ADAPTATION AND COMPENSATION: APPROACHES TO MICROPHONE AND SPEAKER INDEPENDENCE IN AUTOMATIC SPEECH RECOGNITION

Evandro B. Gouvêa, Pedro J. Moreno, Bhiksha Raj, Thomas M. Sullivan, and Richard M. Stern

Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

This paper describes recent efforts by the CMU speech group to address the important problems of robustness to changes in environment and speaker. Results are presented in the context of the 1995 ARPA common Hub 3 evaluation of speech recorded through different microphones at different signal-to-noise ratios (SNRs). For speech that is considered to be of high quality we addressed the problem of speaker variability through a speaker normalization technique. For speech recorded at lower SNRs, we used a combination of environmental compensation techniques previously developed in our group. Speaker normalization reduced the relative error rate for clean speech by 3.5 percent, and the combination of environmental compensation with the use of noise-corrupted speech in the training process reduced the relative error rate for noisy speech by 54.9 percent.

1. INTRODUCTION

Considerable progress has been made in the field of large vocabulary speech recognition in recent years. Good recognition accuracy, however, is far more difficult to achieve when the incoming speech has been recorded in an adverse acoustical environments. For example, a compact implementation of SPHINX-II [3] achieved an error rate of 6.5% in "clean environments" using the evaluation set of the 1994 ARPA common 5000-word CSR evaluations. For speech recorded through "secondary" microphones providing lower SNRs, the best error rate obtained without any environmental compensation was 12.4% [4].

The doubling of error rate observed here is due to two factors: (1) a mismatch between the recording conditions of the training speech and the speech being recognized, and (2) the inherent loss of information due to the presence of noise. While information lost because of additive noise is in principle irretrievable, it is possible to reduce the mismatch between training and testing conditions. Practically all compensation methods are aimed at accomplishing reduction in mismatch.

A second major source of error in large vocabulary speech recognition systems is variability among speakers. This problem is usually approached by either modifying the internal models of a recognition system to adapt them to a new speaker or by normalizing the representation of speech from a new speaker to match more closely those from previously-defined prototype speakers.

The 1995 ARPA Hub 3 task was designed to evaluate the performance of speech recognition systems on unlimited-vocabulary read speech, both in clean and noisy recording conditions. Speech was recorded in sessions of 15 sentences using a variety of microphones, one of which was a Sennheiser 410 close-talking microphone (representing a "clean" recording environment). Although session boundaries were known, it was not known *a priori* whether a given session was recorded using a noisy microphone or using the close-talking microphone.

In this paper we describe and compare the performance of a series of training and compensation procedures that were developed to improve the recognition accuracy of the CMU SPHINX-II speech recognition system, especially in noisy recording environments, in the context of the 1995 ARPA Hub 3 task. We also describe our attempts at compensating for the variability between speakers using a speaker normalization technique that is applied to clean speech and by using session-wise speaker adaptation for speech recorded in noisy conditions. All experiments were performed using the development and evaluation sets of the 1995 ARPA Hub 3 speech corpus.

In Section 2 we describe the technique used to separate clean speech from noisy speech, the first step in the processing of incoming speech. In Section 3 we describe our approach to the recognition of clean speech, namely, speaker normalization. In Section 4 we describe in detail the compensation techniques used for the noisy speech. These techniques involve cepstra compensation approaches (*e.g.* the CDCN algorithm) as well as methods that modify the statistics of the internal distributions of the HMMs (*e.g.*the STAR algorithm). In Section 5 we present and discuss official results for the 1995 ARPA Hub 3 task. We also describe in this section some additional experiments that evaluate the performance of the STAR algorithm and direct Baum-Welch adaptation.

2. SEPARATION OF CLEAN AND NOISY SPEECH

Since it was not known *a priori* whether a given session of speech was recorded over the close-talking microphone or over some other (more noisy) microphone, incoming speech in a given session was first separated into two classes, "clean" (representing speech from the Sennheiser 410 microphone) and "noisy" (representing all other speech). Classification was performed on the basis of a single feature, the difference between the minimum and maximum values of the zeroth-order cepstral coefficient during the course of an utterance.

The zeroth-order cepstral coefficient is a function of the energy in the frame. Therefore, the minimum value of the zeroth-order cepstral coefficient in an utterance is a function of the noise energy, while its maximum is a function of the signal energy. It follows that the difference between the maximum and the minimum zeroth-order cepstral coefficients is a measure of the signal-to-noise ratio (SNR).

The classifier modeled the zeroth-order cepstral coefficient using a Gaussian mixture density with 8 components. Separate models were created for noisy and clean speech, using development set data. The test set utterances were classified as clean or noisy based on a maximum likelihood criterion.

The algorithm achieved perfect classification of the data in the development set, when the classification was performed on a session basis (*i.e.* when all 15 sentences from a given session were used for classification). When the classification was performed on a per-sentence basis, the classification error was only 0.33%.

Sessions classified as being "clean" were processed differently from sessions classified as being "noisy", as is described in Sections 3 and 4, respectively.

3. PROCESSING OF CLEAN SPEECH

Speaker variability is a major source of performance degradation on speech recognition systems, which is why speaker-dependent speech recognition systems typically outperform speaker independent systems. Speaker normalization techniques attempt to address this problem by mapping features representing speech from a new speaker to those of a previously-determined standard speaker. For example, systems that use a bandpass filter bank to accomplish peripheral frequency analysis can partially account for speaker variability by warping the center frequencies of the analysis filters. The approach we adopted operates in similar fashion, as we modify the center frequencies of the triangular weighting functions used to derive the spectral energy estimates in narrow frequency bands from which mel-frequency cepstral coefficients are derived by performing an inverse Fourier transform (*e.g.* [*2*]). We exploit such warping functions in an attempt to achieve speaker normalization.

The speaker normalization algorithm we implemented was previously described by Roth *et al.* [6] and Wegmann *et al.* [7], and was motivated by a series of seminal experiments by Cohen *et al.* [2]. The first step in our implementation consists of an iterative derivation of the Gaussian mixture model that statistically represents a prototype speaker. At each step in this iteration, we first compute the Gaussian mixture model for the prototype speaker and then find the optimal warping function for each speaker. This optimal warping function is chosen to be the one that maximizes the *a posteriori* log-likelihood computed using this Gaussian mixture model. After choosing the optimal warping function for all speakers in the set, we re-compute the Gaussian mixture model, find new optimal warping functions based on the new Gaussian mixture model, and proceed in this manner until convergence is achieved. Convergence in this case means that for consecutive iterations the same optimal warping function is chosen for every speaker.

The second step of the implementation concerns the derivation of speaker-normalized HMMs from the optimally-warped training set. The training steps of our implementation consist of the iterative derivation of the Gaussian mixture model and the derivation of the HMMs.

During recognition, the best warping function is obtained for each new speaker and the warped utterances are recognized using the previously-derived HMMs.

3.1. Training

All experiments performed during system training used the Wall Street Journal SI-284 corpus. The 284 speakers in this corpus were partitioned into two subsets. Each subset was alternately worked on, finding the optimal warping function for each speaker and computing the Gaussian mixture model that best fitted the optimally-warped speakers in that subset. The optimal warping functions for the speakers in one set were estimated using the Gaussian mixture model produced with the other set. We partitioned the training set in an attempt to avoid the fine tuning of warping functions to the training set..

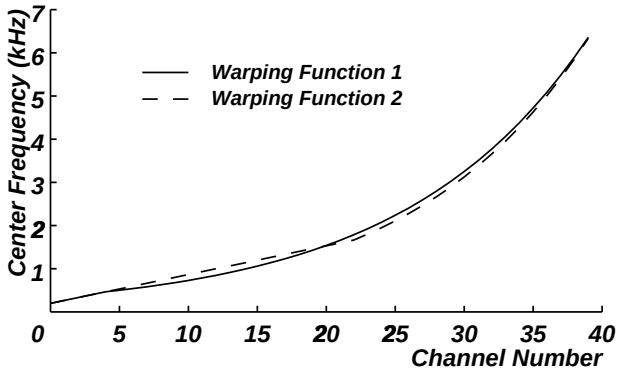

Figure 1. Comparison of two frequency-warping functions used in speaker adaptation.

Twelve frequency-warping functions were chosen arbitrarily. They are linear at low frequencies and exponentially-shaped at higher frequencies. Figure 1 depicts two extreme warping functions used in this procedure. The optimal warping function for each of the speakers in a subset was chosen in the following manner: all utterances by a particular speaker were warped using each of the 12 pre-selected warping functions; the warping function that maximized the average *a posteriori* likelihood of the incoming cepstra was selected as the optimal warping function for that speaker; this *a posteriori* likelihood was computed on a 64-component Gaussian mixture distribution produced from all the normalized training data in the other subset. Figure 2 illustrates the training procedure.

The procedure described here relies on the existence of a Gaussian mixture model representing the prototype normalized speaker. To bootstrap the training procedure we used all the unwarped data in one of the subsets to construct an initial Gaussian mixture distribution. Warping functions and Gaussian models were iteratively derived in the subsequent normal training procedure.

Convergence is achieved when in successive iterations all speakers are matched to exactly the same warping function as in the previous iteration. When convergence is achieved both the optimal warping associated with each speaker and the Gaussian mixture models for the two subsets do not change any more with further iterations. At the end of the iterative process, we have two Gaussian mixture models, and the choices of warping function for each

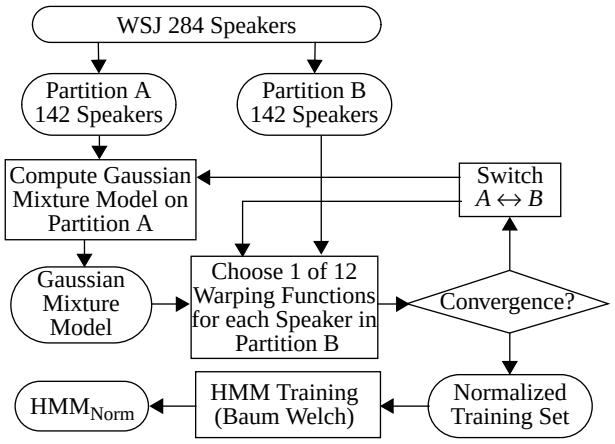

Figure 2. Block diagram of the training process for speaker normalization.

speaker in both partitions of the training data. The estimation of the Gaussian mixture model was performed one final time on the entire set of data using the selected warping functions to produce a single Gaussian mixture model that was used for the recognition phase.

Figure 3 shows the number of speakers for which the best-matched warping function changes from iteration to iteration, as a function of iteration number. Curves for each of the two subsets used for training are plotted separately.

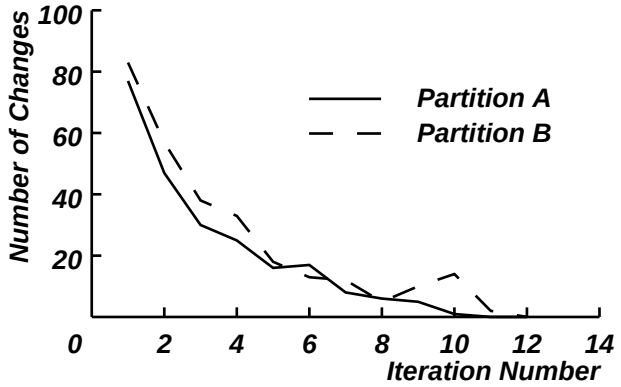

Figure 3. Changes in speakers across cluster, at each iteration of the normalization algorithm.

As an informal measure of the consistency of the algorithm's convergence, we checked for possible correlations between the warping functions chosen for a speaker and the speaker's gender. Figure 4 shows the distributions of labels associated with each of the warping functions, separated by speaker gender. Labels represent warping frequencies linearly spaced along the frequency axis. We can see a clear separation of these distributions, with medians towards one end for male speakers and towards the other end for female speakers.

With optimal warping functions iteratively selected for each speaker, we created HMMs for a generic normalized speaker. The number of iterations of the Baum-Welch algorithm was chosen based on word error rate. This progression is shown on Figure 5.

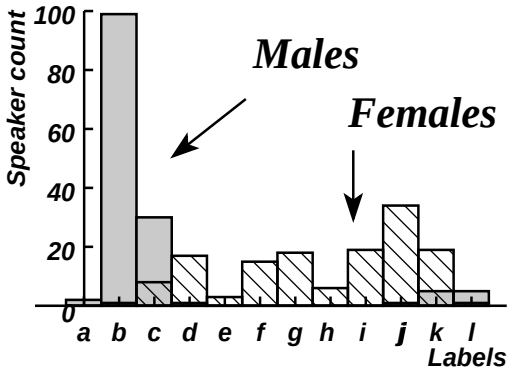

Figure 4. Speaker gender distributed by labels. Labels represent warping frequencies linearly spaced.

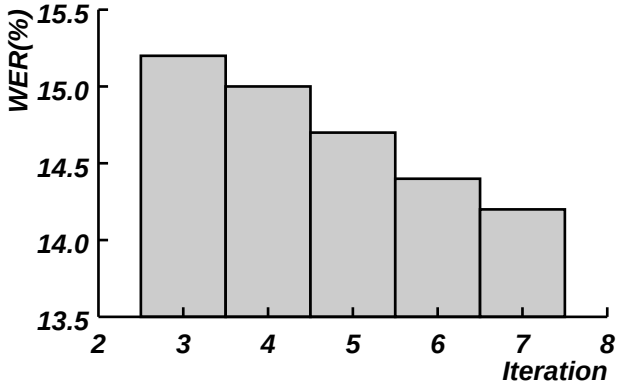

Figure 5. Word error rate with iterations of Baum-Welch. Note the scale of the vertical axis.

3.2. Testing

Using the final Gaussian mixture model obtained in the training process, we chose the warping function that maximized the likelihood across all sentences in the session for each speaker in the test set. The mel-cepstral parametrization of the speaker using this warping function was then used to for recognition.

4. PROCESSING OF NOISY SPEECH

Our approach for recognizing speech that is identified as being noisy is based on the premise that the best recognition accuracy is obtained when the training and testing conditions are comparable. We attempted to achieve this first by training HMMs to model the noise conditions of the noisy speech, and second by further adapting these "noisy" HMMs on a session basis using a combination of CDCN [1] and STAR [5].

4.1. Training "Noisy" HMMs

Two sets of acoustic models, representing males and females, were generated. These models were obtained from the SI-284 WSJ0 and WSJ1 corpora, corrupted by additive noise at a global SNR of 12.5 dB. The noise used to corrupt the clean speech was colored to simulate the noise conditions of the development test set. The additive

noise was generated by passing white noise through a 512-point FIR filter that had a power spectrum equal to the estimated power spectrum of the background recording noise (Figure 6).

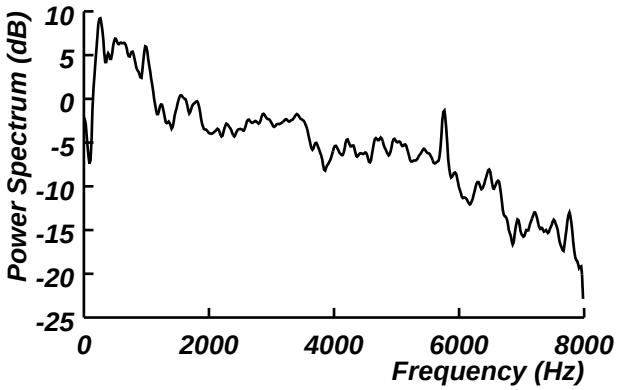

Figure 6. Power spectrum of the colored noise used to contaminate the WSJ training set.

The power spectrum of the background noise was estimated as the average of the power spectra of the 3-second samples of ambient noise supplied by NIST for each of the speakers and microphones in the development set normalized by the estimate of the channel spectrum for that microphone. The transfer function of the channel spectrum for a microphone was estimated as the ratio of the average power spectrum of the samples of speech recorded through the microphone divided by the average power spectrum of samples of speech from Microphone A in the development set. The 12.5-dB SNR noise-corrupted files were then used to obtain HMMs for speech at a nominal SNR of 12.5 dB. Five iterations of the Baum-Welch algorithm were performed to produce gender-dependent HMMs

4.2. Environmental Compensation

The session based environmental compensation procedure was performed in two steps.

CDCN: The first stage of compensation consisted of running the testing utterances through CDCN [1]. CDCN (Figure 7) is a maximum-likelihood algorithm that uses a generic model of unknown additive noise and unknown linear filtering for the noisy speech. The algorithm attempts to estimate the parameters of the noise and filter that best map a set of reference statistics, normally the statistics of the speech used to train the HMMs, onto the noisy speech. The cepstrum of the compensated speech is then estimated from the noisy speech using an MMSE criterion [1].

In this particular implementation of CDCN, the reference statistics used by the algorithm were computed on the 12.5-dB speech used to train the "noisy" HMMs. This causes the algorithm to map any noisy sentence to a "standard" environment with a 12.5-dB SNR.

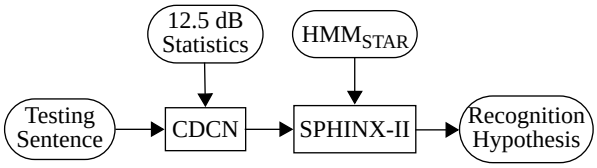

Figure 7. Block diagram of the compensation model using CDCN.

STAR: STAR is an algorithm that uses stereo pairs, *i.e.* sentences recorded simultaneously in the testing environment and in the training environment. Stereo sets of speech from the training environment and noisy CDCN-normalized speech were used to adapt the means and covariances of the HMM distribution using STAR.

Because appropriate stereo recordings of the noisy and clean speech in this dataset are not available, pseudo-stereo pairs were generated to simulate the conditions of the CDCN-normalized noisy data (Figure 8). Specifically, 100 utterances of clean speech from WSJ0 were corrupted to the noise and channel conditions of the session. We used the noise samples provided by NIST to estimate the spectrum of the noise. The channel characteristics were estimated by computing the difference between the mean cepstral coefficients of the actual 15 sentences in a session and the corresponding means from the 12.5-dB training data. This corrupted speech was then compensated to the 12.5-dB reference statistics using CDCN. These data represented the equivalent of the CDCN-compensated noisy data in our pseudo-stereo set.

The same 100 clean utterances from WSJ0 were also corrupted by additive noise at an SNR of 12.5 dB in a manner identical to that used to obtain data for training the 12.5 dB HMMs. These data represented the training-environment counterpart of the noise-corrupted CDCN-compensated data in the "stereo" set.

The combination of the corrupted CDCN-compensated utterances and the corresponding 12.5-dB corrupted utterances was used to compute new HMMs from the original (12.5-dB) HMMs using STAR.

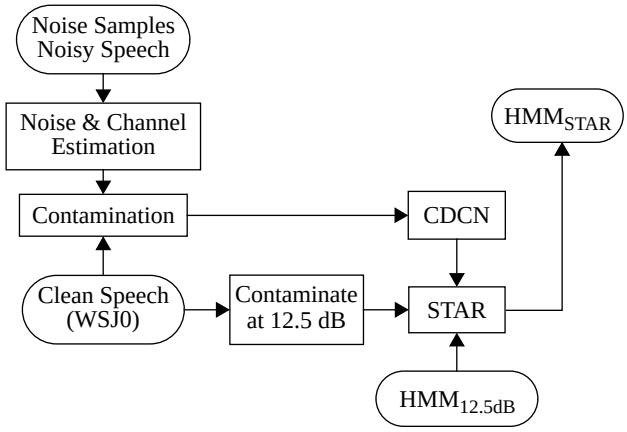

Figure 8. Block diagram of the HMM derivation using STAR.

In recognizing new speech, the incoming waveform is compensated using CDCN and then recognized using the STAR-adapted models.

5. SYSTEM PERFORMANCE ON THE 1995 ARPA HUB 3 EVALUATION DATA

5.1. Baseline Performance

The officially-reported results for the primary test set for the 1995 ARPA CSR Hub 3 evaluation were 13.9% and 29.2% word error rates for the clean and noisy subsets of the evaluation data, respectively. We also ran several additional experiments to evaluate the improvements in recognition error rate provided by each of the components of our compensation strategy, and these results are summarized in Table 1 below.

Recognition times for the H3 task were about 70 times real time for noisy speech and 29 times real time for clean speech. This includes the time required to select the optimal warping in the case of clean speech, and the time required for CDCN-compensation and STAR adaptation in the case of the noisy speech. If we did not adapt the 12.5 dB-HMMs using STAR, *(i.e.* if we used only a combination of CDCN and the 12.5 dB models), the recognition times for noisy speech were only about 36 times real time.

5.2. Performance without Adaptation to Speakers and Environments

To evaluate the improvement provided by the speaker-normalization procedure we recognized the clean speech of the evaluation set on gender-specific models generated without speaker normalization. We also recognized the noisy evaluation data using clean speech HMMs and on the 12.5-dB HMMs to evaluate the improvement obtained by using the 12.5-dB models over the baseline system. These and other results described below are tabulated in Table 1. It was found that the speaker normalization procedure used for the evaluation reduced the relative error rate for clean speech by 3.5 percent. Environmental compensation reduced the relative error rate for noisy speech by 11.2 percent and the combination of noise-corrupted HMMs and environmental compensation reduced the relative error rate for noisy speech by 54.9 percent.

5.3. Impact of Pseudo-Stereo Data Used in STAR Adaptation

To evaluate the improvement obtained by adapting the 12.5-dB models to the CDCN-compensated speech we ran experiments without using the STAR algorithm (*i.e.* we directly recognized the CDCN-compensated speech using the 12.5 dB HMMs). Surprisingly, the recognition accuracy without STAR was found to be greater than that obtained using STAR adaptation of the models. We attribute this anomaly to the fact that our method of generating pseudo-stereo pairs for STAR adaptation was imperfect. While the pseudo-stereo pairs were obtained by corrupting clean speech with estimated channel and noise conditions for the session to be recognized, this corruption was done in the mel-frequency log-spectral domain where the components are actually obtained by integrating the power spectrum over the Mel frequency bands. To confirm this hypothesis we later conducted a test in which the STAR adaptation of the models was performed with "perfect" stereo pairs calculated from direct comparisons of the clean speech and noisy speech in the evaluation sets. (According to the rules of the evaluation, this side information was not available to the recognition system.) The recognition error rate obtained with perfect stereo, 20.6%, was considerably lower than the rate obtained with the pseudo-stereo pairs.

5.4. Baum-Welch Session Adaptation

After the official evaluations were completed we performed a series of "session-adaptation" experiments in which we adapted the 12.5-dB HMMs to each speaker's session using the Baum-Welch algorithm. Unlike other recognition systems used in the 1995 Hub 3 evaluation, SPHINX-II uses a semi-continuous HMM structure with only 256 elements in its codebook of distributions [3]. This small number of parameters makes the use of clustering techniques unnecessary.

Like most other adaptation or training techniques, Baum-Welch session adaptation requires orthographic transcriptions. Since transcriptions are not available for test data, we generated the transcriptions automatically by recognizing the noisy speech in a first pass using the non-adapted 12.5-dB HMMs. With these transcriptions we followed the normal adaptation procedure to produce a new set of adapted mean vectors and covariance matrices. With these new statistics we performed recognition on the same data used for adaptation. The procedure could be iterated as new and more accurate transcriptions were produced with each new set of means and variances. The final recognition error rate obtained in this case was 23.3% for the noisy subset and 12.4% for the clean subset of the 1995 H3 evaluation set, as shown in Table 1.

Since transcriptions generated by the decoder are inevitably errorful, we estimated the lower bound in error rate that can be provided by this technique by running a second experiment in which we assume that the Baum-Welch adaptation procedure could make use of "perfect" knowledge of the correct transcriptions. In these experiments we adapted the means and variances of all the Gaussians of the 12.5-dB HMMs, using all 15 sentences in a given session and the correct transcriptions. The recognition error was observed to stabilize after 5 iterations of Baum-Welch adaptation to 16.4% and 8.6%, respectively, for the noisy and clean subsets of the 1995 H3 evaluation set.

	Clean Speech	**Noisy Speech**
Clean-speech HMMs, no speaker or environment adaptation	14.4%	64.8%
Clean-speech HMMs, with speaker normalization (official result)	**13.9%**	–
12.5-dB HMMs, no environment adaptation	–	32.9%
12.5-dB HMMs, CDCN and STAR (official result)	–	**29.2%**
12.5-dB HMMs, CDCN only	–	27.5%
12.5-dB HMMs, STAR with "perfect" stereo information	–	20.6%
Baum-Welch adaptation with decoder-generated transcriptions	12.4%	23.3%

	Clean Speech	Noisy Speech
Baum-Welch adaptation with "perfect" transcriptions	8.6%	16.4%

Table 1: Error rates on the 1995 ARPA Hub 3 evaluation using alternate speaker and environment adaptation strategies. Officially-reported scores for the primary tasks are shown in bold face. Contrast conditions labelled "perfect" include side information not available in the actual evaluation.

6. SUMMARY AND CONCLUSIONS

In this paper we describe several procedures that have been employed to ameliorate the adverse effects of unknown microphones in noisy environments and speaker variability in large-vocabulary speech recognition systems. The training and adaptation procedures used for the official ARPA evaluations provided relative decreases in error rate of 3.5 and 54.9 percent for clean and noisy speech, respectively.

Despite the considerable benefits that can be provided by conventional speaker and environment adaptation, we also observed that the greatest improvement in recognition accuracy can be obtained by simply re-training the HMMs using techniques such as session-based Baum-Welch adaptation. Nevertheless, Baum-Welch adaptation is not a viable alternative for most systems that require real-time operation, and for such systems a combination of a well-trained HMM and a compensation technique such as CDCN would provide the best recognition accuracy.

ACKNOWLEDGMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. Evandro Gouvêa has a scholarship from the CNPq (Conselho Superior de Desenvolvimento Científico e Tecnológico) - Brazil. We thank Sam-Joo Doh, Juan Huerta, Uday Jain and Matthew Siegler for their help on this research and the rest of the speech group for their contributions to this work.

REFERENCES

1. Acero, A., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.
2. Cohen, J., Kamm, T., Andreou, A., "An Experiment on Vocal Tract Variability", *Proceedings of the CAIP Workshop: Frontiers of Speech Recognition*, Aug 1994.
3. Huang, X., Alleva, F. A., Hon, H.-W., Hwang, M.-Y., Lee, K.-F. and Rosenfeld, R., "The Sphinx-II Speech Recognition System: An Overview", Computer Speech and Language, Vol.2, pp. 137-48.
4. Moreno, P. J., Jain, U., Raj, B., Stern, R. M., "Approaches to microphone independence in Automatic Speech Recognition", *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, pp. 74-6, Jan. 1995.
5. Moreno, P. J., Raj, B., and Stern, R. M., "A Unified Approach to Robust Speech Recognition", *EUROSPEECH-95*, Madrid, Spain, pp. 481-4,Sep. 1995.
6. Roth, R., Gillick, L., Orloff, J., Scattone, F., Gao, G., Wegmann, S., and Baker, J., "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer", *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, pp. 116-120, Jan. 1995.
7. Wegmann, S., McAllaster, D., Orloff, J., Peskin, B., "Speaker Normalization on Conversational Telephone Speech", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I-339 – I-342, May 1996.