A Neural Network System for Large-Vocabulary Continuous Speech Recognition in Variable Acoustic Environments

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

+CAIP Center, Rutgers University
* David Sarnoff Research Center

Performance of speech recognizers is typically degraded by deleterious properties of the acoustic environment, such as multipath distortion (reverberation) and ambient noise. The degradation becomes more prominent as the microphone is positioned more distant from the speaker, for instance, in a teleconferencing application. Mismatched training and testing conditions, such as frequency response, microphone, signal-to-noise ratio (SNR), and room reverberation, also degrade recognition performance. Among available approaches to handling mismatches between training and testing conditions, a popular one is to retrain the speech recognizer under new environments. Hidden Markov models (HMM) have to date been accepted as an effective classification method for large vocabulary continuous speech recognition, e.g., the ARPA-sponsored SPHINX and DECIPHER. Retraining of HMM-based recognizers is a complex and tedious task. It requires recollection 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.

ARPA Contract (No. DABT63-93-C-0037), entitled "A neural network system for large-vocabulary continuous speech recognition in variable acoustic environments," aims to develop a system of microphone arrays (MA) and neural networks (NN) for robust speech recognition. The system will expand the power and advantages of existing ARPA speech recognizers to practical acoustic environments where users need not be encumbered by hand-held, body-worn, or tethered microphone equipment, and must have freedom of movement. (Examples include Combat Information Centers, large group conferences, and mobile hands-busy eyes-busy maintenance tasks.) Use of MA provides autodirective sound pickup that is higher in quality than conventional microphones used at distances. NN processors learn and compensate for environmental interference, and to adapt the testing condition to the training condition. Recognition performance in hostile acoustic environments can thereby be elevated without the need to retrain the recognizer.

The Contract covers joint research between the CAIP Center and Sarnoff Research Center. It commenced July 19, 1993 and will run for 3 years. To date, a new speech corpus has been created to evaluate the recognition performance of synergistically-integrated systems of MA, NN, and ARPA speech recognizers. (The corpus is available to ARPA contractors upon request.) The corpus consists of 50 male and 30 female speakers. Each speaker speaks 20 isolated command-words, 10 digits, and 10 continuous sentences selected from 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 microphone. The other is for simultaneous recording of a desk-mounted microphone (PZM 160) and the line array microphone. The recording environment is a hard-walled laboratory room of 6x6x2.7 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. The recording is done with an Ariel ProPort with a sampling frequency of 16 kHz and 16-bit linear quantization.

A preliminary NN design is also completed. It consists of 13 multi-layer perceptrons (MLP), one for each of the 13 cepstrum coefficients used in the SPHINX recognizer. Each MLP has 3 layers. The input layer has 9 input nodes, covering the current speech frame and four preceding and four following frames. There are 4 sigmoid nodes in the hidden layer and 1 linear node in the output layer. The NN is trained using backpropagation methods when microphone-array speech and close-talking speech are both available. The trained MLP's are then utilized to transform cepstrum coefficients of array speech to those appropriate to high-quality, close-talking speech. The transformed cepstrum features are used as inputs to the SPHINX speech recognizer.

Promising results have been achieved for recognition of isolated utterances. Recognition accuracy under reverberant conditions is increased from from 13% for desk-mounted microphone to 85% for the MA/NN system. It is also found that the system of MA and NN in unmatched training/testing conditions yields higher word accuracies than those obtained with a *retrained* SPHINX when using array input for both training and testing, i.e., a *matched* training/testing condition. Similar results have been obtained for speaker identification.

Various NN architectures are being explored to identify an optimal one. Dynamic-time-warping (DTW) classification techniques are also being included for experimental comparison. Because DTW recognizers have a simple back end, the capability of the MA/NN system in improving recognition performance can be assessed more reliably. Evaluation experiments for continuous speech are under way. Work is planned to use the DECIPHER recognizer for comparisons. Also planned is a study of feature measures most appropriate for NN adaptation. The CAIP Center has ongoing NSF projects on developing 2-D and 3-D microphone arrays. These new array microphones will later be used in the present work. By the end of the Contract, we will implement and demonstrate a reliable, real-time, hands-free prototype speech recognition system for laboratory evaluation. The Contract work will directly impact the successful applications of ARPA automatic speech recognition systems in adverse practical environments.