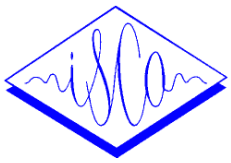

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SYSTEM OF MICROPHONE ARRAYS AND NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION IN MULTIMEDIA ENVIRONMENTS

*Qiguang Lin, Ea-Ee Jan, ChiWei Che, and Bert de Vries**

CAIP Center, Rutgers University, Piscataway, NJ 08855, USA
* David Sarnoff Research Center, Princeton, New Jersey, USA

ABSTRACT

Hands-free operation of speech processing systems is sometimes desired to avoid encumbrance of the user by tethered microphone equipment. This paper explores the use of array microphones and neural networks (MANN) for robust speech recognition in real-world environments, such as large-group conferencing. Microphone arrays (MA) provide high-quality, hands-free sound pickup under severe acoustical conditions; and neural network (NN) processors "learn" the characteristics of environmental interference and transform features of MA-enhanced signal to those obtained under close-talking conditions. In this study, both real-room collected and computer-simulated reverberant speech signals are used to evaluate the power and advantages of MANN for direct deployment of speech recognition technology in adverse practical environments.

1. INTRODUCTION

Multimedia conferencing and interaction with an information system should be as easy and natural as face-to-face communication with a human [1]. This suggests distant-talking sound pick-up in a reverberant, noisy environment, which would contribute to spatial realism and allow the user not to be tethered or encumbered by microphone equipment. However, because of multipath distortion (reverberation) and ambient noise in an enclosure, the sound signal captured by conventional microphones at distance is much distorted. Distorted signal (low SNR's) typically degrades performance of speech recognizers. The degradation becomes more prominent as the microphone is positioned in a more distant place from the speaker, for instance, in a large-group audio/video conferencing application.

Performance of speech recognition also degrades when there is a mismatch between a target application condition and the training condition of a recognizer. Usually, the recognizer is *retrained* for that specific application condition to overcome the mismatch problem. However, retraining is a tedious and costly process, especially for those Hidden Markov Model based and speaker-independent speech recognition systems. A huge amount of speech data from many talkers under corresponding conditions needs to be collected and model parameters need to be estimated. Therefore, it is very impractical to retrain the recognizer for each specific condition.

To expand speech recognition technology to multimedia conferencing applications, we are developing an integrated system of microphone arrays and neural networks (MANN). These two synergistic components address (1) speech enhancement by microphone array (MA) and (2) feature adaptation by neural network (NN) computing, respectively. When features (e.g., cepstrum coefficients) of *enhanced* signals, captured by MA under the testing condition, are transformed, via a trained NN, to those obtained under the training condition of the recognizer, a matched training and testing condition is thereby approximated. Therefore, the MANN system has the following advantages: (i) It allows for effective sound capture under distant-talking, reverberant, and noisy environments; (ii) It approximates a matched training and testing condition without the necessity to retrain a speech recognizer; (iii) Training an NN is faster and more efficient than training a large-vocabulary, speaker independent speech recognizer.

In this paper, we study speech enhancement and feature adaptation by MANN as applied to recognition of isolated words. The speech recognizer is based on dynamic-time-warping (DTW) techniques. Both real-room collected and computer-simulated degraded speech signals are used in the experiments. The recognizer itself is trained with close-talking microphone input. Although the objective of our research is for robust large-vocabulary continuous speech recognition, use of a DTW-based speech recognizer enables an effective and efficient assessment of the power and advantages of MANN to elevate recognition performance under adverse acoustical environments. The efficiency comes from the relatively simple back end of DTW speech recognizers. In addition, a DTW recognizer is appropriate for recognition of a small size of vocabulary and can be used to control set-up and data display of conferencing facilities.

The remainder of this paper is organized as follows. In Section 2, a brief description of the MANN system is presented. In Section 3 and 4, collection/generation of impaired speech signal is described. Experimental results of isolated-word recognition using MANN are given and analyzed in Sections 5 for both collected and simulated speech data. Finally in Section 6, concluding remarks are presented.

2. SYSTEM OF MANN

Figure 1 schematically shows the overall system design of MANN for robust speech recognition, incorporating a microphone array, a neural network processor, and a speech recognizer. The MANN system is typically used when the recognition module is trained with close-talking microphone input and is tested with distant-talking microphone-array input, i.e., an unmatched training/testing condition.

The MANN system operates as follows. First, a simultaneous recording with the close-talking microphone and with the microphone array is needed to train the neural network processor. Typically, 10-second stereo speech material is sufficient to train a speaker-dependent NN. Cepstrum coefficients of the close-talking microphone input and array

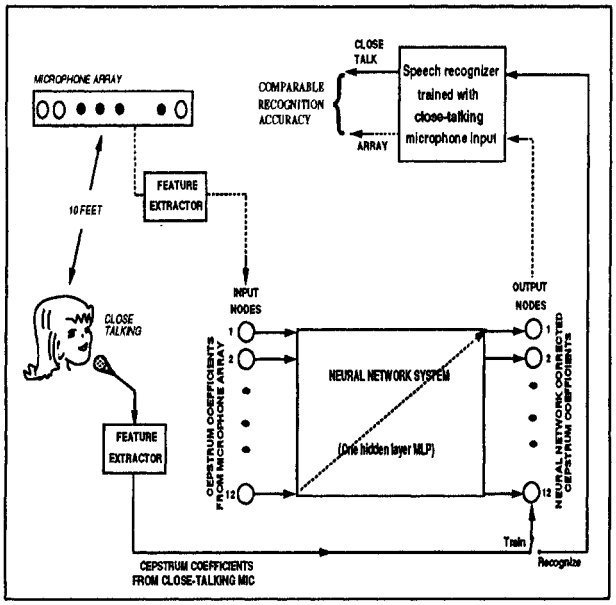

Figure 1: Block diagram of the robust speech recognition system, MANN. The neural network processor is trained using simultaneously recorded speech. The trained neural network processor is then used to transform spectral features of array input to those appropriate to close-talking. The transformed spectral features are inputs to the speech recognition system. No retraining or modification of the speech recognizer is necessary. The training of a speaker-dependent neural net typically requires about 10 seconds of signal.

input are then calculated and used as the target output and the input, respectively, to train the neural network [2] [3].

During recognition, the trained NN transforms cepstrum coefficients of the array input to those corresponding to the close-talking microphone input. The transformed cepstrum coefficients are the input to the speech recognition system. Because of the good sound quality of MA and the feature adaptation by NN, speech recognition performance can be elevated for the unmatched training and testing condition. For experiments under matched training/testing conditions, the NN adaptation is not necessary. Note that the zeroth-order cepstrum coefficient, which is a measure of gain, is not used in the present study.

One of the NN paradigms we use in this study is a fully-connected multi-layer perceptron (MLP). The MLP has 3 layers. The input layer and output layer have 12 nodes each: one for each of the 12 cepstrum coefficients. The hidden layer has 40 nodes associated with nonlinear (sigmoid) activation functions. The activation function at the output layer is linear. The NN is speaker-dependently trained using a backpropagation algorithm.

3. COLLECTED SPEECH DATA

A "hands-free" speech database has been recently created at the CAIP Center for evaluation of the integrated system of a microphone array, a neural network, and a speech recognizer. 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. 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 of 29 harmonically nested gradient microphone sensors. The other is for simultaneous recording of the head-mounted close-talking microphone and a desk-mounted microphone (PCC 160). Both the desk-mounted microphone and the line array microphone are placed 3 meters from the subjects. The recording is done with an Ariel ProPort with a sampling frequency 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. The measured ambient noise level was approximately 50 dB on the A scale of a sound-level meter and 71 dB on the C scale. The reader is referred to Lin et al [4] for more details about the CAIP speech corpus.

4. SIMULATED SPEECH DATA

There are various designs of microphone arrays [5]. Some of these are yet to be constructed. (We are currently constructing a large-scale 3-dimensional multiple-beamforming/matched-filter microphone array. We will extend this research when the construction is completed.) In addition, it is very time-consuming to collect simultaneous data under a variety of acoustical conditions of interest. Therefore, we have used a computer model of room acoustics to generate reverberant speech captured by simulated matched-filter arrays, for evaluation of MANN.

4.1. Image Model Of Room Acoustics

The walls of conventional rooms are large and the roughness dimensions of the wall surfaces are small compared to acoustic wavelengths of interest. Therefore, the walls constitute effective reflectors (mirrors) and acoustic wave propagation, from source to receiver, can be determined by ray tracing techniques. A computer model of room acoustics has recently been implemented by Jan and Flanagan [6], based on the image technique of computing source-receiver impulse responses. In the model, image sources are determined in accordance with Snell's Law. The strength of the images depends upon the absorption of the walls that each ray path encounters. For simplicity, the absorption is assumed to be frequency independent in the model.

4.2. Matched-filter Array

A matched-filter is the time inverse of the impulse response of the system to be matched. In a matched-filter array, a matched-filter is dedicated to each microphone to achieve spatial volume selectivity and mitigation of noise interference [5]. The array output is the summation of outputs from each matched-filter. The impulse response from the desired focal point to each receiver in the array is required to implement the matched-filter array system. This response can be calculated from the room geometry or measured in actual rooms. For a source located at the focal point emitting a signal $s(t)$, the temporal output of the matched-filter array is

$$\begin{aligned} O_f(t) &= \sum_{n=1}^{N} s(t) * h_{nf}(t) * h_{nf}(-t) \\ &= s(t) * \sum_{n=1}^{N} h_{nf}(t) * h_{nf}(-t) \end{aligned} \quad (1)$$

where $h_{nf}(t)$ is the impulse response from the focal point to the nth sensor, N is the total number of sensors, and * denotes convolution. The term denoted by the summation in Eq. (1) is recognized as the autocorrelation of the impulse response from focus to sensor. When the source is off the focal position, the temporal output of the array is

$$O_o(t) = s(t) * \sum_{n=1}^{N} h_{ns}(t) * h_{nf}(-t) \qquad (2)$$

where $h_{ns}(t)$ is the impulse response from the source to the nth sensor. In Eq. (2), the term denoted by summation is recognized as the cross-correlation of the impulse responses from focus to sensor and from source to sensor. One sees that the size of the focal volume for retrieval of low distortion signals is conditioned by the spatial correlation of the impulse responses $h_{nf}(t)$ and $h_{ns}(t)$.

4.3. Generation Of Degraded Speech

Using the image model of room acoustics, a large enclosure, $20 \times 16 \times 5$ m, is simulated. It is highly reverberant, with all the absorption coefficients being set to 0.1. The reverberation time is on the order of 1.6 seconds. Image sources up to fifth order are included in generating reverberant speech. Inputs to the room simulation are the close-talking microphone input taken from the command-word subset of the CAIP "hands-free" corpus. The signal source is placed at the point (14,9.5,1.7) m. A competing noise source of variable intensity, to produce different signal-to-competing-noise ratios (SCNR), can be turned on or off. The noise is generated by a Gaussian random number generator and is located at (3.0,5.0,1.0) m. The noise sequence is different for different sentences.

The following sound pickup systems are used to capture the degraded speech:

1. *Single microphone system.* A single microphone receiver is located at (10,0.5,1.7) m. The overall system impulse response is simply the impulse response from the source to the receiver. The microphone is an omni-directional pressure sensor.
2. *Matched-filter arrays.* Two 31×31 matched-filter arrays are placed on orthogonal walls. Centers of the arrays are at (10,0,1,7) m and (0,8,1.7) m, respectively. Microphone elements are omni-directional and are uniformly separated at a distance of 4 cm between adjacent sensors. The overall system impulse response is calculated from Eqs. (1) and (2) with appropriate temporal alignment [5].

It should be pointed out that in the above positioning, the noise source is closer to the microphone sensors than is the speech (signal) source. This unfavorable positioning results in an even lower SNR at a specified SCNR.

5. RECOGNITION RESULTS

In this section, performance of MANN is evaluated in the context of isolated-word recognition using a DTW based recognizer. Both real-room collected and computer-simulated degraded speech signals are used in the experiments. The recognizer itself is trained with close-talking microphone input.

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

Table 1: Word recognition accuracies in %. The DTW recognizer has been trained with close-talking microphone input.

5.1. For Collected Speech Data

DTW approaches require that end-points of the utterance be determined first. In order to assure that our results will not be contaminated by errors in end-point detection of noisy speech, end-points of close-talking speech are automatically determined by the two-level approach. The end-points are then used to infer starting and ending points of the array speech simultaneously recorded. (Attempts have also been made to automatically detect end-points of array speech [4] [7]). Recall that the array was positioned 3 meters away from the subjects, which accounts for approximately 9 ms delay (or about one frame of 10 ms long) in array input, relative to close-talking microphone input.

The DTW recognizer is speaker dependent and is trained with one of the two sets of clean speech of the CAIP database. The measured features are 12th-order LPC-derived cepstral coefficients over a frame of 16 msec. (The zeroth-order coefficient is not used.) The frame is Hamming-windowed and the consecutive windows overlap by 8 msec. The Euclidean distance is utilized as the distortion measure. The recognizer is tested on the other set of close-talking speech and on the array speech (with the originally calculated and NN corrected cepstral coefficients).

The recognition results, pooled over 10 male speakers, are presented in Table 1. As expected, a high word recognition accuracy is obtained with close-talking microphone input as the testing material, which corresponds to a matched training and testing condition. On the other hand, low performance of the array speech in Table 1 is due to the mismatch in sound capture between training and testing conditions. We have compared frequency responses of the head-mount microphone and the array using silence intervals. If the difference resulting from the recording devices is compensated for by digital filtering, a word accuracy as high as 83% is noted for the array speech. From Table 1, it is also seen that feature adaptation by NN is capable of adjusting channel differences and further elevating performance of speech recognition. The elevated score, 94%, is comparable to 98% obtained under the favorable condition.

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 with its own designated NN. The input to the neural net is the cepstral data from the microphone array; the target cep-

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 2: 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.

stral 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), (ii) a non-linear feedforward net, (iii) and a non-linear recurrent network. The input layer of all nets consists of a tapped delay line (9 frames).

Experimental results are summarized in Table 4. 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. We will continue the exploration of different NN paradigms for feature adaptation within the framework of MANN. Specifically, we will evaluate the performance of various network architectures in terms of word recognition accuracy.

5.2. For Simulated Speech Data

In the experiment with simulated reverberant speech, we compare the performance of speech recognition as a function of sound pickup systems and SCNR's. No NN equalization is applied. End-points of distant-talking speech are inferred from the input signal to the room model.

Table 2 gives the word error rates in %, averaged over 20 male and 5 female speakers. It is seen from this Table that matched-filter arrays outperform a single microphone in distant-talking sound pick-up. It can also be seen that the performance of the matched-filter arrays is relatively stable when SCNR drops from ∞ to 0 dB, while the single microphone system simply does not work at SCNR = 0 dB. The results illustrate the capability of matched-filter arrays for combating interfering noise simultaneously present in the room [5]. This capability stems from the selectivity in spatial volume, see also Section 4.2.

Mic Type	*SCNR (dB)*	
	∞	0
1 Single mic	15.	91.
2 MF-MA's	7.8	10.2

Table 2: Word recognition error rates in % as a function of sound pickup systems and levels of competing noise. The DTW recognizer is trained with close-talking microphone input and tested with distant-talking microphone input. (2 MF-MA's: Two matched-filter microphone arrays.)

6. CONCLUSIONS

We have examined performance of MANN for robust speech recognition in variable acoustic environment, including distant-talking, room reverberation, ambient noise interference, and mismatch between the training and testing conditions.

The above evaluation results suggest that the MANN system can (a) provide high-quality sound pickup at distance and effectively mitigate environmental acoustic interference, and (b) without retraining the recognizer, elevate word recognition accuracies of speech recognizers in variable acoustic environments to levels comparable to those obtained for close-talking, high-quality speech. In particular, a matched-filtered array is effective to achieve spatial volume selectivity and mitigation of noise interference. Similar results have also been obtained for studies on speaker recognition [3].

7. ACKNOWLEDGMENTS

This work is supported by ARPA Contract DABT63-93-C-0037. The work is also in part supported by NSF Grant MIP-9121541.

References

[1] Flanagan, J., "Technologies for Multimedia Communications," *Proc. of The IEEE*, pp. 590-603, 1994.

[2] Che, C., Lin, Q., Pearson, J., de Vries, B., and Flanagan, J.: "Microphone arrays and neural networks for robust speech recognition," *Proc. ARPA Human Language Technology Workshop*, March, 1994, Princeton, New Jersey.

[3] Lin, Q., Jan, E., Che, C. and Flanagan, J., "Microphone array and neural network system for speaker identification," *Proc. ARPA Spoken Language Systems Technology Workshop*, March, 1994, Princeton, New Jersey.

[4] Lin, Q., C. Che, and J. French: "Description of CAIP speech corpus," *Proc. Int'l Conf. on Spoken Language Processing*, September, 1994, Japan.

[5] 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.

[6] Jan, E. and Flanagan, J., "Image characterization of acoustic multipath in concave enclosures," CAIP Technical Report-162, Rutgers University CAIP Center, July, 1993.

[7] Srivastava, S., Che, C., and Lin, Q., "End-point detection of microphone-array speech signals," *J. Acous. Soc. Amer. 95 (A)*, pp. 2887, 1994.