ROBUST DISTANT-TALKING SPEECH RECOGNITION*

Q. Lin[1], *C. Che*[1], *D.-S. Yuk*[1], *L. Jin*[1], *B. de Vries*[2], *J. Pearson*[2], *and J. Flanagan*[1]

[1]CAIP Center, Rutgers University, Piscataway, NJ 08855
[2]David Sarnoff Research Center, Princeton, NJ 08543

Most contemporary speech recognizers are designed to operate with close-talking speech and they work best in a quiet laboratory condition. There is an apparent need to render environment robustness to these systems. The objective of the paper is to explore utility of existing speech recognition technology in adverse "real-world" environments for distant-talking applications. A synergistic system consisting of Microphone Array and Neural Network (MANN) is utilized to mitigate environmental interference introduced by reverberation, ambient noise, and channel mismatch between training and testing conditions. The MANN system is evaluated with experiments on continuous distant-talking speech recognition. The results show that the MANN system elevates the word recognition accuracy to a level which is competitive with a *retrained* speech recognizer and that the neural network compensation performs better than some previously researched techniques.

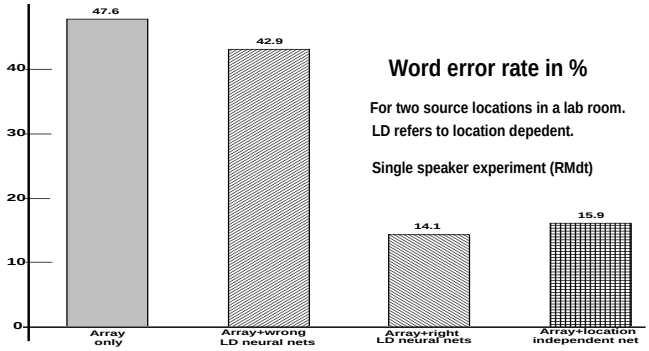

Figure 6. *Effects of source locations on distant-talking recognition (12 ft) using location-dependent (LD) and location-independent NNs.*

reduction in error rate is achieved by the NN compensation technique. For example, the error rate for speaker-independent (gender-dependent) NNs is 11.0%, which is further reduced to 10.5% when the NN compensation is augmented by the SDCN compensation. In Figure 5, it can also be seen that recognition error rate obtained with the speaker-independent NNs is comparable to that of a *retrained* recognizer (7.2%), and is within 7% when compared with the error rate achieved in a close-talking, quite condition.

4.3. Effects of source locations

One of the apparent advantages of microphone array sound capture is the user's freedom of movement around the workplace. It is thus important to quantify effects of source locations on distant-talking speech recognition. For this purpose, another set of distant-talking version of the RM database is collected with the new source located at about 2 m from the source location shown in Figure 2.

Two location-dependent NNs are first trained using exclusively the data from the corresponding source location. As shown in Figure 6, the recognition performance is almost independent of source location when a proper NN was used to compensate for environmental interference. An averaged (over the two locations) word recognition error rate of 14.1% is obtained. It is also shown in Figure 6 that a location-dependent NN cannot be used to compensate for distortion corresponding to another source location. In our measurements, compensation by a wrong NN (42.9%) is equivalent to no compensation processing (47.6%).

Next, a location-independent NN is trained using the adaptation data collected from the both source locations. From Figure 6, it is seen that the location-independent NN is competitive to location-independent NNs.

5. CONCLUSIONS

Increased attention has been devoted to robust distant-talking speech recognition because of its advantageous feature of hands-free operation [2, 5, 4, 8]. Of particular interest is the ability to directly deploy existing speech recognizers trained on close-talking so that expensive and tedious retraining of the recognizers can be avoided.

Two issues then arise. One is how to capture sound signals at distances and the second is how to approximate a matched training and testing condition for the recognizer. In the present research, a line beamforming array is used for distant-talking sound pick up. To compensate for mismatches between close-talking and distant-talking, a neural network is adapted using simultaneously recorded speech signals. Our experiments show that adapting a neural network requires much less speech data than (re-)training a large vocabulary speech recognizer [2, 5, 9]. Our experiments also show that the MANN system elevates recognition performance of distant-talking in a noisy and reverberant environment to a level which is competitive to a *retrained* recognizer.

Both speaker-dependent and speaker-independent (but gender-dependent) NNs have been explored. Because more adaptation data are available for training a speaker-independent NN, better performance is achieved. Speaker-independent NNs have an additional advantage that there is no need to decide who is speaking, and hence make best use of existing speaker independent speech recognition systems.

Effects of source locations on distant-talking speech recognition have also been investigated. Future work will be extended to quantifying effects of automatic source location on the MANN system, and to development of more general region-dependent, location independent NNs. The MANN system will also be compared with other compensation techniques, such as the PMC method of [8].

REFERENCES

[1] Acero, A., and Stern, R., "Environmental robustness in automatic speech recognition," *ICASSP 90*, pp. 849-852.

[2] Che, C., Lin, Q., Pearson, J., de Vries, B., and Flanagan, J., "Microphone Arrays and Neural Networks for Robust Speech Recognition," *Proc. 1994 ARPA HLT Workshop*, New Jersey, pp. 342-347.

[3] Flanagan, J., Berkley, D., Elko, G., West, J., and Sondhi, M., "Autodirective microphone systems," *Acustica 73*, 1991, pp. 58-71.

[4] Giuliani, D., Matassoni, M., Omologo, M., Svaizer, P., "Robust continuous speech recognition using a microphone array," *Eurospeech 95*, pp. 2021-2024, Madrid, Spain, September 1995.

[5] Lin, Q., Che, C., Pearson, J., de Vries, B., and Flanagan, J., "Experiments of distant-talking speech recognition," *Proc. 1995 ARPA SLT Workshop*, pp. 187-192.

[6] Lin, Q., Che, C., and French, J. "Description of the CAIP speech corpus," *Proc. ICSLP 94*, pp. 1823-1826.

[7] Lin, Q., Jan, E., and Flanagan, J. "Microphone arrays and speaker identification," *IEEE-Trans. Speech & Audio Proc.*, Vol. 2, No. 4, 1994, pp. 622-629.

[8] Nakamura, S., Takiguchi, T., Shikano, K. "Noise and room acoustics distorted speech recognition by HMM composition," elsewhere in *ICASSP 96.*

[9] Yuk, D., Che, C., Jin, L., and Lin, Q. "Toward environment-independent continuous speech recognition," elsewhere in *ICASSP 96.*

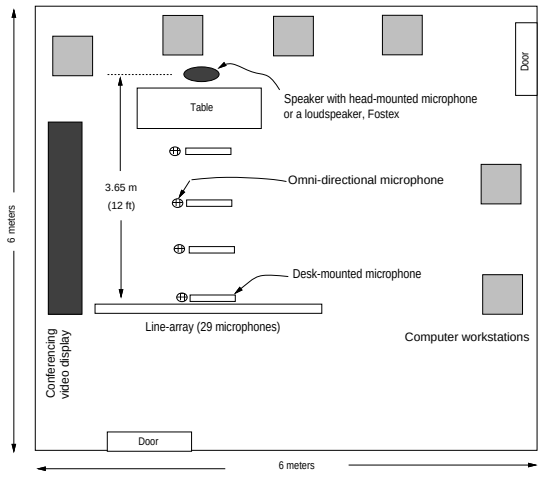

Figure 3. *A top view of the recording environment where the distant-talking speech corpus, RMdt, is collected.*

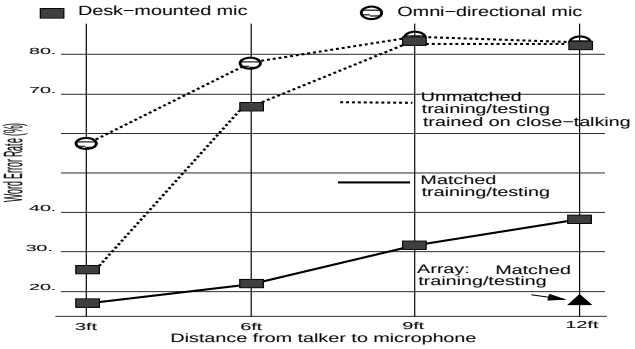

Figure 4. *Word error rates as a function of the distance between the source and the microphone. Recognizer: 3-state triphone models with a Gaussian mixture per state.*

Crown PCC microphone is superior to the omni-directional microphone when the distance is 6 ft and below. At larger distances, the two microphones lead to a similar error rate. It should be noted that the playback volume is kept the same no matter where the microphone is positioned. Background noise contributes to the rapid increase in the error rate when the microphones are moved from 3 ft to 6ft under the unmatched training and testing conditions. It should also be noted that detailed word error rates will be different for different acoustic environments.

Figure 4 also shows that in terms of recognition performance, the line array when positioned 12 ft from the sound source is equivalent to the high-quality desk-top microphone when positioned at a distance of 3 ft.

4.2. Effects of NN compensation

In the ARPA RM database, each of the 12 speakers has a separate set of 10 adaptation sentences, originally designed for speaker adaptation experiments. These sentences are used in the present study to train NNs. The trained NNs are then used for feature adaptation to approximate a matched training and testing condition of the speech recognizer.

In our previous studies [2, 5], speaker-dependent NNs have been used. Each of the NNs is trained using 10 sentences from individual speakers, and may not be well trained because of a limited amount of adaptation speech data. Generally, more speech data are available from multiple talkers than from a single talker. A speaker-independent NN can thus be more adequately trained than a speaker-dependent variant, which in turn can lead to better recognition performance of the MANN system. Furthermore, a speaker-independent NN takes full advantage of speaker-independent speech recognizers without the need to know who has spoken.

	Speaker-dependent (SD)	Gender-dependent (GD)
7 Males	16.9%	13.7%
5 Females	19.1%	16.4%

Table 1. *Distant-talking (array at 12 ft) speech recognition word error rates, using either speaker-dependent NNs or gender-dependent NNs to compensate for environmental mismatches. For speaker-dependent NNs, the results are obtained by averaging all speakers in either gender.*

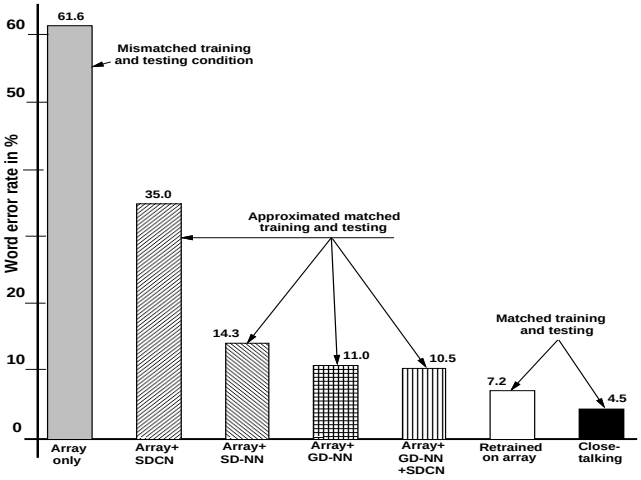

Figure 5. *Distant-talking (12 ft) word recognition error rates in % for different compensation techniques.*

The relative advantages between speaker-dependent and speaker-independent (gender-dependent) NNs are compared in Table 1. The female gender-dependent NN is trained using 50 sentences (each of 5 females has 10 sentences), and the male gender-dependent NN is trained using 70 sentences (each of 7 males has also 10 sentences). It can be seen that the error rate is reduced by 3.3% absolute per cent for male and by 2.7% absolute per cent for female when gender-dependent NNs are utilized.

Word recognition error rates for different conditions are given in Figure 5. The speech recognizer uses gender-independent, 3-state continuous density triphone HMMs, with 7 Gaussian mixtures per state. The HMMs are trained with close-talking speech unless explicitly specified. From Figure 5, it can be seen that without any processing, the distant-talking array speech produces an error rate of 61.6% with a recognizer trained on close-talking speech. The SDCN method [1] reduces the error rate to 35%. Greater

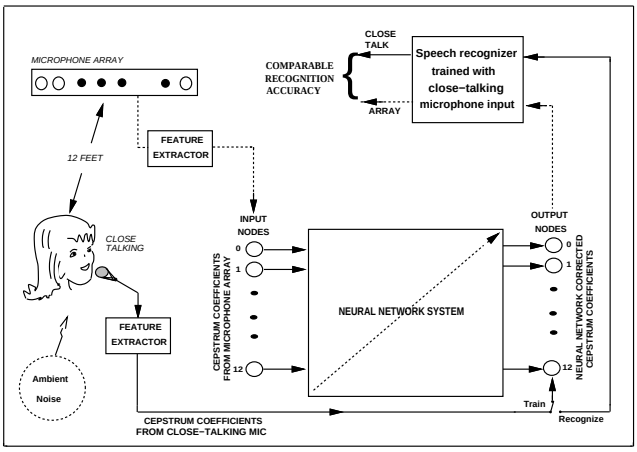

Figure 1. *Block diagram of the MANN system for distant-talking speech recognition using a recognizer trained on close-talking, see text.*

range from some 250 Hz up to 7000 Hz.

2.2. Neural network processors

One of the NN processors we have explored in detail is a fully connected multi-layer perceptron (MLP). The MLP has 3 layers, as depicted in Figure 2. The input and output layers have 13 nodes each, with a node for one of the 13 cepstrum coefficients. The number of hidden nodes can be varied. The neuron activation function of the hidden nodes is a nonlinear sigmoid function, and the activation function of the output layer is a linear function to accommodate the dynamic range of cepstrum coefficients. The input layer incorporates a tapped delay line which allows simultaneous use of previous and preceding frames. Separate measurements show that a input window size of 5 to 7 frames leads to better distant-talking recognition performance. The reader is referred to Yuk et al [9] for more detailed descriptions of different neural network architectures.

Both speaker dependent and speaker independent MLPs have been explored. The NN is trained using a backpropagation method to establish a mapping function of the cepstrum coefficients between distant-talking and close-talking. Therefore, simultaneously-recorded (stereo) data of distant-talking and close-talking is needed for training the NN.

From transformed cepstrum coefficients, the corresponding delta and delta delta cepstrum coefficients can be readily calculated.

3. DISTANT-TALKING CORPUS

To evaluate the MANN system for continuous distant-talking speech recognition, a speech corpus has been generated via loudspeaker playback (Fostex) of the ARPA RM database. Figure 3 shows a top view of the recording environment. The sound signal is captured using the following 3 microphone systems: (i) beamforming line array microphone; (ii) desk-top microphone (Crown PCC160); and (iii) single omni-directional microphone. For the desk-mounted microphone and the omni-directional microphone, 4 distances are selected between the loudspeaker and the sensor. On the other hand, the line array is always fixed at a distance of 12 ft.

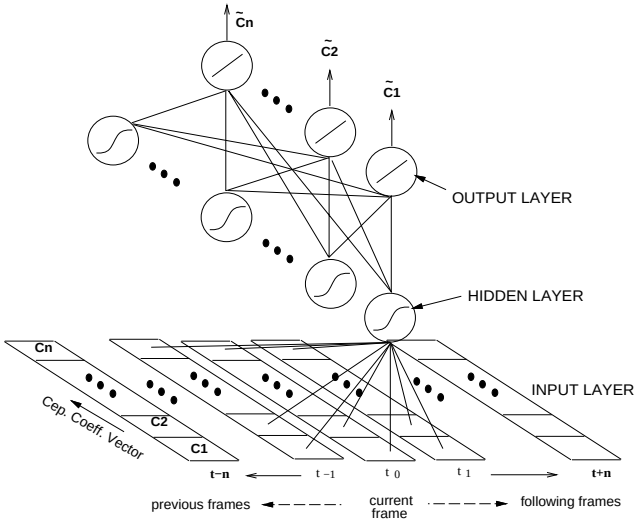

Figure 2. *A feedforward NN (MLP) for mapping the cepstral coefficients of distant-talking to those of close-talking. The neural network accommodates a varying input window length which is typically set to 7 frames.*

The recording enclosure is a hard-walled laboratory room at the CAIP Center. As can be seen in Figure 3, the room is pretty crowded. It has a reverberation time of approximately 0.5 second and the sound pressure level of ambient noise is 50 dBA or 71 dBC, indicating an interfering noise spectrum more tense in lower audio frequencies. A separate corpus with 80 live speakers has also been acquired in the same enclosure [6]. Both corpora are available to speech researchers upon request.

4. EXPERIMENTAL RESULTS

In the following recognition experiments, the Entropic HTK V1.5 speech recognizer is employed. The recognizer uses gender-independent, continuous-density triphone HMM models. The measured acoustic features consist of 12 lifted Mel Frequency Cepstrum Coefficients (MFCC) with mean removal, normalized energy, and their first and second derivatives computed using a 25-ms Hamming window moving every 10 ms. That is, the feature set has 39 elements for each individual frame. The language model is based on word-pair grammar.

The recognizer is trained with all the ARPA RM1 7200 sentences from 12 speakers (7 males and 5 females). During testing, the 300 sentences from the ARPA Oct89 testset are used (25 sentences per each of 12 speakers).

4.1. Effects of talking distance

Figure 4 shows word recognition error rates as a function of the distance between the speaker and the microphone. Results for both matched and unmatched training and testing are included. In the unmatched case, the speech recognizer is trained on close-talking. It can be seen that a matched training/testing condition produces a much lower error rate than unmatched conditions, especially for a distance greater than 3 ft. For mismatched conditions, the

ROBUST DISTANT-TALKING SPEECH RECOGNITION*

Q. Lin[1], *C. Che*[1], *D.-S. Yuk*[1], *L. Jin*[1], *B. de Vries*[2], *J. Pearson*[2], *and J. Flanagan*[1]

[1]CAIP Center, Rutgers University, Piscataway, NJ 08855
[2]David Sarnoff Research Center, Princeton, NJ 08543

ABSTRACT

Most contemporary speech recognizers are designed to operate with close-talking speech and they work best in a quiet laboratory condition. There is an apparent need to render environment robustness to these systems. The objective of the paper is to explore utility of existing speech recognition technology in adverse "real-world" environments for distant-talking applications. A synergistic system consisting of Microphone Array and Neural Network (MANN) is utilized to mitigate environmental interference introduced by reverberation, ambient noise, and channel mismatch between training and testing conditions. The MANN system is evaluated with experiments on continuous distant-talking speech recognition. The results show that the MANN system elevates the word recognition accuracy to a level which is competitive with a *retrained* speech recognizer and that the neural network compensation performs better than some previously researched techniques.

1. INTRODUCTION

High-quality speech input and a matched training/testing condition are two important factors determining performance of speech recognition systems. Therefore, most existing speech recognizers are designed to operate with close-talking microphone input and they work best under a quiet laboratory condition with matched training and testing. The recognition performance is typically degraded when these recognizers are directly deployed for distant-talking speech recognition in variable acoustic environments. The degradation is due to (a) deteriorated speech signal because of multi-path distortion and ambient noise interference; and (b) a mismatched training/testing condition of the recognizers.

This paper describes recent progress in applying existing speech recognition systems, trained on close-talking, to distant-talking speech recognition. To mitigate environmental mismatches between close-talking and distant-talking, a front-end system consisting of a microphone array and a neural network (MANN) is developed. Two synergistic components are included in the MANN system: (1) speech enhancement by microphone arrays to mitigate room reverberation and noise interference; and (2) feature adaptation by neural network processing to approximate a matched training/testing condition for the recognizer.

The MANN system has the following advantages. It allows the user to speak at distances from the microphone without being encumbered by hand-held, body-worn, or tethered microphone equipment. This hands-free advantage is appreciated and sometime necessitated in many hands-busy, eyes-busy applications. Examples include large group conferencing where hands-free sound pick contributes to virtually face-to-face meeting atmosphere. The MANN system also allows more efficient adaptation to new application environments. This is because only the neural network needs to be adapted. As will be shown later, adapting a neural network requires much less speech data than (re-)training a large vocabulary, speaker-independent speech recognizer.

The MANN system has been evaluated on a distant-talking version of the ARPA Resource Management database (RMdt) [5]. In this paper, experimental results are presented and analyzed. The neural network compensation technique is also compared with other compensation techniques. The remainder of this paper is organized as follows: Components of the MANN system are outlined in Section 2; Acquisition of distant-talking speech data is described in Section 3; Experimental results are presented and analyzed in Section 4; Finally, conclusions are given in Section 5.

2. THE MANN SYSTEM

Figure 1 schematically shows the overall system design for robust speech recognition in variable environments, incorporating a microphone array (MA), a neural network (NN), and a speech recognizer which has been trained on close-talking speech. The goal is to achieve a recognition accuracy for distant-talking in noisy/reverberant enclosures which is competitive to that obtained in close-talking, quiet laboratories.

2.1. Beamforming microphone arrays

The microphone array we use is a one-dimensional beamforming line array composed of first-order gradient microphones. It uses direct-path arrivals to produce a single-beam delay-and-sum beamformer [3, 7]. Beamforming arrays effectively combat reverberation and multi-path distortion because the confined beam "sees" fewer sound images in reflecting surfaces. The array consists of 29 gradient sensors, which are non-uniformly positioned in a line (harmonically nested over four octaves). The array has a frequency

*This work is supported by ARPA Contract #DABT63-93-C-0037.