[26] H. Meng and V. W. Zue, "A Comparative Study of Acoustic Representations of Speech for Vowel Classification Using Multi-Layer Perceptrons," *Proc. ICSLP-90*, pp. 1053-1056, November, 1990.

[27] A. Nadas, D. Nahamoo, and M.A. Picheny, "Speech Recognition Using Noise-Adaptive Prototypes," *IEEE Trans. Acoustics, Speech and Signal Processing*, 37, pp.1495-1503, October, 1989.

[28] L. Neumeyer and M. Weintraub, "Probabilistic Optimum filtering for Robust Speech Recognition," *Proc. ICASSP-94*, pp. I-417 - I-420, April, 1994.

[29] Y. Ohshima, *Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing*, Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1993.

[30] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang and M. Allerhand,, "Complex Sounds and Auditory Images," In *Auditory Physiology and Perception*, Cazals, Y., Horner, K., and Demany, L., Eds., pp. 429-446, Pergamon Press, 1992.

[31] P. M. Peterson, it Adaptive Array Processing for Multiple Microphone Hearing Aids, RLE TR No. 541, Res. Lab. of Electronics, MIT, Cambridge, MA.

[32] A. Sankar and C.-H. Lee, "Stochastic Matching for Robust Speech Recognition," *IEEE Signal Processing Letters*, 1: 124-125, August, 1994.

[33] S. Seneff, "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing," *Journal of Phonetics*, 16: 55-76, January, 1988.

[34] T. G. Stockham, T. M. Cannon and R. B. Ingebretsen, "Blind Deconvolution Through Digital Signal Processing," *Proc. IEEE*, 63, pp. 678-692, April, 1975.

[35] T. M. Sullivan and R. M. Stern, "Multi-microphone Correlation-based Processing For Robust Speech Recognition," *ICASSP-93*, pp. II-91-II-94, April, 1993.

[36] D. Van Compernolle, "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings," *ICASSP-90*, pp. 833-836, April, 1990.

[37] A. P. Varga and R. K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," *ICASSP-90*, pp. 845-848, April, 1990.

[38] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[14] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, 87: 1738-1752, April, 1990.

[15] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, 2: 578-589, October, 1994.

[16] B. A. Hanson, T. H. Applebaum and J.-C. Junqua, "Robust Speech Recognition Under Adverse Conditions," in *Advanced Topics in Automatic Speech and Speaker Recognition*, C.-H. Lee and F. K. Soong, Kluwer Academic Publishers, Boston, 1995.

[17] M. J. Hunt and C. Lefebvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech," *Proc. ICASSP-89*, 1: 262-265, June, 1989.

[18] B.-H. Juang, "Speech Recognition in Adverse Environments," *Computer Speech and Language*, 5:275-294, July, 1991.

[19] K.-F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, 1989.

[20] H. C. Leung, B. Chigier and J. R. Glass, "A comparative study of signal representations and classification techniques for speech recognition," *Proc. ICASSP-93*, pp. 680-683, April, 1993.

[21] Q. Lin, C. Che, B. de Vries, J. Pearson and J. L. Flanagan, "Experiments on Distant-Talking Speech Recognition," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 187-192, January, 1995, Austin, TX, Morgan Kaufmann, J. R. Cohen, Ed.

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

[23] F.-H. Liu, R. M. Stern, X. Huang and A. Acero, "Efficient Cepstral Normalization For Robust Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 69-74, Princeton, NJ, Morgan Kaufmann, M. Bates, Ed., March, 1993.

[24] F.-H. Liu, R. M. Stern, A. Acero and P. Moreno, "Environment Normalization For Robust Speech Recognition Using Direct Cepstral Comparison," *Proc. ICASSP-94*, pp. II-61 - II-64, April, 1994.

[25] R. F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea," *Proc. ICASSP-82*, pp. 1282-1285, May, 1982.

REFERENCES

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

[2] A. Acero and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition," *Proc. ICASSP-90*, pp. 849-852, 1990.

[3] V. L. Beattie, *Hidden Markov Model State-Based noise compensation*, Ph.D. Thesis, Curchill College, Cambridge University, 1992.

[4] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *ICASSP-79*, pp. 208-211, April, 1979.

[5] M. Bodden, "Modeling human sound-source localization and the cocktail-party-effect," *Acta Acustica*: 1:43-55, February, 1993.

[6] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, 27, pp. 113-120, April, 1979.

[7] J. R. Cohen, "Application of an Auditory Model to Speech Recognition," *J. Acoust. Soc. Amer.*, 85: 2623-2629, June, 1989.

[8] S. B. Davis, and P. Mermelstein, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech and Signal Processing*, 28, pp.357-366, August, 1980.

[9] Y. Ephraim, "Statistical-model-based Speech Enhancement Systems," *Proc. IEEE*, 80: 1526-1555, October, 1992.

[10] J. L. Flanagan, J. D. Johnston, R. Zahn and G.W. Elko, "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms," *J. Acoust. Soc. Amer.*, 78: 1508-1518, November, 1985.

[11] J. L. Flanagan, D. A. Berkeley, G. W. Elko, J. E. West and M. M. Sondhi, "Autodirective microphone systems," *Acustica.*, 73: 58-71, February, 1991.

[12] M. J. F. Gales and S. J. Young, "Cepstral Parameter Compensation for HMM Recognition in Noise," *Speech Communication*, 12: 231-239, July, 1993.

[13] O. Ghitza, "Auditory Nerve Representation as a Front-End for Speech Recognition in a Noisy Environment," *Comp. Speech and Lang.*, 1: 109-130, December, 1986.

6 SUMMARY AND CONCLUSIONS

In this chapter we have reviewed a number of techniques that individually and collectively provide substantial reduction of speech recognition error rates in difficult acoustical environments including unknown additive noise and/or unknown linear filtering. We compared empirically-derived and structurally-based approaches to acoustical pre-processing. Empirical compensation approaches are quite easy to implement, but they require prior access to examples of simultaneously-recorded speech in the training and testing domains. Model-based compensation procedures require a valid parametric characterization of the testing environment, but they do not require access to "stereo" databases. The performance of model-based compensation procedures also converges more rapidly in new testing environments. Finally, cepstral high-pass filtering procedures provide substantial robustness at almost zero cost, and are recommended universally. We also note that the use of microphone arrays can provide a further improvement in recognition accuracy that is complementary to the benefit provided by acoustical pre-processing techniques. Finally, we also discuss several issues concerning the use of signal processing algorithm based on models of the human auditory periphery, which so far have not yet provided substantial quantitative reductions in recognition error rate.

Acknowledgements

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005, and by the Motorola Corporation. 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.

This review represents the contributions of many colleagues and friends. We thank Tom Sullivan, in particular, for the use of his experimental data in Sec. 4. We also thank Beth Cummins, Evandro Gôuvea, Pedro Moreno, and Bhiksha Raj for their help and advice in the preparation of this manuscript. Finally, we thank Raj Reddy and all of the members of the CMU speech group for their encouragement and support over the years, as well as Chin-Hui Lee for his advice and encouragement in completing this chapter.

ies. The results in the lower panel of Fig. 7 demonstrate that the mean rate and GSD outputs of the Seneff model provide lower error rates than conventional LPC cepstra when the system is trained using the CLSTLK microphone and tested using the PZM6FS microphone. This indicates that the Seneff model provides additional robustness in cases where clean speech is corrupted by linear filtering as well as additive noise, as previously noted by Hunt and Lefebvre [17] and Meng [26]. Nevertheless, use of conventional LPC-derived cepstral features combined with the CDCN algorithm (indicated by the dashed curves with the circular symbols) produced error rates that were equal to or better than the results ahieved using either output of the Seneff model for these data. We have also explored several ways of combining auditory models with compensation algorithms like CDCN, and we have failed to identify any combination of processing schemes in which the use of auditory processing provides any additional improvement in recognition accuracy beyond the accuracy obtained using conventional cepstral processing with appropriate environmental compensation [29].

While these observations are disappointing, there are several possible reasons why the benefit obtained by auditory modelling to date has been limited. For one thing, the Hidden Markov models used in SPHINX implicitly assume that the incoming features can be characterized by multivariate Gaussian pdfs. This is reasonably true for cepstral features, but far less so for the outputs of the auditory models. Indeed, Leung *et al.* [20] showed that significantly better phoneme classification accuracy can be obtained using a neural network-based classifier compared to the accuracy obtained using an HMM, presumably because the neural-net classifier makes no assumptions concerning the form of the pdfs of the features. In addition, most experiments that evaluate the recognition accuracy obtained using physiologically-motivated front ends (including the one summarized by the data in Fig. 7) simply convert the auditory model outputs into a spectrum-like display, similar to the information provided by cepstral coefficients. In reality there is a great deal more information available provided by the auditory models (especially when detailed timing information is taken into account), and it is quite possible that better recognition accuracy can be obtained using other aspects of the outputs of the auditory models. Nevertheless, the comparisons of Fig. 7 do underscore the need to evaluate auditory models in terms of the extent to which they provide improvement in recognition accuracy beyond the accuracy that now can be obtained by the best possible conventional environmental compensation procedures, rather than by the improvement that auditory models provide relative to baseline processing. The approach of auditory modeling continues to merit further attention, particularly with the goal of resolving these issues.

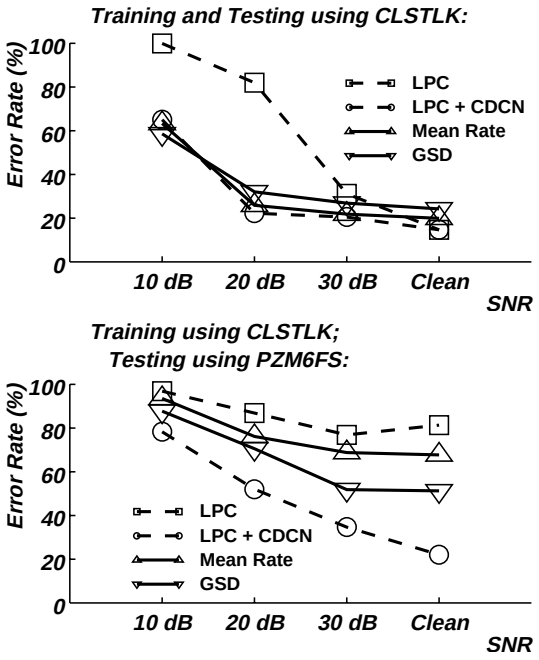

Figure 7 Comparison of error rates obtained on the Census task using conventional LPC-derived cepstral features, with and without CDCN, with results obtained using the mean rate and synchrony outputs of the Seneff auditory model. The SPHINX-I recognition system was trained using the CLSTLK microphone, and testing using either the CLSTLK microphone (upper panel) or the Crown PZM6FS microphone (lower panel).

trained using the CLSTLK microphone in all cases, and tested using either the CLSTLK microphone (upper panel) or the Crown PZM6FS microphone (lower panel). White noise was artificially added to the speech signals, and data are plotted as a function of global SNR [13].

When the SPHINX-I system is trained and tested using the CLSTLK microphone, best performance is obtained using conventional LPC-based signal processing for "clean" speech, as seen in the upper panel of Fig. 7. As the SNR decreases, however, error rates obtained using either the mean rate or GSD outputs of the Seneff model increase more gradually than error rates obtained with baseline LPC processing, confirming similar findings from previous stud-

of data acquisition and the need to be able to process much greater amounts of data. We expect that the use of microphone arrays will become much more widespread as cheaper and faster signal processing platforms become available. Nevertheless, the development of efficient multiple-microphone algorithms that are able to improve speech recognition accuracy in reverberant acoustical environments remains a significant unsolved technical challenge.

5 PHYSIOLOGICALLY-MOTIVATED SIGNAL PROCESSING

Another significant trend in robust speech recognition has been an increased interest in the use of peripheral signal processing schemes that are motivated by human auditory physiology and perception (*e.g.* [7, 13, 14, 25, 30, 33]). Recent evaluations indicate that with "clean" speech, such approaches tend to provide recognition accuracy that is comparable to that obtained with conventional LPC-based or DFT-based signal processing schemes. When the quality of the incoming speech (or the extent to which it resembles the speech used in training the system) decreases, these auditory models can provide greater robustness with respect to environmental changes [17, 26]. Despite the apparent utility of such processing schemes, no one has a deep-level understanding of why they work as well as they do, and in fact different researchers choose to emphasize rather different aspects of the peripheral auditory system's response to sound in their work. Most auditory models include a set of linear bandpass filters with bandwidth that increases nonlinearly with center frequency, a nonlinear rectification stage that frequently includes short-term adaptation and lateral suppression, and, in some cases, a more central display based on short-term temporal information. We have estimated that the number of arithmetic operations of some of the currently-popular auditory models ranges from 35 to 600 times the number of operations required for conventional LPC-based processing [29].

Figure 7 compares error rates for SPHINX-I on the Census task using both conventional LPC-derived cepstra (without CMN), with and without CDCN, and the mean rate and synchrony outputs of the Seneff auditory model [33]. The LPC results were obtained using the standard 12 LPC-based cepstral coefficients (and their derivatives) that are normally input to the SPHINX-I system, and the auditory model results were obtained using an implementation of the 40-channel mean-rate output of the Seneff model, and with the 40-channel outputs of Seneff's Generalized Synchrony Detectors (GSDs) [33]. The system was

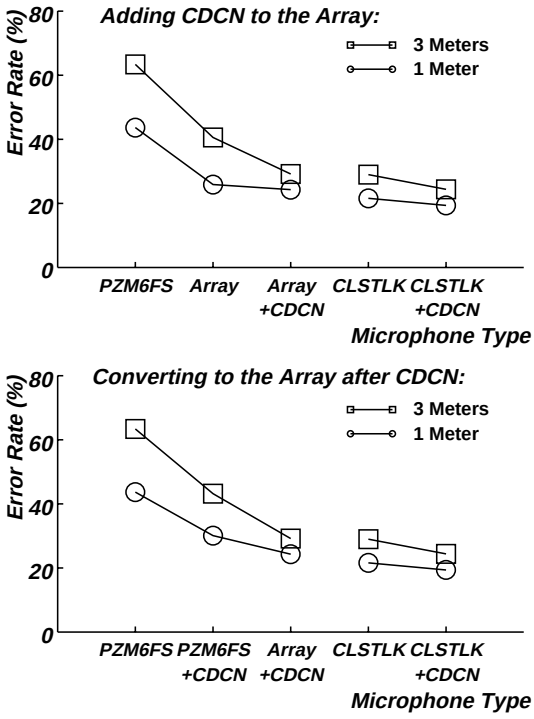

Figure 6 Comparison of recognition accuracy obtained on a portion of the Census task using the omnidirectional Crown PZM6FS, the 23-microphone array developed by Flanagan, and the CLSTLK microphone, each with and without CDCN.

These results have been replicated by Lin and his colleagues [21], who demonstrated good recognition accuracy using a combination of the Flanagan array and a neural network postprocessing stage. The neural network in these experiments performed the necessary function of compensation for spectral coloration introduced by the microphone-array processing algorithms, as did the CDCN algorithm in our experiments.

In summary, the use of arrays of microphones has considerable potential in providing an additional complementary degree of robustness to a systems that are already using acoustical pre-processing. We speculate that the major impediments to the adoption of microphone arrays up to now has been the high cost

of degradation. Consequently, these algorithms can provide good improvement in SNR when signal degradations are caused by additive independent noise sources, but they do not perform well in reverberant environments when the distortion is at least in part a delayed version of the desired speech signal (*e.g.* [31]). (This problem can be avoided by only adapting during non-speech segments [36]). A third type of approach to microphone array processing is the use of cross-correlation-based algorithms (*e.g.* [5, 35]). These algorithms are appealing because they are based on human binaural hearing, and because cross-correlation is an efficient way to identify and isolate the direction of a strong signal source. We believe that signal processing techniques based on human binaural perception are worth pursuing, but their effectiveness for automatic speech recognition remains to be conclusively demonstrated.

Figure 6 describes results obtained from a pilot evaluation of the microphone array developed by Flanagan and his colleagues at AT&T Bell Laboratories. The Flanagan array [10] is a one-dimensional delay-and-sum beamformer which uses 23 unevenly-spaced microphones. We compared the recognition accuracy for the alphanumeric Census task obtained using the Flanagan array to the accuracy observed using the CLSTLK and PZM6FS microphones. The utterances were recorded in a sparsely-furnished laboratory at the Rutgers CAIP Center with an estimated reverberation time between 500 and 750 ms. Simultaneous recordings were made of each utterance using three microphones: the CLSTLK microphone, the Crown PZM6FS, and the Flanagan array with the input low-pass-filtered at 8 kHz. Recordings were made with the speaker seated at distances of 1, 2, and 3 meters from the PZM6FS and Flanagan array microphones, while wearing the CLSTLK microphone in the usual fashion at all times.

When the Flanagan array is used in conjunction with the CDCN algorithm, the resulting error rates are very close to the error rates obtained with the CLSTLK microphone without CDCN. In other words, the combination of microphone arrays and acoustical pre-processing can completely close the "gap" in performance noted at the end of the first subsection of Sec. 3.4 between results obtained testing using the desktop PZM6FS microphone and results obtained using the CLSTLK microphone. It is also interesting to note that the improvements provided by Flanagan array and the CDCN algorithm are complementary: if one is already using the Flanagan array, the error rate can be decreased by adding CDCN (upper panel of Fig. 6), and if one is already using CDCN, the error rate can be reduced by replacing the PZM6FS microphone by the Flanagan array (lower panel of Fig. 6).

ing environment. Figure 5 compares recognition accuracy as a function of the amount of environment-specific speech data available for adaptation using two compensation algorithms. The first algorithm is an empirical compensation algorithm known as BSDCN [22], which is closely related to the SDCN algorithm described in Sec. 3.1. The second algorithm is the model-based CDCN procedure described in Sec. 3.2. Recognition accuracy using the real-time CDCN algorithm converges with only about 2 seconds of adapting speech, while the BSDCN algorithm requires at least 60 seconds of adapting speech to reach asymptotic levels of recognition accuracy. This is consistent with intuition, as compensation procedures based on a structural model of degradation (such as CDCN) are based on the estimation of only a small number of model parameters. Empirical compensation algorithms such as the BSDCN algorithm, on the other hand, must learn all relevant aspects of the testing environment by observation. We believe that convergence time (or conversely, the amount of testing data needed for convergence) represents another facet of the tradeoff between empirical and structural approaches to adaptation: the empirical approaches can be applied to a wider variety of environments, but they require much more data to be effective.

4 MULTIPLE MICROPHONE ARRAYS

Further improvements in recognition accuracy can be obtained in difficult environments by combining acoustical pre-processing with arrays of multiple microphones. The use of microphone arrays is motivated by a desire to improve the effective SNR of speech as it is input to the recognition system. Close-talking microphones, for example, produce higher SNRs than desktop microphones under normal circumstances because they pick up a relatively small amount of additive noise, and because the incoming signal is not degraded by reverberated components of the original speech.

Several different types of array-processing strategies have been applied to automatic speech recognition. The simplest approach is that of the delay-and-sum beamformer, in which delays are inserted in each channel to compensate for differences in travel time between the desired sound source and the various sensors (*e.g.* [10, 11]). A second option is to use an adaptation algorithm based on minimizing mean square energy such as the Frost or Griffiths-Jim algorithm [38]. These algorithms provide the opportunity to develop nulls in the direction of noise sources as well as more sharply focused beam patterns, but they assume that the desired signal is statistically independent of all sources

Use of Physical Parameters versus Presumed Phonemic Identity for Empirical Compensation

We have also measured the recognition accuracy obtained by adding either MPDCN or MFCDCN to a recognition system that already makes use of CMN when secondary microphones are used for testing [24]. In our comparisons using the 1993 WSJ1 database, a combination of MFCDCN and MPDCN with CMN reduces error rates compared to the use of CMN alone 40.2 percent for all of the 10 secondary microphones in the evaluation set. The effects of MPDCN and MFCDCN were somewhat complementary in that the addition of MPDCN to a recognition system that already includes CMN and MFCDCN provides a further decrease of recognition error rate by 11.7 percent. We conjecture that MPDCN complements the effects of MFCDCN because the MPDCN compensation vectors are based on a partitioning of the incoming cepstral vectors that is somewhat different from the partition that is obtained using MFCDCN.

Convergence Times for Model-Based Compensation versus Empirical Compensation

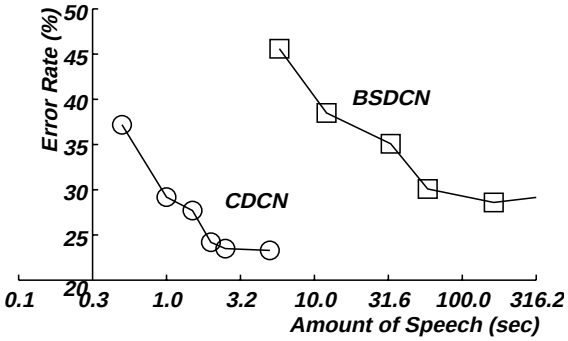

Figure 5 Comparison of the recognition accuracy of the empirical BSDCN algorithm and the model-based CDCN algorithm. Results are shown as a function of the amount of speech in the testing environment available for adaptation.

An important figure of merit for environmental compensation algorithms is the amount of data in the testing environment needed for the algorithm to converge. For example, speech recognition over switched telephone networks must be performed on the basis of only a very small amount of speech in the test-

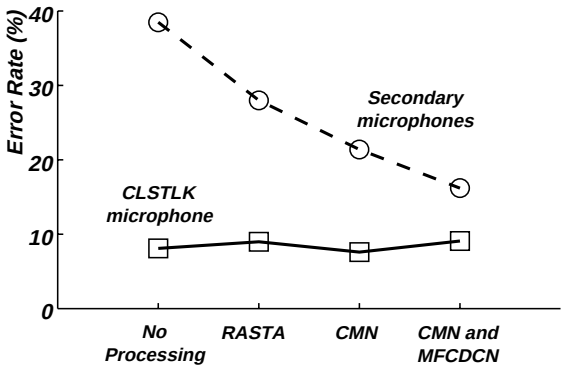

Figure 4 Comparison of the effects of MFCDCN, cepstral mean normalization (CMN), and the RASTA algorithm on recognition accuracy of the Sennheiser HMD-414 microphone (solid curve) and the secondary microphones (dashed curve). Results were obtained by training on the CLSTLK microphone and testing on the 1992 ARPA WSJ0 CSR evaluation data.

while the dashed curve describes results obtained training using the CLSTLK microphone and testing using an ensemble of non-closetalking and telephone-bandwidth "secondary" microphones. Results are obtained using two types of cepstral high-pass filtering algorithm described in Sec. 3.3 along with the MFCDCN algorithm described in Sec. 3.1. Use of the RASTA and CMN algorithms reduced the error rates observed while using the secondary microphones by 27.2 percent and 44.4 percent, respectively, with respect to the baseline error rates. Adding MFCDCN processing to CMN provided an additional decrease of 24.3 percent in error rate. We believe that the MFCDCN algorithm provides greater recognition accuracy than cepstral high-pass filtering because it incorporates an ensemble of different cepstral compensation vectors that depend on the specific SNR and VQ codeword identity of each incoming cepstral vector. This is equivalent to using all of the compensation vectors represented by the curves of Fig. 3 rather than just a single time-averaged vector.

compensation for the effects of noise and filtering. We note that if the desktop microphone is used for testing, the system performs equally well regardless of which microphone had been used for training. Hence, the degradation in performance due to mismatches between training and testing conditions is eliminated by use of the CDCN algorithm (at least for this pair of environments and this test set). The remaining difference between the compensated error rate with the PZM6FS microphone (25.1 percent) and the error rate obtained training and testing with the CLSTLK microphone (14.7 percent) arises because speech recorded using the PZM6FS microphone has a lower SNR. We will revisit this issue in our discussion of microphone arrays in Sec. 4 below.

Empirical versus Model-Based Compensation

Comparisons of recognition accuracy obtained using the empirical compensation procedure MFCDCN described in Sec. 3.1 with the model-based compensation procedure CDCN described in Sec. 3.2 are included in [23]. These comparisons were obtained using SPHINX-II and Version 0 of the 5000-word 1992 Wall Street Journal evaluation set (WSJ0). In this study the error rates observed using the empirical MFCDCN algorithm were approximately the same as those obtained using the model-based CDCN algorithm. This is not true in general: empirical compensation generally works well when the environments used to "train" the compensation procedure are similar to those used in evaluating the system, while model-based compensation procedures work well when the structural model that is assumed is actually representative of the data in the testing set. In the case of the WSJ0 task, the environments used in training and testing are quite acoustically similar, and the structural model assumed by CDCN (shown in Fig. 2) is indeed valid. CDCN would have outperformed MFCDCN if there were greater differences between training and testing conditions, while MFCDCN would have worked substantially better than CDCN for testing conditions (such as nonlinear distortion) for which the simple model of degradation shown in Fig. 2 is invalid.

Empirical Compensation versus Cepstral high-pass Filtering

Figure 4 depicts results from the ARPA 1992 WSJ0 database that compare recognition accuracy obtained using direct cepstral comparison with that obtained using two types of cepstral high-pass filtering [the RASTA algorithm and cepstral mean normalization (CMN)] [23]. The solid curve describes recognition accuracy obtained by training and testing using the CLSTLK microphone,

of speech in the training and testing environments. The high-pass nature of both the RASTA and CMN filters forces the average values of cepstral coefficients to be zero in the training and testing environments individually, which, of course, implies that the average cepstra in the two environments are equal to each other.

Cepstral high-pass filtering can also be thought of as a degenerate case of compensation based on direct cepstral comparison. Consider, for example, the compensation vectors with frequency response depicted in Fig. 3. Cepstral high pass filtering produces the same effect that would have been achieved if all of the compensation vectors for a particular testing environment are combined into a *single* compensation vector, weighted in proportion to the percentage of frames having the set of physical parameters (or presumed phoneme identity) corresponding to each of the original compensation vectors. As Fig. 3 indicates, actual cepstral compensation vectors depend on the SNR, VQ codeword location, and/or phonemic identity of the individual frames of the testing utterances. Hence neither CMN nor RASTA can compensate directly for all of the combined effects of additive noise and linear filtering.

In general, cepstral high-pass filtering is so cheap and effective that it is currently embedded in some form in virtually all systems that are required to perform robust speech recognition.

3.4 Performance of Compensation Algorithms

We now compare and discuss the performance of some of the acoustical preprocessing algorithms described in the previous sections. While the comparisons presented are (of necessity) far from comprehensive, they do serve to illustrate the capabilities and limitations of various alternative approaches to compensation via acoustical pre-processing.

Joint versus Independent Compensation for the Effects of Noise and Filtering

Figure 1 includes recognition error rates obtained using the model-based CDCN algorithm that compensates jointly for the effects of additive noise and linear filtering. When the SPHINX-I system is trained using the CLSTLK microphone and tested using the PZM6FS microphone, the use of joint compensation reduces the recognition error rate by 28.7 percent relative to independent

$$\hat{\mathbf{x}}_i = \sum_{k=0}^{K-1} f_i[k](\mathbf{z} - \hat{\mathbf{q}} - \hat{\mathbf{r}}[k]), \text{ with } i = 0, 1, \cdots, N-1 \tag{11}$$

We have found that these equations normally converge within a very small number of iterations, although they are not guaranteed to do so.

Although model-based compensation is somewhat more computationally intensive than compensation based on empirical comparisons, the bulk of the computational cost is incurred in estimating the distortion parameters $\mathbf{q}$ and $\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q})$. Since distortion due to noise and filtering changes relatively slowly, it is generally not necessary to compute new values for these parameters for every incoming speech frame. The compensation itself (Eq. 11) must be applied to each incoming frame, but this does not entail great computational cost.

Model-based compensation can provide fast and efficient compensation if the assumptions built into the structural model are valid. We compare representative results using model-based compensation with empirical compensation in Section 3.4 below.

3.3 Cepstral high-pass Filtering

The third major adaptation technique is cepstral high-pass filtering, which provides a remarkable amount of robustness at almost zero computational cost. The development of these algorithms was originally motivated by a desire to emphasize the transient aspects of speech representations, as discussed in the chapter in this volume by Hanson *et al.* [16].

In the well-known *Relative Spectral Processing* or *RASTA* processing [15], a high-pass (or band-pass) filter is applied to a log-spectral representation of speech such as the cepstral coefficients. *Cepstral mean normalization* (CMN) is an alternate way to high-pass filter cepstral coefficients. high-pass filtering in CMN is accomplished by subtracting the short-term average of cepstral vectors from the incoming cepstral coefficients.

Algorithms like RASTA and CMN are effective in compensating for the effects of unknown linear filtering in the absence of additive noise because under these circumstances the ideal cepstral compensation vector $\mathbf{v}[SNR, k, \phi, e]$ is a constant that is independent of SNR and VQ cluster. Such a compensation vector is, in fact, equal to the long-term average difference between all cepstra

By making several simplifying assumptions concerning the form of the covariance matrices Σ_k, the correction vectors can be iteratively estimated by first assuming initial values of $\hat{\mathbf{n}}^{(0)}$ and $\hat{\mathbf{q}}^{(0)}$ for $j = 0$. Updated values for $\hat{\mathbf{n}}$ and $\hat{\mathbf{q}}$ are obtained by iterating

$$\mathbf{r}^{(j)}[k] = IDFT(\ln(1 + \exp(DFT[\hat{\mathbf{n}}^{(j)} - \hat{\mathbf{q}}^{(j)} - \mathbf{c}[k]]))) \tag{7}$$

$$f_i[k] = \frac{\exp(-\frac{d_i^2[k]}{2\sigma_s^2})}{\sum_{k=0}^{K-1} \exp(-\frac{d_i^2[k]}{2\sigma_s^2})} \tag{8}$$

where $f_i[k]$ is the weighting constant for Gaussian mixture k in frame i, and the distances $d_i[k]$ are given by $d_i[k] = ||\mathbf{z}_i - \hat{\mathbf{q}}^{(j)} - \mathbf{c}[k] - \hat{\mathbf{r}}^{(j)}[k]||$. The new estimates for $\hat{\mathbf{n}}^{(j+1)}$ and $\hat{\mathbf{q}}^{(j+1)}$ are

$$\hat{\mathbf{n}}^{(j+1)} = \frac{\sum_{i=0}^{N-1} f_i[0]\mathbf{z}_i}{\sum_{i=0}^{N-1} f_i[0]} \tag{9}$$

and

$$\hat{\mathbf{q}}^{(j+1)} = \frac{\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} f_i[k](\mathbf{z}_i - \mathbf{c}[k] - \mathbf{r}^{(j)}[k])}{\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} f_i[k]} \tag{10}$$

Equations (7) through (10) are iterated until $\mathbf{n}$ and $\mathbf{q}$ converge. Once convergent estimates for $\mathbf{n}$ and $\mathbf{q}$ have been obtained, the clean speech vectors are estimated using

and testing domains. An alternate approach to compensation is the use of a parametric model of degradation, combined with optimal estimation of the parameters of the model. For example, Ephraim [9] has presented a unified view of statistical model-based speech enhancement that can be applied to speech enhancement (for human listeners), speech coding, and enhanced robustness for automatic speech recognition systems. Varga and Moore [37] and Gales and Young [12] have also developed algorithms that modify the parameters of HMMs to characterize the effects of noise on speech. Sankar and Lee [32] have used an arbitrary parametric functions to reduce distortions between training and testing environments of the incoming features or model parameters of the HMM. Most of the above approaches have been developed primarily to ameliorate the effects of pure additive noise on speech. Acero's Codeword-Dependent Cepstral Normalization (CDCN) algorithm [1, 2] is similar in principle, except that it was developed explicitly to provide for joint compensation for the effects of additive noise combined with linear filtering.

The CDCN algorithm assumes the model of environmental degradation shown in Fig. 2. The algorithm attempts to reverse the effects of the linear filter with transfer function $H(f)$ and the additive noise with power spectrum $P_n(f)$ by solving two independent problems. The first problem is that of estimating the parameters $\mathbf{q}$ and $\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q})$, the cepstral vectors describing the effects of the noise and filtering in Eq. (2). This is accomplished using ML parameter estimation. The second problem is estimation of the uncorrupted cepstral vector $\mathbf{x}$ for a particular input frame, given the corrupted observation vector $\mathbf{z}$ and the distortion parameters $\mathbf{q}$ and $\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q})$. MMSE parameter estimation is used for this task. In effect, these two operations determine the values of $\mathbf{q}$ and $\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q})$ that when applied in inverse fashion map the set of input cepstra $\mathbf{z}$ into a set of compensated cepstral coefficients $\mathbf{x}$ that are as "close" as possible to the VQ codeword locations encountered in the training data. CDCN is typically implemented on a sentence-by-sentence basis.

We typically use the common representation of Gaussian mixtures for the probability density function (pdf) of the speech signal

$$p(\mathbf{x}) = \sum_{k=0}^{K-1} P[k] p(\mathbf{x}|k) = \sum_{k=0}^{K-1} P[k] N_{\mathbf{X}}(\mathbf{c}[k], \Sigma_k) \tag{6}$$

where the mixture component locations $\mathbf{c}[k]$ are obtained by vector quantizing the cepstral coefficients of speech in the training domain.

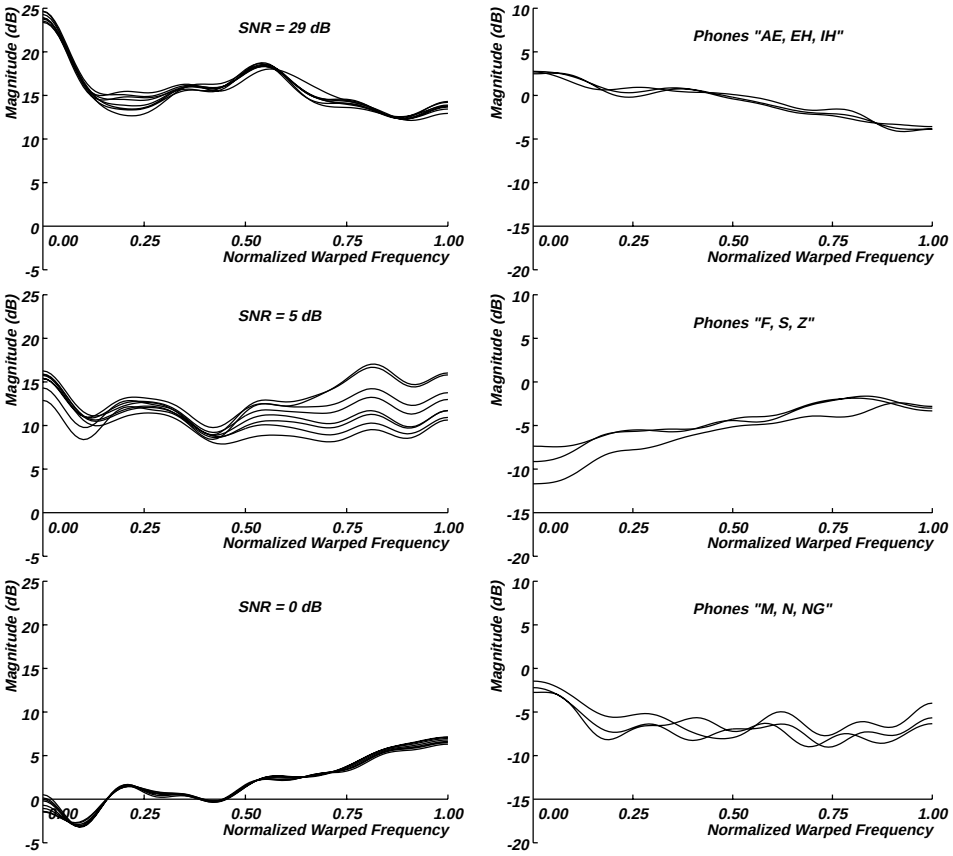

Figure 3 Power spectra of compensation vectors used by the FCDCN algorithm (left panels) and by the PDCN algorithm (right panels). The FCDCN compensation vectors are based on three different SNRs and eight VQ codeword locations at each SNR. The PDCN compensation vectors are based on three different sets of phonemes. The training environment is the standard CLSTLK microphone, while the testing environment is the unidirectional desktop PCC-160 microphone.

3.2 Model-Based Compensation

The compensation algorithms described in the previous section depend on frame-by-frame empirical comparisons of cepstral coefficients in the training

vironments used to develop compensation vectors most closely resembles the actual testing environment. The ensemble of compensation vectors that is appropriate for that most likely environment is then applied to the incoming data. If the incoming speech is not from one of the environments used to develop compensation vectors, recognition accuracy can be further improved by interpolating among the several "closest" environments. Environmental classification need not be perfect for these algorithms to be effective. The "multiple-environment" versions of FCDCN and PDCN are referred to as MFCDCN and MPDCN.

Figure 3 illustrates some typical compensation vectors produced by the MFCDCN and MPDCN algorithms. The standard CLSTLK microphone was used for the training data, and the unidirectional desktop PCC-160 desktop microphone was used in the testing environments. The left column of Fig. 3 depicts MFCDCN compensation vectors, plotted at the extreme SNRs of 0 and 29 dB, as well as at 5 dB. Compensation vectors are plotted for 8 VQ cluster locations at each value of SNR. The curves are obtained by calculating the cosine transforms of the cepstral compensation vectors, $\mathbf{v}[SNR, k, \phi, e]$, and they provide an estimate of the effective spectral profile of the compensation vectors. The horizontal frequency axis is warped nonlinearly according to the mel scale [8]. The maximum frequency corresponds to the Nyquist frequency, 8,000 Hz. We note that the spectral profiles of the compensation vectors vary with SNR. This confirms our assertion that the vectors needed to compensate for the effects of linear filtering (which are dominant at high SNRs) are different from the vectors needed to compensate for the effects of additive noise (which dominate at low SNRs). Furthermore, at intermediate SNRs (such as 5 dB), additional improvement in recognition accuracy can be obtained by developing separate compensation vectors for the different VQ clusters within a given SNR. Compensation vectors for speech frames with SNRs that are greater than 10 dB are very similar in appearance to the compensation vectors shown for 29 dB.

The right column of Fig. 3 depicts similar compensation vectors for the phoneme-based MPDCN algorithm. The right panel depicts MPDCN compensation vectors that are appropriate for three vowels (AE, EH, and IH), three fricatives (F, S, and Z), and three nasals (M, N, and NG). While the overall shape of the MPDCN compensation vectors may primarily reflect differences in the average power, the details of the spectral shapes differ, and the use of phoneme-based compensation in addition to SNR-based compensation can indeed provide further reduction in recognition error rate.

Dependent Cepstral Normalization (SDCN) [2]. Compensation vectors for the SDCN algorithm are developed using a stereo database with simultaneously-recorded speech in the training and testing environments. Individual frames are partitioned into subsets according to the SNR in each frame in the testing environment. (SNR is normally estimated from the total signal power for a given frame). Compensation vectors corresponding to a given range of SNRs are estimated by calculating the average difference between cepstral vectors in the training and testing environments for all frames with that particular range of SNRs. The ensemble of compensation vectors constitutes an empirical characterization of the differences between the training and testing environments. When a new test utterance is presented to the classifier, the SNR is estimated for each frame of the input speech, and the appropriate compensation vector is added to the cepstral coefficients derived from the input speech for that frame.

The *Fixed Codeword-Dependent Cepstral Normalization* (FCDCN) algorithm [2] produces greater recognition accuracy by developing a more fine-grained set of compensation vectors for a particular testing environment. Compensation vectors for FCDCN are obtained by first partitioning the frames of speech from a stereo development corpus according to SNR, as with SDCN. A second partitioning of the development corpus is then obtained by vector quantizing (VQ) the cepstral coefficients at each SNR in the testing environment. Individual compensation vectors are developed for each VQ cluster location at each SNR.

The *Phone-Dependent Cepstral Normalization* (PDCN) algorithm [24] is similar in philosophy, but it makes use of a different type of partitioning of the input frames. Compensation vectors are obtained that depend on the presumed phoneme to which a given frame belongs. Phoneme hypotheses are obtained by running an initial pass of the HMM decoder without compensation. The PDCN algorithm is similar in concept to the method proposed by Beattie and Young [3], except that the latter authors base compensation directly on decoder state and they use a general approach that is more mathematical and less empirical.

Environment-Independent Algorithms

The compensation algorithms described above all are designed to work in the particular testing environment from the stereo database that was used to develop the compensation vectors. A degree of environmental independence can be obtained if several stereo training databases are available using different testing environments. Separate ensembles of compensation vectors can then be developed for each environment for which stereo data are available. Environment-independent compensation is performed by first determining which of the en-

where $\mathbf{v}[SNR, k, \phi, e]$ refers to the additive cepstral compensation vectors. In general, these vectors can depend on instantaneous SNR, the specific vector-quantized (VQ) cluster location k that is nearest to the incoming feature vector (as discussed below), the presumed phonemic identity ϕ, and the specific testing environment e.

Applying the compensation is equally simple, as the compensation vector is just added to the incoming cepstral vector to produce $\hat{\mathbf{x}}$, an estimate of the original cepstral vector, $\mathbf{x}$:

$$\hat{\mathbf{x}} = \mathbf{z} + \mathbf{v}[SNR, k, \phi, e] \tag{5}$$

The goal of compensation is normally to provide relief from the effects of both additive noise and linear filtering, which affect different speech frames differently. Because of this, we have found it advantageous to separate the incoming speech on a frame-by-frame basis into different classes according either to physical parameters such as SNR (as estimated in the testing environment) or according to presumed phonemic identity. Individual compensation vectors are calculated for each of the various classes, and as each incoming speech frame is processed, the additive compensation vector is applied that is appropriate for that particular class.

At high SNRs, the compensation vectors $\mathbf{v}[SNR, k, \phi, e]$ primarily compensate for the effects of linear filtering, because under these circumstances the vector $\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q})$ in Eq. (2) is approximately zero. At the lowest SNRs, the vectors primarily compensate for the effects of additive noise, because under these circumstances Eq. (2) is dominated by the effects of the additive noise . At intermediate SNRs, the compensation vectors perform a combination of compensation for the effects of noise and filtering. Compensation using direct cepstral comparison is generally rather simple to apply, although its utility is limited by the coverage of the stereo training data.

We discuss some aspects of implementations of this approach performed at CMU in the sections below. Similar and complementary work performed at other sites include [27, 28].

Environment-Dependent Compensation Algorithms

Processing by direct cepstral comparison can be best illustrated by first considering the simplest cepstral comparison algorithm developed at CMU, *SNR*-

empirical compensation by direct cepstral comparison, (2) model-based compensation by cepstral remapping, and (3) compensation via cepstral high-pass filtering.

Compensation by direct cepstral comparison is totally data driven, and requires a "stereo" database that contains time-aligned samples of speech that had been simultaneously recorded in the training environment and in representative testing environments. The success of data-driven approaches depends on the extent to which the putative testing environments used to develop the parameters of the compensation algorithm are in fact representative of the actual testing environment.

Compensation by cepstral remapping is a model-based approach. Statistical estimation theory is applied to estimate the parameters representing the effects of noise and filtering in the model for acoustical degradation depicted in Fig. 2. Compensation is then provided by applying the appropriate inverse operations. The success of model-based approaches depends on the extent to which the model of degradation used in the compensation process accurately describes the true nature of the degradation to which the speech had been subjected.

As the name implies, compensation by high-pass filtering implies removal of the steady-state components of the cepstral vector, as is discussed in the chapter by Hanson *et al.* [16] in this volume. The amount of compensation provided by high-pass filtering is more limited than the compensation provided by the two other types of approaches, but the procedures employed are so simple that they should be included in virtually every current speech recognition system.

We now discuss each of these approaches in greater detail.

3.1 Empirical Cepstral Compensation

Empirical cepstral comparison procedures assume the existence of "stereo" databases containing speech that had been simultaneously recorded in the training environment and one or more prototype testing environments. In general, cepstral vectors are calculated on a frame-by-frame basis from the speech in the training and testing environments, and compensation vectors are obtained by computing the differences between average cepstra in the two environments:

$$\mathbf{v}[SNR, k, \phi, e] = \overline{\mathbf{x}} - \overline{\mathbf{z}} \tag{4}$$

sume that the speech signal $x[m]$ is first passed through a linear filter $h[m]$ whose output is then corrupted by uncorrelated additive noise $n[m]$. We characterize the power spectral density (PSD) of the processes involved as

$$P_z(f) = P_x(f)|H(f)|^2 + P_n(f) \tag{1}$$

If we let the cepstral vectors $\mathbf{x}$, $\mathbf{n}$, $\mathbf{z}$, and $\mathbf{q}$ represent the Fourier series expansions of $\ln P_x(f)$, $\ln P_n(f)$, $\ln P_z(f)$, and $\ln|H(f)|^2$, respectively, Eq. (1) can be rewritten with some algebraic manipulation as

$$\mathbf{z} = \mathbf{x} + \mathbf{q} + \mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q}) \tag{2}$$

In this representation the cepstral vectors $\mathbf{z}$ (representing the observed speech) are considered to have been obtained by additive perturbations of the original speech cepstra $\mathbf{x}$. The additive perturbation $\mathbf{q}$ represents the effects of linear filtering while the other additive vector

$$\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q}) = IDFT(\ln(1 + e^{DFT[\mathbf{n}-\mathbf{q}-\mathbf{x}]})) \tag{3}$$

represents the effects of additive noise. In other words, the effect of both additive noise and linear filtering can be represented by additive perturbations to the cepstral representation, although the characterization of the effects of noise as an additive perturbation in the cepstral domain is not a very natural one. In general, our goal is to estimate $\mathbf{x}$, the cepstral representation of $x[m]$ from $\mathbf{z}$, the cepstral representation of $z[m]$. Among other methods, this can be accomplished by estimating $r(\mathbf{x}, \mathbf{n}, \mathbf{q})$ and $\mathbf{q}$, the cepstral parameters characterizing the effects of unknown additive noise and unknown linear filtering, and performing the appropriate inverse operations. Performing compensation in the cepstral domain (as opposed to the spectral domain) has the advantage that a smaller number of parameters needs to be estimated. In addition, cepstral-based features are widely used by current speech recognition systems.

3 ACOUSTICAL PRE-PROCESSING

In this section we examine several types of cepstral compensation algorithms. We have found it convenient to group these algorithms into three classes: (1)

It can be seen from Figure 1 that the use of spectral normalization and spectral subtraction provides increasing degrees of improvement to the recognition accuracy obtained in the "cross" conditions when training and testing environments differ. From these results we can identify two distinct goals of environmental compensation: (1) to eliminate the degradation experienced in the "cross" conditions, and (2) to eliminate the degradation in accuracy experienced when training and testing using the PZM6FS microphone, compared to the error rate obtained when training and testing using the CLSTLK microphone.

We performed additional experiments that evaluated recognition accuracy on this task while applying both spectral subtraction and spectral normalization in sequence. We found that a simple cascade of these two procedures provided no further improvement in error rate beyond that obtained with spectral subtraction alone. We believe the failure to obtain further improvement in recognition accuracy arises from at least two reasons. First, both subtraction and normalization process different frequency components independently, and there is no constraint that ensures that the across-frequency nature of the compensated features is speech-like. In addition, the effects of additive noise and linear filtering combine nonlinearly in the cepstral domain used to derive the features used in classification. Because of this nonlinear interaction, we argue that it is necessary to compensate *jointly* (rather than independently) for the effects of noise and filtering. Such joint compensation is facilitated by the use of the analytical model of degradation described in the next section.

2.3 A Model of Environmental Degradation

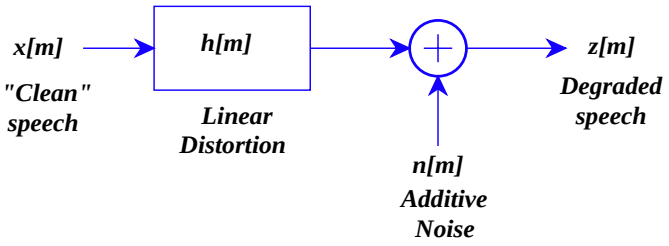

Figure 2 A model of environmental distortion including the effects of additive noise and linear filtering.

Figure 2 describes the implicit model for environmental degradation used in many signal processing algorithms developed at CMU and elsewhere. We as-

2.2 Independent Compensation for Additive Noise and Linear Filtering

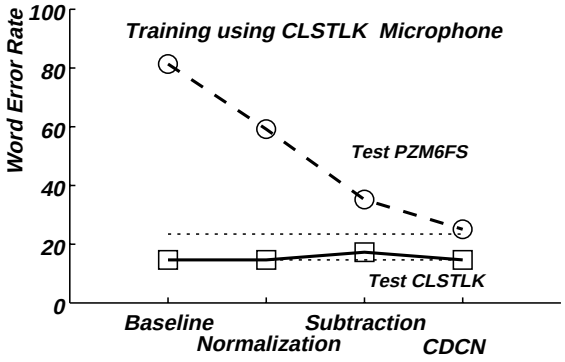

Figure 1 Comparison of error rates obtained by training and testing the SPHINX-I system on the close-talking Sennheiser HMD224 microphone (CLSTLK) and the omnidirectional desktop Crown PZM6FS. Error rates are compared using no environmental compensation, spectral normalization, spectral subtraction, and the CDCN compensation algorithm, for each of the two microphones, on the CMU Census task.

We first consider the ability of spectral subtraction and spectral normalization algorithms applied in isolation to ameliorate the effects of additive noise and linear filtering. Figure 1 summarizes experimental results obtained from a series of initial experiments using a small alphanumeric database called the Census database [2]. This database consists of 1018 training utterances and 140 testing utterances, all recorded simultaneously (*i.e.* in "stereo") using a close-talking Sennheiser HMD224 microphone (CLSTLK), and an omnidirectional desk-top microphone, the Crown PZM6Fs (PZM6FS). The recognition system used was an implementation of the original discrete-HMM system SPHINX-I [19], with between-word statistics eliminated to provide more rapid training and testing. The system was trained using the CLSTLK microphone and tested using the two microphones, in either the baseline condition, or with the use of spectral normalization and spectral subtraction. The upper dotted horizontal line indicates the baseline word error rate obtained when the system was trained and tested using the PZM6FS; the lower horizontal indicates the baseline error rate obtained by training and testing using the CLSTLK microphone. Results are also included for the CDCN algorithm, which is discussed in Sec. 3.2 below.

2 SOURCES OF ENVIRONMENTAL DEGRADATION

2.1 Additive Noise and Linear Filtering

There are many sources of acoustical distortion that can degrade the accuracy of speech recognition systems. For many speech recognition applications the two most important sources of acoustical degradation are *unknown additive noise* (from sources such as machinery, ambient air flow, and speech babble from background talkers) and *unknown linear filtering* (from sources such as reverberation from surface reflections in a room, and spectral shaping by microphones or by the vocal tracts of individual speakers). Other sources of degradation of recognition accuracy include transient interference to the speech signal (such as the noises produced by doors slamming or telephones ringing), nonlinear distortion (arising from sources such as carbon-button microphones or the random phase jitter in telephone systems), and "co-channel" interference by individual competing talkers. Until now, most research in robust recognition has been directed toward compensation for the effects of additive noise and linear filtering.

Research in robust speech recognition has been strongly influenced by earlier work in speech enhancement. Two seminal speech enhancement algorithms have proved to be especially important in the development of strategies to cope with unknown noise and filtering. The first technique, *spectral subtraction*, was introduced by Boll [6] to compensate for additive noise. In general, spectral subtraction algorithms attempt to estimate the power spectrum of additive noise in the absence of speech, and then subtract that spectral estimate from the power spectrum of the overall input (which normally includes the sum of speech plus noise). The algorithm was later extended by Berouti *et al.* [4] and many others, primarily with the goal of avoiding "musical noise" by "over-subtraction" of the the noise spectrum. The second major technique is *spectral normalization*, introduced by Stockham *et al.* [34] to compensate for the effects of unknown linear filtering. In general, spectral normalization algorithms first attempt to estimate the average power spectra of speech in the training and testing domains, and then apply the linear filter to the testing speech to "best" converts its spectrum to that of the training speech. Improvements and extensions of spectral subtraction and spectral normalization algorithms continue to be introduced to this date.

automobiles, on a factory floor, or outdoors demand an even greater degree of environmental robustness.

The goal of this chapter is to provide a comprehensive review of research at Carnegie Mellon University (CMU) and elsewhere that is directed toward rendering speech recognition systems more robust with respect to environmental variation. Historically, the greatest amount of effort in robust recognition has been devoted to *acoustical pre-processing* algorithms, which typically modify either the features extracted from incoming speech or the representation of these features by the recognition system in order to reduce mismatches between training and testing conditions. In recent years, however, other complementary approaches to robust recognition are becoming increasingly popular. For example, *arrays of multiple microphones* have the ability to improve speech recognition accuracy by improving the signal-to-noise ratio (SNR) when signal and noise sources arrive from spatially-distinct sources. Other research groups have focussed on the use of *signal processing algorithms based on human audition*, motivated by the observation that the feature set developed by the human auditory system is remarkably robust.

We begin this chapter with a description of some of the sources of degradation that reduce the accuracy of speech recognition systems in Sec. 2, and we briefly review some of the classical approaches to environmental robustness in that section. In Sec. 3 we describe three approaches to acoustical pre-processing for environmental robustness: (1) empirical approaches in which compensation parameters are estimated by direct comparison of speech features in the training and testing environments, (2) model-based approaches in which parameters of a structural model of acoustical degradation are obtained by optimal estimation, and (3) cepstral high-pass filtering, which enables the system to obtain a more limited amount of compensation in a very computationally-efficient fashion. In Secs. 4 and 5, respectively, we compare recognition results obtained using acoustical pre-processing to results obtained using microphone arrays and physiologically-motivated signal processing strategies. Finally, we summarize our findings in Sec. 6.

14

SIGNAL PROCESSING FOR ROBUST SPEECH RECOGNITION

Richard M. Stern, Alejandro Acero*, Fu-Hua Liu, Yoshiaki Ohshima*****

Carnegie Mellon University, Pittsburgh, PA 15213, USA

** Microsoft Corporation*
*** IBM Thomas J. Watson Laboratory*
**** IBM Tokyo Research Laboratory*

ABSTRACT

This chapter compares several different approaches to robust automatic speech recognition. We review ongoing research in the use of acoustical pre-processing to achieve robust speech recognition, discussing and comparing approaches based on direct cepstral comparisons, on parametric models of environmental degradation, and on cepstral high-pass filtering. We also describe and compare the effectiveness of two complementary methods of signal processing for robust speech recognition: microphone array processing and the use of physiologically-motivated models of peripheral auditory processing. This chapter includes comparisons of recognition error rates obtained when the various signal processing algorithms considered are used to process inputs to CMU's SPHINX speech recognition system.

1 INTRODUCTION

The development of robust speech recognition systems that maintain a high level of recognition accuracy in difficult and dynamically-varying acoustical environments is becoming increasingly important as speech technology is becoming a more integral part of practical applications. Results of numerous 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.* [1, 2, 18]), even in a relatively quiet office environment. Applications such as speech recognition over telephones, in