SIGNAL PROCESSING FOR ROBUST SPEECH RECOGNITION

Fu-Hua Liu, Pedro J. Moreno, Richard M. Stern, Alejandro Acero

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

ABSTRACT

This paper describes a series of cepstral-based compensation procedures that render the SPHINX-II system more robust with respect to acoustical environment. The first algorithm, phone-dependent cepstral compensation, is similar in concept to the previously-described MFCDCN method, except that cepstral compensation vectors are selected according to the current phonetic hypothesis, rather than on the basis of SNR or VQ codeword identity. We also describe two procedures to accomplish adaptation of the VQ codebook for new environments, as well as the use of reduced-bandwidth frequency analysis to process telephone-bandwidth speech. Use of the various compensation algorithms in consort produces a reduction of error rates for SPHINX-II by as much as 40 percent relative to the rate achieved with cepstral mean normalization alone, in both development test sets and in the context of the 1993 ARPA CSR evaluations.

1. INTRODUCTION

A continuing problem with current speech recognition technology is that of lack of robustness with respect to environmental variability. For example, the use of microphones other than the ARPA standard Sennheiser HM-414 "close-talking" headset (CLSTLK) severely degrades the performance of systems like the original SPHINX system, even in a relatively quiet office environment [*e.g.* 1,2]. Applications such as speech recognition in automobiles, over telephones, on a factory floor, or outdoors demand an even greater degree of environmental robustness.

In this paper we describe and compare the performance of a series of cepstrum-based procedures that enable the CMU SPHINX-II [8] speech recognition system to maintain a high level of recognition accuracy over a wide variety of acoustical environments. We also discuss the aspects of these algorithms that appear to have contributed most significantly to the success of the SPHINX-II system in the 1993 ARPA CSR evaluations for microphone independence (Spoke 5) and calibrated noise sources (Spoke 8).

In previous years we described the performance of cepstral mapping procedures such as the CDCN algorithm, which is effective but fairly computationally costly [2]. More recently we discussed the use of cepstral highpass-filtering algorithms [such as the popular RASTA and cepstral-mean-normalization algorithms (CMN) [6]. These algorithms are very simple to implement but somewhat limited in effectiveness, and CMN is now part of baseline processing for the CMU and many other systems.

In this paper we describe several new procedures that when used in consort can provide as much as an additional 40 percent improvement over baseline processing with CMN. These techniques include:

- Phone-dependent cepstral compensation
- Environmental interpolation of compensation vectors
- Codebook adaptation
- Reduced-band analysis for telephone-bandwidth speech.
- Silence codebook adaptation

In Sec. 2 we describe these compensation procedures in detail, and we examine their effect on recognition accuracy in Secs. 3 and 4.

2. ENVIRONMENTAL COMPENSATION ALGORITHMS

We begin this section by reviewing the previously-described MFCDCN algorithm, which is the basis for most of the new procedures discussed. We then discuss blind environment selection and environmental interpolation as they apply to MFCDCN. The complementary procedures of phone-dependent cepstral normalization and codebook adaptation are described. We close this section with brief description of reduced-bandwidth analysis and silence-codebook adaptation, which are very beneficial in processing telephone-bandwidth speech and speech recorded in the presence of strong background noise, respectively.

2.1. Multiple Fixed Codeword-Dependent Cepstral Normalization (MFCDCN)

Multiple fixed codeword-dependent cepstral normalization (MFCDCN) provides additive cepstral compensation vectors that depend on signal-to-noise ratio (SNR) and that also vary from codeword to codeword of the vector-quantized (VQ) representation of the incoming speech at each SNR [6]. At low SNRs these vectors primarily compensate for effects of additive noise. At higher SNRs, the algorithm compensates for linear filtering, while at intermediate SNRs, they compensate for both of these effects. Environmental independence is provided by computing compensation vectors for a number of different environments and selecting the compensation environment that results in minimal residual VQ distortion.

Compensation vectors for the chosen testing environment are applied to normalize the utterance according to the expression

$\hat{\mathbf{x}}_t = \mathbf{z}_t + \mathbf{r}[k_t, l_t, e]$ where k_t, l_t, t, and e are the VQ codeword index, instantaneous frame SNR, time frame index and the index of the chosen environment, respectively, and $\hat{\mathbf{x}}_t$, $\mathbf{z}_t$, and $\mathbf{r}$ are the compensated (transformed) data, original data and compensation vectors, respectively.

2.2. Blind Environment Selection

In several of the compensation procedures used, including MFCDCN, one of a set of environments must be selected as part of the compensation process. We considered three procedures for environment selection in our experiments.

The first procedure, referred to as *selection by compensation*, applies compensation vectors from each possible environment successively to the incoming test utterance. The environment e is chosen that minimizes the average residual VQ distortion over the entire utterance. In the second approach, referred to as *environment-specific VQ*, environment-specific cookbooks are generated from the original uncompensated speech. By vector quantizing the test data using each environment-specific codebook in turn, the environment with the minimum VQ distortion is chosen. The third procedure, referred to as *Gaussian environment classifier*, models each environment with mixtures of Gaussian densities. Environment selection is accomplished so that the test data has the highest probability from the corresponding classifier. This latter approach is similar to one proposed previously by BBN [7].

All three methods produce similar speech recognition accuracy.

2.3. Interpolated FCDCN (IFCDCN)

In cases where the testing environment does not closely resemble any of the particular environment used to develop compensation parameters for MFCDCN, interpolating the compensation vectors of several environments can be more helpful than using compensation vectors from a single (incorrect) environment. As in MFCDCN, compensation vectors used in the *interpolated fixed codeword-dependent cepstral normalization* algorithm (IFCDCN) are precomputed for environments in the training database in the estimation phase. Compensation vectors for new environments are obtained by linear interpolation of several of the MFCDCN compensation vectors:

$$\hat{r}[k,l] = \sum_{e=1}^{E} f_e \cdot r[k,l,e]$$

where $\hat{r}[k,l]$, $r[k,l,e]$, and f_e are the estimated compensation vectors, the environment-specific compensation vector for the e^{th} environment, and the weighting factor for the e^{th} environment, respectively.

The weighting factors for each environment are also based on residual VQ distortion:

$$f_e = \frac{p(e|\bar{Z})}{\sum_{j=1}^{E} p(j|\bar{Z})} = \frac{exp\{D_e/(2\sigma^2)\}}{\sum_{j=1}^{E} exp\{D_j/(2\sigma^2)\}}$$

where σ is the codebook standard deviation using speech from the CLSTLK microphone, $\bar{Z}$ represents the testing utterance, and D_j and D_e are the residual VQ distortions of the j^{th} and e^{th} environments. We have generally used a value of 3 for E.

2.4. Phone-Dependent Cepstral Normalization (PDCN)

In this section we discuss an approach to environmental compensation in which additive cepstral compensation vectors are selected according to the current phoneme hypothesis in the search process, rather than according to physical parameters such as SNR or VQ codeword identity. Since this phoneme-based approach relies on information from the acoustic-phonetic and language models to determine the compensation vectors, it can be referred to as a "back-end" compensation procedures, while other approaches such as MFCDCN which work independently of the decoder can be regarded as "front-end" compensation schemes.

Estimation of PDCN compensation Vectors. In the current implementation of *phone-dependent cepstral normalization* (PDCN), we develop compensation vectors that are specific to individual phonetical events, using a base phone set of 51 phonemes, including silence but excluding other types of non-lexical events. This is accomplished by running the decoder in supervised mode using CLSTLK data and correction transcriptions. All CLSTLK utterances are divided into phonetic segments. For every phonetic label, a difference vector is computed by accumulating the cepstral difference between the CLSTLK training data, $\mathbf{x}_t$, and its noisy counterpart, $\mathbf{z}_t$. Compensation vectors are computed by averaging the corresponding difference vector as follows,

$$c[p] = \frac{\sum_{u=1}^{A}\sum_{t=1}^{T_u}(\mathbf{x}_t - \mathbf{z}_t)\,\delta(f_t - p)}{\sum_{u=1}^{A}\sum_{t=1}^{T_u}\delta(f_t - p)}$$

where f_t is the phoneme for frame t, p the phoneme index and T_u length of the uth utterance out of A sentences.

Compensation of PDCN in Recognition. The SPHINX-II system uses the senone [4,8], a generalized state-based probability density function, as the basic unit to compute the likelihood from acoustical models. The probability density function for senone s in frame t for the cepstral vector $\mathbf{z}_t$ of incoming speech can be expressed as

$$Pr(\mathbf{z}_t|s) = \sum_{m_t=1}^{B} w_{m_t} N(\mathbf{z}_t; \mu_{m_t}, \sigma_{m_t})$$

where m_t stands for the index of the best B Gaussian mixtures of senone s for cepstra vector $\mathbf{z}_t$, and μ_{m_t}, σ_{m_t}, and w_{m_t} are the corresponding mean, standard deviation, and weight for the m_t^{th} mixture of senone s. Multiple compensated cepstral vectors are formed in PDCN by adding various compensation vectors to incoming cepstra, $\hat{\mathbf{x}}_{t,p}(t)$, where $\hat{\mathbf{x}}_{t,p} = \mathbf{z}_t + \mathbf{c}[p]$ on a frame-by-frame basis for the presumed phoneme index, p.

The amount of computation needed for this procedure is reduced because in SPHINX-II, each senone corresponds to only one distinctive base phoneme. A cepstral vector can be normalized with a proper PDCN compensation factor corresponding to the particular base phonetical identity. As a result, senone probabilities can be

calculated by the presumed phonetic identity that corresponds to a given senone. Using this approach, the senone probability in PDCN is re-written as

$$Pr(\hat{\mathbf{x}}_{t,p}|\, s) = \sum_{n_t=1}^{B} w_{n_t} N(\hat{\mathbf{x}}_{t,p}; \mu_{n_t}, \sigma_{n_t})$$

where n_t is the index of the best B Gaussian mixtures for senone s at frame t with respect to the PDCN-normalized cepstral vector $\hat{x}_{t,p}$, for the corresponding phonetic label p for senone s.

Interpolated PDCN (IPDCN). PDCN, like SDCN and FCDCN [3,6], assumes the existence of a database of utterances recorded in stereo in the training and testing environments. In situations where no data from any particular testing environment is available for estimation, IPDCN is desirable. Based on an ensemble of pre-computed PDCN compensation vectors, IPDCN applies to the incoming utterance an interpolation of compensation vectors from several of the closest environments (IPDCN). The interpolation is performed in the same way that it was for IFCDCN. In the current implementation, we use the 3 closest environments with the best 4 Gaussian mixtures in interpolation.

2.5. Codebook Adaptation (DCCA and BWCA)

A vector quantization (VQ) codebook, which is a set of mean vectors and/or co-variance matrices of cepstral representations, also exhibits some fundamental differences when mismatches are encountered between training and testing environments [7]. This suggests that when such mismatches exist, the codebook can be "tuned" to better characterize the cepstral space of testing data. In this section, we propose two different implementations of such codebook adaptation.

Dual-Channel Codebook Adaptation (DCCA). *Dual-Channel Codebook Adaptation (DCCA)* exploits the existence of speech that is simultaneously recorded using the CLSTLK microphone and a number of secondary microphones. From the viewpoint of front-end compensation, the senone probability density function can be expressed as the Gaussian mixture

$$p_{s_t} = \sum_{k=1}^{B} w_k N(\hat{\mathbf{x}}_t; \mu_k, \sigma_k) = \sum_{k=1}^{B} w_k N(\mathbf{z}_t + \delta\mathbf{z}_t; \mu_k, \sigma_k)$$

where k, $\mathbf{z}_t$, $\delta\mathbf{z}_t$, $\hat{\mathbf{x}}_t$, μ_k, σ_k are the mixture index among top B mixtures, noisy observation vector, compensation vector, compensated vector, mean vector for the k^{th} mixture, and variance, respectively. The senone probability density function is re-written as

$$\begin{aligned} p_{s_t} &= \sum_{k=1}^{B} w_k N(\mathbf{z}_t; \mu_k + \delta\mu_k, \sigma_k + \delta\sigma_k) \\ &= \sum_{k=1}^{B} w_k N(\mathbf{z}_t; \hat{\mu}_k, \hat{\sigma}_k) \end{aligned}$$

where $\delta\mu_k$ and $\delta\sigma_k$ are deviations from cepstral space of target noisy environment to that of the reference training environment for means and variance in the corresponding Gaussian mixtures.

In implementing DCCA, VQ encoding is performed on speech from the CLSTLK microphone processed with CMN. The output VQ labels are shared by the CLSTLK data and the corresponding data in the secondary (or target) environment. For each subspace in the CLSTLK training environment, we generate the corresponding means and variances for the target environment. Thus, a one-to-one mapping between the means and variances of the cepstral space of the CLSTLK training condition and that of the target condition is established.

Recognition is accomplished by shifting the means of the Gaussian mixtures according to the relationships

$$p_{s_t} = \sum_{k=1}^{B} w_k N(\mathbf{z}_t + \delta\mathbf{z}_t; \mu_k, \sigma_k) = \sum_{k=1}^{B} w_k N(\mathbf{z}_t; \mu_k - \delta\mathbf{z}_t, \sigma_k)$$

Baum-Welch Codebook Adaptation (BWCA). There are many applications in which stereo data simultaneously recorded in the CLSTLK and target environments are unavailable. In these circumstances, transformations can be developed between environments using the contents of the adaptation utterances using the Baum-Welch algorithm.

In Baum-Welch codebook adaptation, mean vectors and covariance matrices, along with senones, are re-estimated and updated using the Baum-Welch algorithm [5] during each iteration of training process. To compensate for the effect of changes in acoustical environments, the Baum-Welch approach is used to transform the means and covariances toward the cepstral space of the target testing environments. This is exactly like normal training, except that only a few adaptation utterances are available, and that number of free parameters to be estimated (*i.e.* the means and variances of the VQ codewords) is very small.

2.6. Reduced Bandwidth Analysis for Telephone Speech

The conventional SPHINX-II system uses signal processing that extracts Mel-frequency cepstral coefficients (MFCC) over an analysis range of 130 to 6800 Hz. This choice of analysis bandwidth is appropriate when the system processes speech recorded through good-quality microphones such as the CLSTLK microphone. Nevertheless, when speech is recorded from telephone lines, previous research at CMU [9] indicates that error rates are sharply decreased when the analysis bandwidth is reduced. This is accomplished by performing the normal DFT analysis with the normal 16,000-Hz sampling rate, but only retaining DFT coefficients after the triangular frequency smothing from center frequencies of 200 to 3700 Hz. Reduced-bandwidth MFCC coefficients are obtained by performing the discrete-cosine transform only on these frequency-weighted DFT coefficients.

To determine whether or not speech from an unknown environment is of telephone nature, we use the *Gaussian environment classifier* approach, as described in Sec. 2.2. Two VQ codebooks are used, one for telephone speech using a wideband front-end analysis and another for non-telephone speech. The speech was classified to maximize environmental likelihood.

2.7. Silence codebook adaptation

When dealing with speech-like noises such as a speech or music in the background the compensation techniques described above pro-

vide only partial recovery. Most of these techniques assume certain statistical features for the noise (such as stationarity at the sentence level), that are not valid. The SPHINX-II recognition system still produces a large number of insertion errors in difficult recognition environments, such as those used in the 1993 CSR Spoke 8 evaluation, even when cepstral compensation is used. We have found that the use of *silence codebook adaptation* (SCA) helps reduce insertion rates in these circumstances by providing better discrimination between speech and speech-like noises.

In SCA processing, the HMM parameters (codebook means, variances, and probabilities) are updated for the silence and noise segments by exposure to training data from a corpus that more closely approximates the testing environment than speech from the CLSTLK microphone. If not enough data are available, an update of the cepstral means only is performed. Further details on how this procedure was implemented for the 1993 CSR Spoke 8 evaluation are provided in Sec. 4.2

3. PERFORMANCE OF ALGORITHMS IN DEVELOPMENTAL TESTING

In this and the following section we describe the results of a series of experiments that compare the recognition accuracy of the various algorithms described in Sec. 2 using the ARPA CSR Wall Street Journal task. The 7000 WSJ0 utterances recorded using the CLSTLK microphone were used for the training corpus, and in most cases the system was tested using the 330 utterances from secondary microphones in the 1992 evaluation test set. This test set has a closed vocabulary of 5000 words.

Two implementations of the SPHINX recognition system were used in these evaluations. For most of the development work and for the official 1993 CSR evaluations for Spoke 5 and Spoke 8, a smaller and faster version of SPHINX-II was used than the implementation used for the official ATIS and CSR Hub evaluations. We refer to the faster system as SPHINX-IIa in this paper. SPHINX-IIa differs from SPHINX-II in two ways: it uses a bigram grammar (rather than a trigram grammar) and it uses only one codebook (rather than 27 phone-dependent codebooks). Spoke 5 and selected other test sets were subsequently re-evaluated using a versions of SPHINX-II that was very similar to the one used in the ATIS and CSR Hub evaluations.

3.1. Comparison of MFCDCN, IFCDCN, PDCN, and IPDCN

We first consider the relative performance of the MFCDCN, IFCDCN, PDCN, and IPDCN algorithms, which were evaluated using the training and test sets described above. Recognition accuracy using these algorithms is compared to the baseline performance, which was obtained using conventional Mel-cepstrum based signal processing in conjunction with cepstral man normalization (CMN).

Table 1 compares word error rates obtained using various processing schemes along with the corresponding reduction of word error rates with the respect to the baseline with CMN. Compensation vectors used for these comparisons were developed from training data that include the testing environments. Table 2 summarizes similar results that were obtained when the actual testing environment was excluded from the set of data used to develop the compensation vectors.

COMPENSATION ALGORITHM	CLSTLK mic	OTHER mics
CMN (baseline)	7.6	21.4
CMN+MFCDCN	7.6	14.5
CMN+IFCDCN	7.8	15.1
CMN+PDCN	7.9	16.9
CMN+MFCDCN+PDCN	7.6	12.8

Table 1: Word error rates obtained on the secondary-mic data from the 1992 WSJ evaluation test using CMN, MFCDCN, and PDCN with and without environment interpolation.

COMPENSATION ALGORITHM	CLSTLK mic	OTHER mics
CMN (baseline)	7.6	21.4
CMN+MFCDCN	7.6	16.1
CMN+MFCDCN+PDCN	7.6	14.8
CMN+IFCDCN	7.6	15.6
CMN+IFCDCN+IPDCN	7.6	13.5

Table 2: Word error rates obtained using CMN, MFCDCN, and PDCN as in Table 1, but with the testing environments excluded from the corpus used to develop compensation vectors.

The results of Table 1 indicate that PDCN when applied in isolation provides a recognition error rate that is not as good as that obtained using MFCDCN. Nevertheless, the effects of PDCN and MFCDCN are complementary in that the use of the two algorithms in combination provides a lower error rate than was observed with either algorithm applied by itself. The results in Table 2 demonstrate that the use of environment interpolation is helpful when the testing environment is not included in the set used to develop compensation vectors. Environmental interpolation degrades performance slightly, however, when the actual testing environment was observed in developing the compensation vectors.

3.2. Performance of Codebook Adaptation

Table 3 compares word error rates obtained with the DCCA and BWCA as described in Sec. 2.5 with error rates obtained with CMN and MFCDCN. The Baum-Welch codebook adaptation was implemented with four iterations of re-estimation, re-estimating the codebook means only. (Means and variances were re-estimated in a pilot experiment, but with no improvement in performance.)

COMPENSATION ALGORITHM	CLSTLK mic	OTHER mics
CMN (baseline)	7.6	21.4
CMN+MFCDCN	7.6	14.5
CMN+IFCDCN	7.8	15.1
CMN+DCCA	7.9	14.2
CMN+MFCDCN+DCCA	7.6	12.3
CMN+BWCA	7.9	16.7
CMN+MFCDCN+BWCA	7.6	13.5

Table 3: Comparison of error rates obtained using codebook adaptation with and without MFCDCN.

Table 4 provides similar comparisons, but with the testing environments excluded from the corpus used to develop compensation vectors (as in Table 2).

COMPENSATION ALGORITHM	CLSTLK mic	OTHER mics
CMN (baseline)	7.6	21.4
CMN+MFCDCN	7.6	16.1
CMN+MFCDCN+DCCA	7.6	15.8
CMN+MFCDCN+BWCA	7.6	15.5
CMN+IFCDCN	7.6	15.6
CMN+IFCDCN+DCCA	7.6	15.0
CMN+IFCDCN+BWCA	7.6	14.6

Table 4: Word error rates as in Table 3, but with the testing environments excluded from the corpus used to develop the compen-

The results of Tables 3 and 4 indicate that the effectiveness of codebook adaptation used in isolation to reduce error rate is about equal to that of MFCDCN. Once again, the use of environmental interpolation is helpful in cases in which the testing environment was not used to develop compensation vectors.

3.3. Reduced-bandwidth Processing for Telephone Speech

Table 5 compares error rates obtained using conventional signal processing versus reduced-bandwidth analysis for the telephone-microphone subset of the development set, and for the remaining microphones.

PROCESSING BANDWIDTH	NON-TELEPHONE mics	TELEPHONE mics
Full bandwidth	11.2	39.0
Reduced bandwidth	22.3	16.5

Table 5: Word error rates obtained using conventional and reduced-bandwidth analysis for the 1992 WSJ test set.

It can be seen in Table 5 that the use of a reduced analysis bandwidth dramatically improves recognition error rate when the system is trained with high-quality speech and tested using telephone-bandwidth speech.

In an unofficial evaluation we applied reduced-bandwidth analysis to the telephone speech data of Spoke 6 (known-microphone adaptation) from the 1993 ARPA CSR evaluation. Using a version of SPHINX-II trained with only the 7000 sentences of the WSJ0 corpus, we observed an error rate for the test set of 13.4%. This results compares favorably with results reported by other sites using systems that were trained with both the WSJ0 and WSJ1 corpora.

4. PERFORMANCE USING THE 1993 CSR EVALUATION DATA

We summarize in this section the results of experiments using the 1993 ARPA CSR WSJ test set, including official results obtained using SPHINX-IIa and subsequent evaluations with SPHINX-II.

4.1. Spoke 5: Microphone Independence

The testing data for Spoke 5 were samples of speech from 10 microphones that had never been used previously in ARPA evaluations. One of the microphones is a telephone handset and another is a speakerphone.

The evaluation system for Spoke 5 first performs a crude environmental classification using the Gaussian environment classifier, blindly separating incoming speech into utterances that are assumed to be recorded either from wideband microphones or telephone-bandwidth microphones and channels. Speech that is assumed to be recorded from a full-bandwidth microphone is processed using a combination of IFCDCN and IPDCN, interpolating over the closest three environments for IFCDCN and over the best four environments for IPDCN. Speech that is believed to be recorded through a telephone channel is processed using a combination of narrow-band processing as described in Sec. 2.6 and MFCDCN. The systems were trained using the CLSTLK microphone.

CONDITION	TESTING MIC	COMPENSATION	SPHX-IIa 11/93	SPHX-II 12/93
P0	Other	ON	15.6	10.4
C1	Other	OFF	21.3	16.8
C2	CLSTLK	ON	10.0	6.6
C3	CLSTLK	OFF	10.1	6.5

Table 6: Comparison of word error rates for Spoke 5 using two different recognition systems, SPHINX-IIa and SPHINX-II.

Recognition error rates on the 5000-word S5 task are summarized in Table 6. The results for SPHINX-IIa are the official evaluation results. The test data were re-run in 12/93 using a version of SPHINX-II that was very similar to that used for the Hub evaluation. Although this evaluation was "unofficial", it was performed without any further algorithm development or exposure to the test set after the official evaluation. We note that the baseline system (without "compensation") already includes CMN.

We believe that one of the most meaningful figures of merit for environmental compensation is the ratio of errors for the P0 and C2 conditions (*i.e.* the ratio of errors obtained with CLSTLK speech and speech in the target environments with compensation enabled). For this test set, switching from speech from the CLSTLK microphone to speech from the secondary microphones causes the error rates to increase by a factor of 1.3 for the 8 non-telephone environments, by a factor of 2.4 for the 2 telephone environments, and by a factor of 1.5 for the complete set of testing data. In fact, in 3 of the 10 secondary environments, the compensated error rate obtained using the secondary miss was within 25 percent of the CLSTLK error rate. Interestingly enough, the ratio of errors for the P0 and C2 conditions is unaffected by whether SPHINX-II or SPHINX-IIa was used for recognition, confirming that in these conditions, the amount of error reduction provided by

environmental compensation does not depend on how powerful a recognition system is used.

4.2. Spoke 8: Calibrated Noise Sources

Spoke 8 considers the performance of speech recognition systems in the presence of background interference consisting of speech from AM-radio talk shows or various types of music at 3 different SNRs, 0, 10 and 20 dB. The speech is simultaneously collected using two microphones, the CLSTLK microphone and a desktop Audio-Technica microphone with known acoustical characteristics. The 0-dB and 10-dB conditions are more difficult than the acoustical environment of Spoke 5, because the background signal for both the AM-radio and music conditions is frequently speech-like, and because it is highly non-stationary. SNRs are measured at the input to the Audio-Technica microphone

The evaluation system used a combination of two algorithms for environmental robustness, MFCDCN, and silence codebook adaptation (SCA). New silence codebooks were created using an implementation of Baum-Welch codebook adaptation, as described in Sec. 2.5. Two cepstral codebooks were developed for SPHINX-IIa, with one codebook representing the noise and silence HMMs, and the other codebook representing the other phones. The normal Baum-Welch re-estimation formulas were used updating only the means for the noise and silence HMMs.

NOISE TYPE	SNR (dB)	CONDITIONS (% error rate)			
		P0	C1	C2	C3
Music	0	58.7	77.9	17.7	16.9
	10	19.5	32.2	14.8	12.0
	20	14.8	15.0	14.3	11.5
AM radio	0	75.5	86.7	29.1	25.9
	10	25.5	36.9	15.4	12.1
	20	15.8	16.6	15.4	12.4

Table 7: Word error rates for Spoke 8 evaluation using SPHINX-IIa. The evaluation system included MFCDCN and SCA. The Audio-Technica is used as the secondary microphone.

The reduced-performance SPHINX-IIa system was used as in Spoke 5, except that two cepstral codebooks were needed to implement silence codebook adaptation, one to model the noise and silence segments, and the other to model the remaining phonemes. Four codebooks were developed to model silence segments for different combinations of SNR and background noise, and the codebook used to provide silence compensation was chosen blindly on the basis of minimizing residual VQ distortion. System parameters were chosen that optimize performance for the 10-dB SNR. Results for Spoke 8 are summarized in Table 7. By comparing the C2 and C3 results using the CLSTLK microphone with the P0 and S1 results using the Audio-Technica mic, we note that very little degradation is observed in the 20-dB condition but that recognition accuracy is quite low for the 0-dB condition, even when the signal is compensated. The use of MFCDCN and SCA improves recognition accuracy by 35.0 percent overall, and the ratio of overall error rates for the C2 and P0 conditions is 1.49, as in Spoke 5. As expected, the AM-radio interference was more difficult to cope with than the musical interference at all SNRs, presumably because it is more speech-like.

5. SUMMARY AND CONCLUSIONS

In this paper we describe a number of procedures that improve the recognition accuracy of the SPHINX-II system in unknown acoustical environments. We found that the use of MFCDFN and phone-dependent cepstral normalization reduces the error rate by 40 percent compared to that obtained with CMN alone. The use of Baum-Welch codebook adaptation with MFCDCN reduces the error rate by 37 percent compared to that obtained with CMN alone. The use of reduced-frequency processing reduces error rates for telephone-bandwidth speech by 58 percent compared to the rate observed for conventional signal processing. The performance of these systems for the 1993 CSR Spoke 5 and Spoke 8 evaluations is described.

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. We thank Raj Reddy, Mei-Yuh Hwang, and the rest of the speech group for their contributions to this work, and Yoshiaki Ohshima in particular for seminal discussions on reduced-bandwidth frequency analysis.

REFERENCES

1. Juang, B.-H., "Speech Recognition in Adverse Environments", *Computer Speech and Language*, 5:275-294, 1991.
2. Acero, A., *Acoustical and Environmental Robustness in Automatic SPeech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.
3. Liu, F.H., Acero, A., and Stern, R.M., "Effective Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering", *ICASSP-92*, pp. 865-868, March 1992.
4. Hwang, M.Y., *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition*, Ph.D. Thesis, Carnegie Mellon University, 1993.
5. Huang, X.D., Ariki, Y., and Jack, M., *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, U.K., 1990.
6. Liu, F.H., Stern, R.M., Huang, X.D., and Acero A., "Efficient Cepstral Normalization for Robust Speech Recognition," *Proceedings of ARPA Speech and Natural Language Workshop*, pp. 69-74, Princeton, March 1993.
7. Schwartz, R., Anastasakos, T., Kubala, F., Makhoul, J., Nguyen, L., and Zavaliagkos, G., "Comparative Experiments on Large Vocabulary Speech Recognition", *Proc. ARPA Human Language Technology Workshop*, March, 1993.
8. Huang, X., Alleva, F., Hon, H., Hwang, M., Lee, K., and Rosenfeld, R., "The SPHINX-II Speech Recognition System: An Overview", *Computer Speech and Language*, 2:137-148, 1993.
9. Moreno, P. J., and Stern, R. M., "Sources of Degradation of Speech Recognition in the Telephone Network", *ICASSP-94*, April, 1994.