# EFFICIENT CEPSTRAL NORMALIZATION FOR ROBUST SPEECH RECOGNITION

*Fu-Hua Liu, Richard M. Stern, Xuedong Huang, Alejandro Acero*

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

## ABSTRACT

In this paper we describe and compare the performance of a series of cepstrum-based procedures that enable the CMU SPHINX-II speech recognition system to maintain a high level of recognition accuracy over a wide variety of acoustical environments. We describe the MFCDCN algorithm, an environment-independent extension of the efficient SDCN and FCDCN algorithms developed previously. We compare the performance of these algorithms with the very simple RASTA and cepstral mean normalization procedures, describing the performance of these algorithms in the context of the 1992 DARPA CSR evaluation using secondary microphones, and in the DARPA stress-test evaluation.

## 1. INTRODUCTION

The need for speech recognition systems and spoken language systems to be robust with respect to their acoustical environment has become more widely appreciated in recent years (*e.g.* [1]). Results of many studies have demonstrated that even automatic speech recognition systems that are designed to be speaker independent can perform very poorly when they are tested using a different type of microphone or acoustical environment from the one with which they were trained (*e.g.* [2,3]), even in a relatively quiet office environment. Applications such as speech recognition over telephones, in automobiles, on a factory floor, or outdoors demand an even greater degree of environmental robustness.

Many approaches have been considered in the development of robust speech recognition systems including techniques based on autoregressive analysis, the use of special distortion measures, the use of auditory models, and the use of microphone arrays, among many other approaches (as reviewed in [1,4]).

In this paper we describe and compare the performance of a series of cepstrum-based procedures that enable the CMU SPHINX-II speech recognition system to maintain a high level of recognition accuracy over a wide variety of acoustical environments. The most recently-developed algorithm is *multiple fixed codeword-dependent cepstral normalization* (MFCDCN). MFCDCN is an extension of a similar algorithm, FCDCN, which provides an additive environmental compensation to cepstral vectors, but in an environment-specific fashion [5]. MFCDCN is less computationally complex than the earlier CDCN algorithm, and more accurate than the related SDCN and BSDCN algorithms [6], and it does not require domain-specific training to new acoustical environments. In this paper we describe the performance of MFCDCN and related algorithms, and we compare it to the popular RASTA approach to robustness.

## 2. EFFICIENT CEPSTRUM-BASED COMPENSATION TECHNIQUES

In this section we describe several of the cepstral normalization techniques we have developed to compensate simultaneously for additive noise and linear filtering. Most of these algorithms are completely data-driven, as the compensation parameters are determined by comparisons between the testing environment and simultaneously-recorded speech samples using the DARPA standard close-etalking Sennheiser HMD-414 microphone (referred to as the CLSTLK microphone in this paper). The remaining algorithm, *codeword-dependent cepstral normalization* (CDCN), is model-based, as the speech that is input to the recognition system is characterized as speech from the CLSTLK microphone that undergoes unknown linear filtering and corruption by unknown additive noise.

In addition, we discuss two other procedures, the RASTA method, and cepstral mean normalization, that may be referred to as cepstral-filtering techniques. These procedures do not provide as much improvement as CDCN, MFCDCN and related algorithms, but they can be implemented with virtually no computational cost.

### 2.1. Cepstral Normalization Techniques

**SDCN**. The simplest compensation algorithm, *SNR-Dependent Cepstral Normalization* (SDCN) [2,4], applies an additive correction in the cepstral domain that depends exclusively on the instantaneous SNR of the signal. This correction vector equals the average difference in cepstra

between simultaneous "stereo" recordings of speech samples from both the training and testing environments at each SNR of speech in the testing environment. At high SNRs, this correction vector primarily compensates for differences in spectral tilt between the training and testing environments (in a manner similar to the blind deconvolution procedure first proposed by Stockham *et al.* [7]), while at low SNRs the vector provides a form of noise subtraction (in a manner similar to the spectral subtraction algorithm first proposed by Boll [8]). The SDCN algorithm is simple and effective, but it requires environment-specific training.

**FCDCN.** *Fixed codeword-dependent cepstral normalization* (FCDCN) [4,6] was developed to provide a form of compensation that provides greater recognition accuracy than SDCN but in a more computationally-efficient fashion than the CDCN algorithm which is summarized below.

The FCDCN algorithm applies an additive correction that depends on the instantaneous SNR of the input (like SDCN), but that can also vary from codeword to codeword (like CDCN)

$$\hat{\boldsymbol{x}} = \boldsymbol{z} + \boldsymbol{r}[k, l]$$

where for each frame $\hat{\boldsymbol{x}}$ represents the estimated cepstral vector of the compensated speech, $\mathbf{z}$ is the cepstral vector of the incoming speech in the target environment, $k$ is an index identifying the VQ codeword, $l$ is an index identifying the SNR, and $\boldsymbol{r}[k, l]$ is the correction vector.

The selection of the appropriate codeword is done at the VQ stage, so that the label $k$ is chosen to minimize

$$\| \boldsymbol{z} + \boldsymbol{r}[k, l] - \boldsymbol{c}[k] \|^2$$

where the $\boldsymbol{c}[k]$ are the VQ codewords of the speech in the training database. The new correction vectors are estimated with an EM algorithm that maximizes the likelihood of the data.

The probability density function of $\mathbf{x}$ is assumed to be a mixture of Gaussian densities as in [2,4].

$$p(\boldsymbol{x}) = \sum_{k=0}^{K-1} P[k]\ (N_{\boldsymbol{x}} c[k], \Sigma_k)$$

The cepstra of the corrupted speech are modeled as Gaussian random vectors, whose variance depends also on the instantaneous SNR, $l$, of the input.

$$p(\boldsymbol{z} | k, \boldsymbol{r}, l) = \frac{C'}{\sigma[l]} \exp\left( -\frac{1}{2\sigma^2} \| \boldsymbol{z} + \boldsymbol{r}[k, l] - c[k] \|^2 \right)$$

In [4] it is shown that the solution to the EM algorithm is the following iterative algorithm. In practice, convergence is reached after 2 or 3 iterations if we choose the initial values of the correction vectors to be the ones specified by the SDCN algorithm.

1. Assume initial values for $\boldsymbol{r}'[k, l]$ and $\sigma^2[l]$.

2. **Estimate** $f[k]$, the *a posteriori* probabilities of the mixture components given the correction vectors $r'[k, l_i]$, variances $\sigma^2[l_i]$, and codebook vectors $\boldsymbol{c}[k]$

$$f_i[k] = \frac{\exp\left( -\frac{1}{2\sigma^2[l_i]} \| \boldsymbol{z}_i + \boldsymbol{r}'[k, l] - \boldsymbol{c}[k] \|^2 \right)}{\sum_{p=0}^{K-1} \exp\left( -\frac{1}{2\sigma^2[l_i]} \| \boldsymbol{z}_i + \boldsymbol{r}'[p, l_i] - \boldsymbol{c}[p] \|^2 \right)}$$

where $l_i$ is the instantaneous SNR of the $i^{th}$ frame.

3. **Maximize** the likelihood of the complete data by obtaining new estimates for the correction vectors $\boldsymbol{r}'[k, l]$ and corresponding $\sigma[l]$:

$$\boldsymbol{r}[k, l] = \frac{\sum_{i=0}^{N-1} (\boldsymbol{x}_i - \boldsymbol{z}_i) f_i[k]\, \delta[l - l_i]}{\sum_{i=0}^{N-1} f_i[k]\, \delta[l - l_i]}$$

$$\sigma^2[l] = \frac{\sum_{i=0}^{N-1} \sum_{k=0}^{K-1} \| \boldsymbol{x}_i - \boldsymbol{z}_i - \boldsymbol{r}[k, l] \|^2 f_i[k]\, \delta[l - l_i]}{\sum_{i=0}^{N-1} \sum_{k=0}^{K-1} f_i[k]\, \delta[l - l_i]}$$

4. Stop if convergence has been reached, otherwise go to Step 2.

In the current version of FCDCN the SNR is varied over a range of 30 dB in 1-dB steps, with the lowest SNR set equal to the estimated noise level. At each SNR compensation vectors are computed for each of 8 separate VQ clusters.

Figure 1 illustrates some typical compensation vectors obtained with the FCDCN algorithm, computed using the standard closetalking Sennheiser HMD-414 microphone and the unidirectional desktop PCC-160 microphone used as the target environment. The vectors are computed at the extreme SNRs of 0 and 29 dB, as well as at 5 dB. These curves are obtained by calculating the cosine transform of the cepstral compensation vectors, so they provide an estimate of the effective spectral profile of the compensation vectors. The horizontal axis represents frequency, warped nonlinearly according to the mel scale [9]. The maximum frequency corresponds to the Nyquist frequency, 8000 Hz. We note that the spectral profile of the compensation vector varies with SNR, and that especially for the intermediate SNRs the various VQ clusters require compensation vectors of different spectral shapes. The compensation curves for 0-dB SNR average to zero dB at low frequencies by design.

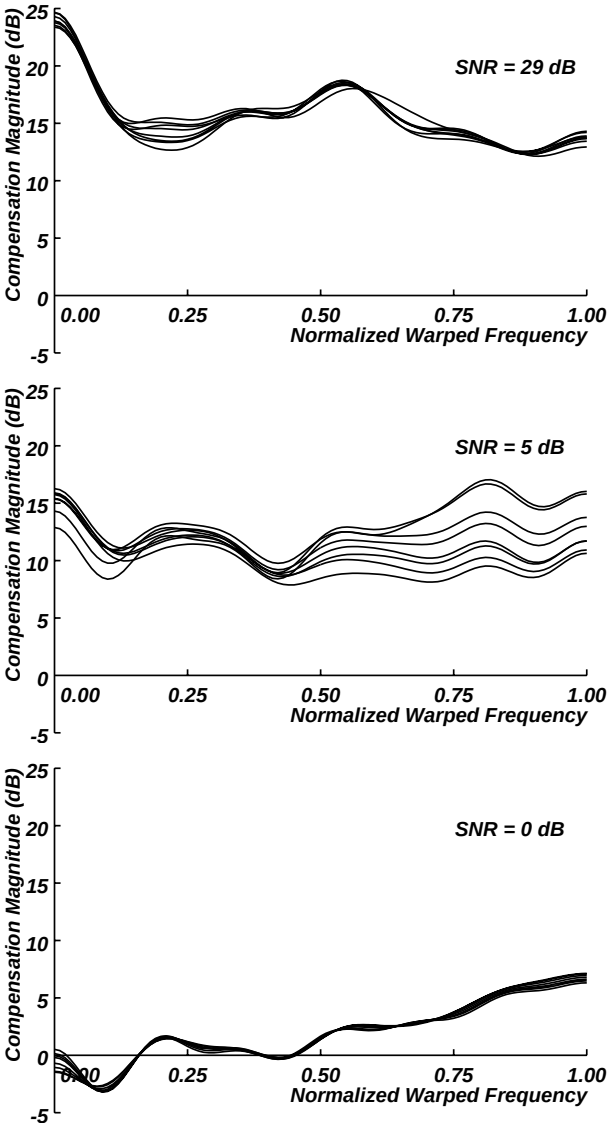


**Figure 1:** Comparison of compensation vectors using the FCDCN method with the PCC-160 unidirectional desktop microphone, at three different signal-to-noise ratios. The maximum SNR used by the FCDCN algorithm is 29 dB.

The computational complexity of the FCDCN algorithm is very low because the correction vectors are precomputed. However, FCDCN does require simultaneously-recorded data from the training and testing environments. In previous studies [6] we found that the FCDCN algorithm provided a level of recognition accuracy that exceeded what was obtained with all other algorithms, including CDCN.

**MFCDCN**. *Multiple fixed codeword-dependent cepstral normalization* (MFCDCN) is a simple extension to the FCDCN algorithm, with the goal of exploiting the simplicity and effectiveness of FCDCN but without the need for environment-specific training.

In MFCDCN, compensation vectors are precomputed in parallel for a set of target environments, using the FCDCN procedure as described above. When an utterance from an unknown environment is input to the recognition system, compensation vectors computed using each of the possible target environments are applied successively, and the environment is chosen that minimizes the average residual VQ distortion over the entire utterance,

$$|| \boldsymbol{z} + \boldsymbol{r}\,[k, l, m] - \boldsymbol{c}\,[k]\,||^2$$

where $k$ refers to the VQ codeword, $l$ to the SNR, and $m$ to the target environment used to train the ensemble of compensation vectors. This general approach is similar in spirit to that used by the BBN speech system [13], which performs a classification among six groups of secondary microphones and the CLSTLK microphone to determine which of seven sets of phonetic models should be used to process speech from unknown environments.

The success of MFCDCN depends on the availability of training data with stereo pairs of speech recorded from the training environment and from a variety of possible target environments, and on the extent to which the environments in the training data are representative of what is actually encountered in testing.

**IMFCDCN.** While environment selection for the compensation vectors of MFCDCN is generally performed on an utterance-by-utterance basis, the probability of a correct selection can be improved by allowing the classification process to make use of cepstral vectors from previous utterances in a given session as well. We refer to this type of unsupervised incremental adaptation as *Incremental Multiple Fixed Codeword-Dependent Cepstral Normalization* (IMFCDCN).

**CDCN.** One of the best known compensation algorithms developed at CMU is *Codeword-Dependent Cepstral Normalization* (CDCN) [2,4]. CDCN uses EM techniques to compute ML estimates of the parameters characterizing the contributions of additive noise and linear filtering that when applied in inverse fashion to the cepstra of an incoming utterance produce an ensemble of cepstral coefficients that best match (in the ML sense) the cepstral coefficients of the incoming speech in the testing environment to the locations of VQ codewords in the training environment.

The CDCN algorithm has the advantage that it does not require *a priori* knowledge of the testing environment (in the form of any sort of simultaneously-recorded "stereo" training data in the training and testing environments). However, it has the disadvantage of a somewhat more computationally demanding compensation process than MFCDCN and the other algorithms described above. Compared to MFCDCN and similar algorithms, CDCN uses a greater amount of structural knowledge about the nature of the degradations to the speech signal in order to improve recognition accuracy. Liu *et al.* [5] have shown that the structural knowledge embodied in the CDCN algorithm enables it to adapt to new environments much more rapidly

than an algorithm closely related to SDCN, but this experiment has not yet been repeated for FCDCN.

## 2.2. Cepstral Filtering Techniques

In this section we describe two extremely simple techniques, RASTA and cepstral mean normalization, which can achieve a considerable amount of environmental robustness at almost negligible cost.

**RASTA.** In *RASTA* filtering [10], a high-pass filter is applied to a log-spectral representation of speech such as the cepstral coefficients. The SRI DECIPHER$^{TM}$ system, for example, uses the highpass filter described by the difference equation

$$\boldsymbol{y}[n] = \boldsymbol{x}[n] - \boldsymbol{x}[n-1] + 0.97\boldsymbol{y}[n-1]$$

where $\boldsymbol{x}[n]$ and $\boldsymbol{y}[n]$ are the time-varying cepstral vectors of the utterance before and after RASTA filtering, and the index $n$ refers to the analysis frames [11].

**Cepstral mean normalization.** *Cepstral mean normalization* (CMN) is an alternate way to high-pass filter cepstral coefficients. In cepstral mean normalization the mean of the cepstral vectors is subtracted from the cepstral coefficients of that utterance on a sentence-by-sentence basis:

$$\boldsymbol{y}[n] = \boldsymbol{x}[n] - \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}[n]$$

where $N$ is the total number frames in an utterance.

Figure 2 shows the low-frequency portions of the transfer functions of the RASTA and CMN filters. Both curves exhibit a deep notch at zero frequency. The shape of the CMN curve depends on the duration of the utterance, and is plotted in Figure 2 for the average duration in the DARPA Wall Street Journal task, 7 seconds. The Nyquist frequency for the time-varying cepstral vectors is 50 frames per second.

Algorithms like RASTA and CMN compensate for the effects of unknown linear filtering because linear filters produce a static compensation vector in the cepstral domain that is the average difference between the cepstra of speech in the training and testing environments. Because the RASTA and CMN filters are highpass, they force the average values of cepstral coefficients to be zero in both the training and testing domains. Nevertheless, neither CMN nor RASTA can compensate directly for the combined effects of additive noise and linear filtering. It is seen in Figure 1 that the compensation vectors that maximize the likelihood of the data vary as a function of the SNR of individual frames of the utterance. Hence we expect compensation algorithms like MFCDCN (which incorporate this knowledge) to be more effective than RASTA or CMN (which do not).

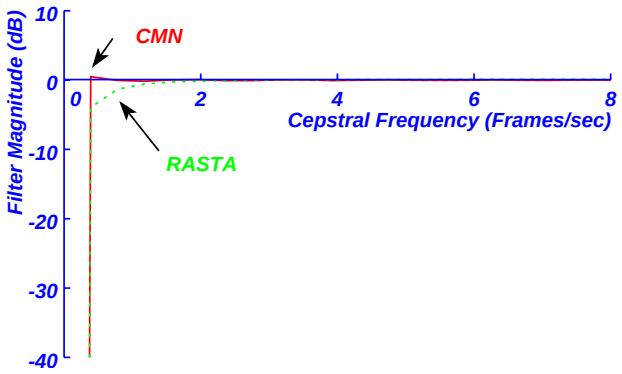


**Figure 2:** Comparison of the frequency response of the highpass cepstral filters implemented by the RASTA algorithm as used by SRI (dotted curve), and as implied by CMN (solid curve). The CMN curve assumes an utterance duration of 7 seconds.

# 3. EXPERIMENTAL RESULTS

In this section we describe the ability of the various environmental compensation algorithms to improve the recognition accuracy obtained with speech from unknown or degraded microphones.

The environmental compensation algorithms were evaluated using the SPHINX-II recognition system [12] in the context of the November, 1992, evaluations of continuous speech recognition systems using a 5000-word closed-vocabulary task consisting of dictation of sentences from the Wall Street Journal. A component of that evaluation involved utterances from a set of unknown "secondary" microphones, including desktop microphones, telephone handsets and speakerphones, stand-mounted microphones, and lapel-mounted microphones.

## 3.1. Results from November CSR Evaluations

We describe in this section results of evaluations of the MFCDCN and CDCN algorithms using speech from secondary microphones in the November, 1992, CSR evaluations.

Because of the desire to benchmark multiple algorithms under several conditions in this evaluation combined with limited resources and the severe time constraints imposed by the evaluation protocol, this evaluation was performed using a version of SPHINX-II that was slightly reduced in performance, but that could process the test data more rapidly than the system described in [12]. Specifically, the selection of phonetic models (across genders) was performed by minimizing mean VQ distortion of the cepstral vectors before recognition was attempted, rather than on the basis of *a posteriori* probability after classification. In addition, neither the unified stochastic engine (USE) described in [12] nor the cepstral mean normalization algorithms were applied. Finally, the CDCN evaluations were conducted without making use of the CART decision tree or alternate

pronunciations in the recognition dictionary. The effect of these computational shortcuts was to increase the baseline error rate for the 5000-word task from 6.9% as reported in [12] to 8.1% for the MFCDCN evaluation, and to 8.4% for the CDCN evaluation.

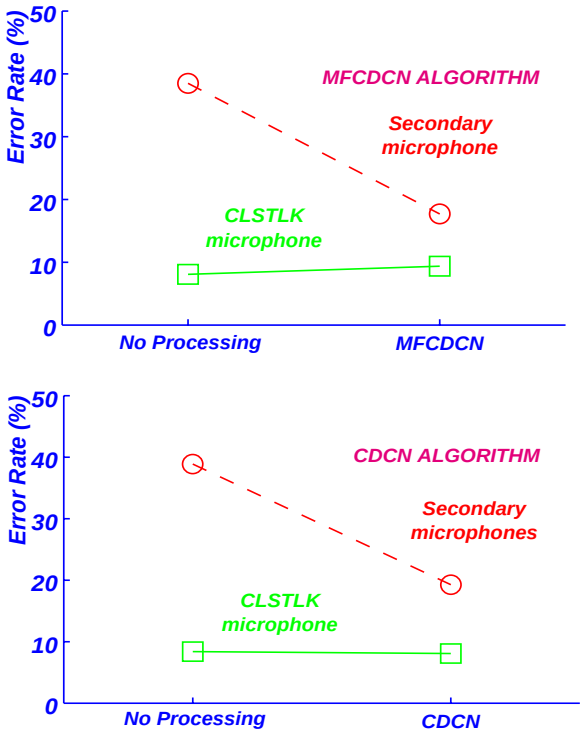


**Figure 3:** Performance of the MFCDCN algorithm (upper panel) and the CDCN algorithm (lower panel) on the official DARPA CSR evaluations of November, 1992

Figure 3 summarizes the results obtained in the official November, 1992, evaluations. For these experiments, the MFCDCN algorithm was trained using the 15 environments in the training set and developmental test set for this evaluation. It is seen that both the CDCN and MFCDCN algorithms significantly improve the recognition accuracy obtained with the secondary microphones, with little or no loss in performance when applied to speech from the close-etalking Sennheiser HMD-414 (CLSTLK) microphone. The small degradation in recognition accuracy observed for speech from the CLSTLK microphone using the MFCDCN algorithm may be at least in part a consequence of errors in selecting the environment for the compensation vectors. Environment-classification errors occurred on 48.8% of the CLSTLK utterances and on 28.5% of the utterances from secondary microphone. In the case of the secondary microphones, however, recognition accuracy was no better using the FCDCN algorithm which presumes knowledge of the correct environment, so confusions appear to have taken place primarily between acoustically-similar environments.

In a later study we repeated the evaluation using MFCDCN compensation vectors obtained using only the seven categories of microphones suggested by BBN rather than the original 15 environments. This simplification produced only a modest increase in error rate for speech from secondary microphones (from 17.7% to 18.9%) and actually improved the error rate for speech from the CLSTLK microphone (from 9.4% to 8.3%).

Figure 4 summarizes the results of a series of (unofficial) experiments run on the same data that explore the interaction between MFCDCN and the various cepstral filtering techniques. The vertical dotted line identifies the system described in [12].

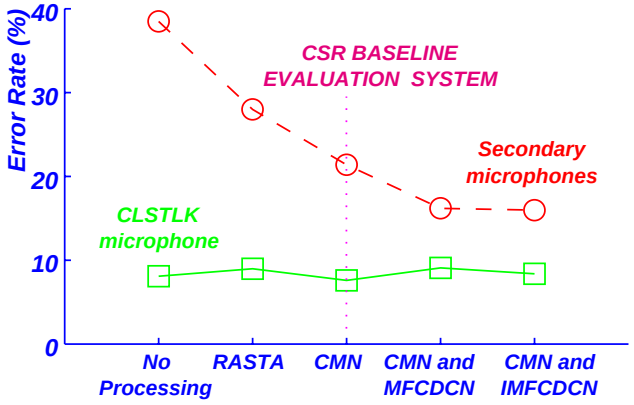


**Figure 4:** Comparison of the effects of MFCDCN, IMFCDCN, cepstral mean normalization (CMN), and the RASTA algorithm on recognition accuracy of the Sennheiser HMD-414 microphone (solid curve) and the secondary microphones (dashed curve). from the November 1992 DARPA CSR evaluation data.

It can be seen in Figure 4 that RASTA filtering provides only a modest improvement in errors using secondary microphones, and degrades speech from the CLSTLK microphone. CMN, on the other hand, provides almost as much improvement in recognition accuracy as MFCDCN, without degrading speech from the CLSTLK microphone. We do not yet know why our results using CMN are so much better than the results obtained using RASTA. In contrast, Schwartz *et al.* obtained approximately comparable results using these two procedures [13].

Finally, adding MFCDCN to CMN improves the error rate from 21.4% to 16.2%, and the use of IMFCDCN provides a further reduction in error rate to 16.0% for this task.

## 3.2. Results from the "Stress Test" Evaluation

In addition to the evaluation described above, a second unofficial "stress-test" evaluation was conducted in December, 1992, which included spontaneous speech, utterances containing out-of-vocabulary words, and speech from unknown microphones and environments, all related to the Wall Street Journal domain.

The version of SPHINX-II used for this evaluation was configured to maximize the robustness of the recognition process. It was trained on 13,000 speaker-independent utterances from the Wall Street Journal task and 14,000 utterances of spontaneous speech from the ATIS travel planning domain. The trigram grammar for the system was derived from 70.0 million words of text without verbalized punctuation and 11.6 million words with verbalized punctuation. Two parallel versions of the SPHINX-II system were run, with and without IMFCDCN. Results obtained are summarized in the Table I below.

| | In Vocab | Out of Vocab | STRESS TOTAL | BASE CSR |
|---|---|---|---|---|
| **5K CLSTLK** | 9.4% | – | **9.4%** | 5.3% |
| **5K other mic** | 13.4% | – | **13.4%** | 17.7% |
| **20K CLSTLK** | 16.8% | 22.8% | **18.1%** | 12.4% |
| **20K other mic** | 23.7% | 24.8% | **24.0%** | – |
| **Spontaneous** | 11.9% | 27.2% | **22.4%** | – |

**Table 1:** Error rates obtained by SPHINX-II in the December, 1992, "Stress-Test" Evaluation. The baseline CSR results are provided for comparison only, and were not obtained using a comparably-configured system.

We also compared these results with the performance of the baseline SPHINX-II system on the same data. The baseline system achieved a word error rate of 22.9% using only the bigram language model. Adding IMFCDCN reduced the error rate only to 22.7%, compared to 20.8% for the stress-test system using IMFCDCN. We believe that the IMFCDCN algorithm provided only a small benefit because only a small percentage of data in this test was from secondary microphones.

In general, we are very encouraged by these results, which are as good or better than the best results obtained only one year ago under highly controlled conditions. We believe that the stress-test protocol is a good paradigm for future evaluations.

## ACKNOWLEDGMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory, under Grant No. N00014-93-2005. 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 and the rest of the speech group for their contributions to this work.

## REFERENCES

1. Juang, B. H. "Speech Recognition in Adverse Environments". Comp. Speech and Lang. **5**:275-294, 1991.

2. Acero, A. and Stern, R. M. "Environmental Robustness in Automatic Speech Recognition". *ICASSP-90*, pages 849-852. April, 1990.

3. Erell, A. and Weintraub, M. Estimation "Using Log-Spectral-Distance Criterion for Noise-Robust Speech Recognition". *ICASSP-90*, pages 853-856. April, 1990.

4. Acero, A. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Kluwer Academic Publishers, Boston, MA, 1993.

5. Liu, F.-H., Acero, A., and Stern, R. M. "Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering". *ICASSP-92*, pages 865-868. March, 1992.

6. Acero, A. and Stern, R. M. "Robust Speech Recognition by Normalization of the Acoustic Space". *ICASSP-91*, pages 893-896. May, 1991.

7. Stockham, T. G., Cannon, T. M,. and Ingebretsen, R. B. "Blind Deconvolution Through Digital Signal Processing". *Proc. IEEE.* **63**:678-692, 1975.

8. Boll, S. F. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". *IEEE Trans. Acoust. Speech. and Sig. Proc.* **2:**113-120, 1979.

9. Davis, S.B, Mermelstein, P. "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoust. Speech. and Sig. Proc.* **28**:357-366, 1980.

10. Hermansky, H., Morgan, N., Bayya, A., Kohn, P. "RASTA-PLP Speech Analysis Technique". *ICASSP-92*, pages 121-124. March, 1992

11. Murveit, H., Weintraub, M. "Speaker-Independent Connected-Speech Recognition Using Hidden Markov Models". *Proc. DARPA Speech and Natural Language Workshop*. February, 1992.

12. Huang, X.,Alleva, F., Hwang. M.-Y., Rosenfeld, R. "An Overview of the SPHINX-II Speech Recognition System". *Proc. DARPA Human Language Technology Workshop*. March, 1993.

13. Schwartz, R., Anastasakos,A., Kubala, F., Makhoul, J. , Nguyen, L. , Zavaliagkos, G. "Comparative Experiments on Large Vocabulary Speech Recognition". *Proc. DARPA Human Language Technology Workshop*, March, 1993.