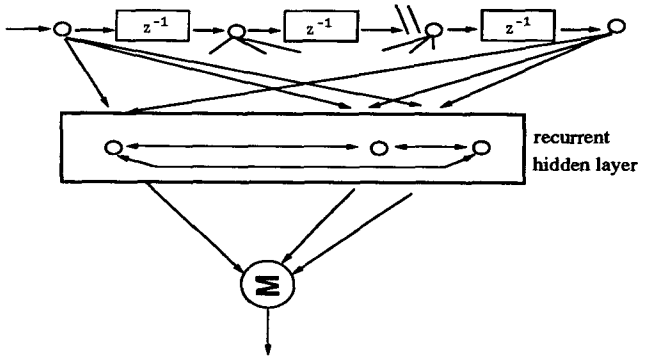

Figure 6: The recurrent network has a similar structure as the 2-layer feedforward.

architecture	final sqe	nflops/epoch	# parameters	adaptation rule
no processing	.12			
adaline (1 tap)	.0952	~14,000	1	delta rule
adaline (5)	.0844	~40,000 (1)	5	delta
adaline (11)	.0825	~80,000 (2)	11	delta
ffwdnet (5,2,1)	.0787	~1924,000 (48)	15	backprop
recnet (5,2r,1)	.0782	~2478,000 (62)	19	bptt
ffwdnet (5,4,1)	.0775	~3772,000 (94)	29	backprop

Figure 7: Experimental results of different neural network configurations. The various runs are ordered by increasing performance. Final sqe (squared error) is the mean sqe per time step on the test database. The ops/epoch denotes the number of floating point operations per epoch during training. The number in brackets is the number of flops per epoch divided by flops/epoch for adaline (5 taps). # parameters denotes the number of adaptive parameters in the network.

7. CONCLUSION AND DISCUSSION

The above evaluation results suggest that the system of microphone array and neural network processors can

- effectively mitigate environmental acoustic interference
- without retraining the recognizer, elevate word recognition accuracies of HMM-based and/or DTW-based speech recognizers in variable acoustic environments to levels comparable to those obtained for close-talking, high-quality speech
- achieve word recognition accuracies, under unmatched training and testing conditions, that exceed those obtained with a retrained speech recognizer using array speech for both retraining and testing, i.e., under em matched training and testing conditions

Similar results have also been achieved for studies on speaker recognition [9, 10].

In future work, we expect to extend the comparative evaluations of different neural network architectures, so that the performance of neural network equalization can be addressed in terms of word recognition accuracy. We also want to extend the evaluation experiments to continuous speech. For comparison, the DECIPHER system will be included, and possibly other advanced ARPA speech recognizers. The CAIP Center has concomitant NSF projects on developing 2-D and 3-D microphone arrays. These new array microphones have better spatial volume selectivity and can provide a high signal-to-noise ratio. They will be incorporated into this study. Further work will compare the system of microphone array and neural network with other existing noise compensation algorithms, such as Codebook Dependent Cepstrum Normalization (CDCN) [17] and Parallel Model Combination (PMC) [18].

8. ACKNOWLEDGMENT

This work is supported by ARPA Contract No: DABT63-93-C-0037. The work is also in part supported by NSF Grant No: MIP-9121541.

References

1. Flanagan, J., Berkley, D., Elko, G., West, J., and Sondhi, M., "Autodirective microphone systems," *Acustica 73*, 1991, pp. 58-71.
2. Flanagan, J., Surendran, A., and Jan, E., "Spatially selective sound capture for speech and audio processing," *Speech Communication, 13*, Nos. 1-2, 1993. pp. 207-222.
3. Silverman, H. F., "Some analysis of microphone arrays for speech data acquisition," *IEEE Trans. Acous. Speech Signal Processing 35*, 1987, pp. 1699-1712.
4. Berkley, D. A. and Flanagan, J. L., "HuMaNet: An experimental human/machine communication network based on ISDN," *AT & T Tech J.*, 1990, pp. 87-98.
5. Che, C., Rahim, M., and Flanagan J. "Robust speech recognition in a multimedia teleconferencing environment," *J. Acous. Soc. Am. 92* (4), pt 2, p. 2476(A), 1992.

6. Sullivan, T. and Stern, R. M., "Multi-microphone correlation-based processing for robust speech recognition," *Proc. ICASSP-93*, April, 1993.

7. Lin, Q., Jan, E., and Flanagan, J., "Microphone-arrays and speaker identification," Accepted for publication in the special issue on Robust Speech Processing of the *IEEE Trans. on Speech and Audio Processing*, 1994.

8. Lin, Q., C. Che, and R. Van Dyck: "Description of CAIP speech corpus," *CAIP Technical Report*, Rutgers University, in preparation, 1994.

9. Lin, Q., Jan, E., Che, C., and Flanagan, J., "Speaker identification in teleconferencing environments using microphone arrays and neural networks," *Proc. of ESCA Workshop on Speaker, Recognition, Identification, and Verification*, Switzerland, April, 1994.

10. Lin, Q., Che, C., Jan, E., and Flanagan, J., "Speaker/speech recognition using microphone arrays and neural networks," paper accepted for the SPIE Conference, San Diego, July, 1994.

11. Rabiner, L. and Sambur, M. "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J. 54*, No. 2, 1975, pp. 297-315.

12. Sakoe, H. and Chiba, S. "Dynamic programming optimization for spoken word recognition." *IEEE Trans. on Acous. Speech Signal Processing 26*, 1978, pp. 43-49.

13. Srivastava, S., Che, C., and Lin, Q., "End-point detection of microphone-array speech signals," Paper accepted for 127th meeting of *Acous. Soc. of Amer.*, Boston, June, 1994.

14. Lee, K.-F., *Automatic Speech Recognition: The Development Of The SPHINX System*, Kluwer Academic Publishers, Boston, 1989.

15. Furui, S. "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol 29, 1979, pp 254-272.

16. de Vries, B., "Short term memory structures for dynamic neural networks," to appear in *Artificial Neural Networks with Applications in Speech and Vision* (Ed. R. Mammone).

17. Liu, F.-H., Acero, A., and Stern, R. M., "Efficient joint compensation of speech for the effect of additive noise and linear filtering," *ICASSP-92*, April, 1992, pp 257-260.

18. Gales, M. J. F. and Young, S. J., "Cepstral parameter compensation for HMM recognition," *Speech Communication 12*, July, 1993, pp. 231-239.