nica microphone than for the Sennheiser microphone in the noisier environment. In the computer room environment, the performance with the Audio-Technica microphone is almost indistinguishable from that of the Sennheiser recording.

Experiment		Word Error	
		Sennheiser	Secondary Microphone
Audio-Technica Recordings	Env 1	6.3	8.5
	Env 2	9.1	17.4
Telephone Handset Recordings	Env 1	8.4	29.1
	Env 2	8.3	28.8

Table 9: Word Error for both Sennheiser and Secondary Microphone with Robust Signal Processing Front End when Recorded in Two Noisy Environments

5. CONCLUSIONS

We have presented a feature-mapping algorithm capable of exploiting nonlinear relations between two acoustic spaces. We have shown how to improve the performance of the recognizer in the presence of a noisy signal by using a small database with simultaneous recordings in the clean and noisy acoustic environments.

We have shown that

- There is no significant difference in speech recognition performance between those obtained with an Audio-Technica microphone and those obtained with a Sennheiser microphone. There is no significant performance degradation in a quiet environment and only a slight degradation in low-noise environments (~59 dBA).
- Multidimensional least-squares filters can be successfully used to exploit the correlation of the features over time and among components of the spectral features at the same time. These filters can be conditioned on both local and global spectral information to improve robust recognition performance.
- Most of the performance loss in converting wide-bandwidth models to telephone speech models is due to the loss of information associated with the telephone bandwidth.
- It is possible to construct acoustic models for telephone speech using a high-quality speech corpus with only a minor increase in recognition word error rate.
- A telephone-bandwidth system trained with high-quality speech can outperform a system that is trained on telephone speech even when tested on telephone speech.
- The variability introduced by the telephone handset does not degrade speech recognition performance.
- Robust signal processing can be designed to maintain speech recognition performance using wide-bandwidth HMM models with a telephone-bandwidth test set.

ACKNOWLEDGMENTS

The authors thank John Butzberger for helping to set up the ATIS experiments, and Vassilios Digalakis for providing genonic HMM models and helping with the telephone-bandwidth experiments.

This research was supported by a grant, NSF IRI-9014829, from the National Science Foundation (NSF). It was also supported by the Advanced Research Projects Agency (ARPA) under Contracts ONR N00014-93-C-0142 and ONR N00014-92-C-0154.

REFERENCES

1. M. Weintraub and L. Neumeyer, “Constructing Telephone Acoustic Models from a High-Quality Speech Corpus,” 1994 IEEE ICASSP.
2. V. Digalakis and H. Murveit, “GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer,” 1994 IEEE ICASSP.
3. H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, “Large-Vocabulary Dictation Using SRI’s DECIPHER Speech Recognition System: Progressive Search Techniques,” 1993 IEEE ICASSP, pp. II319-II322.
4. H. Murveit, J. Butzberger, and M. Weintraub, “Performance of SRI’s DECIPHER™ Speech Recognition System on DARPA’s CSR Task,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 410-414.
5. Murveit, H., J. Butzberger, and M. Weintraub, “Reduced Channel Dependence for Speech Recognition,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 280-284.
6. A. Erell and M. Weintraub, “Spectral Estimation for Noise Robust Speech Recognition,” 1989 DARPA SLS Workshop, pp. 319-324.
7. A. Erell and M. Weintraub, “Recognition of Noisy Speech: Using Minimum-Mean Log-Spectral Distance Estimation,” 1990 DARPA SLS Workshop, pp. 341-345.
8. B.H. Juang and L.R. Rabiner, “Signal Restoration by Spectral Mapping,” 1987 IEEE ICASSP, pp. 2368-2371.
9. H. Gish, Y.L. Chow, and J.R. Rohlicek, “Probabilistic Vector Mapping of Noisy Speech Parameters for HMM Word Spotting,” 1990 IEEE ICASSP, pp. 117-120.
10. K. Ng, H. Gish, and J.R. Rohlicek, “Robust Mapping of Noisy Speech Parameters for HMM Word Spotting,” 1992 IEEE ICASSP, pp. II-109-II-112.
11. A. Acero, “Acoustical and Environmental Robustness in Automatic Speech Recognition,” Ph.D. Thesis, Carnegie-Mellon University, September 1990.
12. R.M. Stern, F.H. Leu, Y. Ohshima, T.M. Sullivan, and A. Acero, “Multiple Approaches to Robust Speech Recognition,” 1992 International Conference on Spoken Language Processing, pp. 695-698.
13. A. Nadas, D. Nahamoo, and M. Picheny, “Adaptive Labeling: Normalization of Speech by Adaptive Transformations based on Vector Quantization,” 1988 IEEE ICASSP, pp. 521-524.
14. Y. Linde, A. Buzo, and R.M. Gray, “An Algorithm for Vector Quantizer Design,” IEEE Trans. Comm., vol. 28, pp. 84-95, January 1980.
15. G. Doddington, “CSR Corpus Development,” 1992 DARPA SLS Workshop, pp. 363-366.
16. S.F. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification,” IEEE Trans. ASSP, Vol. 29, pp. 254-272, April 1981.

4. WSJ EXPERIMENTAL RESULTS

Another series of experiments was performed on the WSJ Speech Corpus [15]. We evaluated our system on the 5000-word-recognition closed-vocabulary speaker-independent speech-recognition tasks: Spoke S5 Unknown Microphone, Spoke S6: Known Microphone, and Spoke S7 Noisy Environment.

The version of the DECIPHER speaker-independent continuous speech recognition system used for these experiments is based on a progressive-search strategy [3] and continuous-density, genonic HMMs [2]. Gender-dependent models are used in all passes. Gender selection uses the models with the higher recognition likelihood.

The acoustic models used by the HMM system were trained with 37,000 sentences of Sennheiser data from 280 speakers, a set officially designated as the WSJ0+WSJ1 many-speaker baseline training. A 5,000 closed-vocabulary back-off trigram language model provided by M.I.T. Lincoln Laboratory for the WSJ task was used. Gender-dependent HMM acoustic models were used.

The front-end processing extracts one long spectral vector consisting of the following six feature components: cepstrum, energy, and their first and second order derivatives. The dimensionality of this feature is 39 (13 * 3) for the wide-bandwidth spectral analysis and 27 (9 * 3) for the telephone-bandwidth spectral analysis. The cepstral features are computed from an FFT filterbank, and subsequent cepstral-mean normalization on a sentence-by-sentence basis is performed.

Before using wide-bandwidth context-dependent genonic HMMs, a robust estimate of the Sennheiser cepstral parameters is computed using POF. The robust front-end analysis is designed for an unknown microphone condition. The POF mapping algorithm estimates are conditioned on the noisy cepstral observations. Separate mappings are trained for each of the 14 microphones in the baseline WSJ0+WSJ1 si_tr_s stereo training, and one mapping for the overall case of single nontelephone mapping. When the default no-transformation zero-mean cepstra are included, this makes a total of 15 estimated feature streams. These feature streams are computed on each test waveform, and the two feature streams with the highest likelihoods (using a simplified HMM for scoring the features) are averaged together (Top2). In all cases the first and second delta parameters are computed on these estimated cepstral values.

Front-End Bandwidth	Signal Processing	Test Set	Word Error (%)
Wide	Standard	Sennheiser	5.8
Telephone	Standard	Sennheiser	9.6
Telephone	Standard	Telephone	10.9
Wide	Robust POF15 Cepstral Mapping	Telephone	11.9

Table 6: Performance on the Aug 1993 WSJ Spoke S6 development test set for simultaneous Sennheiser/telephone recordings

The results in Table 6 show that most of the loss in performance between recognizing on high-quality Sennheiser recordings and on local telephone speech is due to the loss of information outside the telephone bandwidth. There is an increase in the word-error rate of 66% when testing on Sennheiser recordings with a wide-bandwidth analysis (5.8%) compared to testing with a telephone-bandwidth analysis (9.6%).

The loss in performance when switching from Sennheiser recordings to telephone recordings is small in comparison to the loss of information due to bandwidth restrictions. There is a 14% increase in the word error rate when testing on the Sennheiser recordings (9.6%) compared to testing on the AT&T telephone recordings (10.9%).

4.1. Official Spoke Results: Unknown Microphone

The results in Table 7 show the speech recognition performance when the secondary microphone condition is unknown. In these experiments, the robust signal processing front end decreased the word error rate from 17.2 to 13.1%.

Experiment	Word Error	
	Sennheiser	Secondary Microphone
Compensation Disabled	6.6	17.2
Compensation Enabled	6.6	13.1

Table 7: Word error rate with and without compensation on both Sennheiser and secondary microphone data

4.2. Official Spoke Results: Known Microphone

The results in Table 8 show no significant difference in speech recognition performance between those obtained with an Audio-Technica microphone and those obtained with the Sennheiser microphone. The robust front-end signal processing has demonstrated for the first time that one can achieve the same performance with a stand-mounted microphone as with a high-quality close-talking microphone, all when trained on a high-quality speech corpus.

Experiment	Word Error	
	Sennheiser	Secondary Microphone
Audio-Technica Recordings	5.9	6.4
Telephone Handset Recordings	7.2	19.1

Table 8: Word Error for both Sennheiser and Secondary Microphone with Robust Signal Processing Front End

4.3. Official Spoke Results: Noisy Environment

The results in Table 9 show the performance when the recordings are made in a noisy environment. The first noisy environment was a computer room (average background noise level of 58 to 59 dBA), and the second noisy environment was a laboratory with mail sorting equipment (average noise level varied from 62 to 68 dBA). The error rates are significantly higher for the Audio-Tech-

.

Acoustic Model Training		Test Set Word Error (%)	
Training Data	Front-End Bandwidth	Sennheiser	Telephone
Sennheiser	Wide	7.8	19.4
Sennheiser	Telephone	9.0	9.7
Telephone	Telephone	10.0	10.3

Table 3: Effect of different training and front-end bandwidth on test set performance. Results are word error rate on the 400 Sentence simultaneous test set.

We can see from Table 3 that there is a 15.4% decrease in performance when using a telephone front end (7.8% increases to 9.0% word error) and testing on Sennheiser data. This is due to the loss of information in reducing the bandwidth from 100-6400 Hz to 300-3300 Hz. However, when we are using a telephone front end, there is only a 7.8% increase in word error when testing on telephone speech compared to testing on Sennheiser speech (9.7% versus 9.0%). This is a very surprising result, and we had expected a much bigger performance difference when Sennheiser models are tested on telephone speech acoustics.

3.3. Multiple Microphones: Single or Multiple Mapping

The POF mapping algorithm can be used in a number of ways when the microphone is unknown. Some of these variations are shown in Table 4.

Experiment		Word Error
Single Mapping Combining All 10 Telephones in Training Data		9.4
Train 10 Mappings, One for Each Telephone; Run 10 Recognizers in Parallel, each using Different Mapping; Select Recognizer with Highest Probability		9.2
Train 10 Mappings, One for Each Telephone; Run 10 Mappings in Parallel and Average Features of Best N Feature-Streams that Have Highest Likelihood	Top1	9.3
	Top2	9.2
	Top3	8.9
	Top4	8.7
Train 15 Mappings for WSJ Corpus; Run 15 Mappings in Parallel and Average Features of Best N Feature-Streams that Have the Highest Likelihood	Top1	9.8
	Top2	9.6
	Top3	10.3
	Top4	10.7

Table 4: Performance on the multiple-telephone handset test set when mapping algorithm is used in different ways.

The differences between the experimental conditions are small, but the trends are different and depend on the mapping and the corpus. These differences depend on the similarities of the different microphones that are used in training conditions, and the relationship between the training and the testing conditions.

When the microphones are all similar (10 telephone mappings), then averaging the features of each mapping helps improve performance. When the microphones are very different (e.g., those in the WSJ corpus), averaging the features of each mapping has a minimum when averaging two best (likelihood) feature streams.

3.4. Multiple Microphones: Conditioning Feature

The next experiment varied the conditioning feature. The conditioning feature is the feature vector used to divide the space into different acoustic regions. In each region of the acoustic space a different linear transformation is trained.

The mapping approach was fixed: we used a single POF mapping for multiple telephone handsets. For this experiment we mapped the cepstrum vector *(c1-c12)* and the cepstral energy *(c0)*. The maximum delay of the filters was kept fixed at *p=2*, and the number of Gaussians was 512. The experimental variable was the feature the estimates were conditioned on. We tried the following conditioning features:

- **Cepstrum**. Same conditioning feature used in the single microphone experiment *(c0-c12)*.
- **Spectral SNR**. This is an estimate of the instantaneous signal-to-noise ratio computed on the log-filterbank energy domain. The vector size is 25.
- **Cepstral SNR**. This feature is generated by applying the discrete cosine transform (DCT) to the spectral SNR. The transformation reduces the dimensionality of the vector from 25 to 12 elements.

The results are shown in Table 5. The baseline result is a 19.4% word error rate. This result is achieved when the same wide-band front end is used for training the models with clean data and for recognition using telephone data. When a telephone front end [1] is used for training and testing, the error decreases to 9.7%. The disadvantage of using this approach is that the acoustic models of the recognizer have to be reestimated. However, the POF-based front end operates on the clean models and results in better performance. The cepstral SNR produces the best result (8.7%). With this conditioning feature we combine the effects of noise and spectral shape in a compact representation.

Experiment	Word Error (%)	Error Ratio
Wide-band front-end	19.4	2.49
Telephone-bandwidth front-end	9.7	1.24
Mapping with cepstrum	9.4	1.20
Mapping with spectral SNR	8.9	1.14
Mapping with cepstral SNR	8.7	1.11

Table 5: Performance for the multiple-telephone handset test set when varying the conditioning feature.

is a probabilistic nonsingular auto-correlation matrix, and

$$\boldsymbol{r}_i = \sum_{n=p}^{N-1-p} \boldsymbol{Y}_n \boldsymbol{x}_n^T p(g_i|\boldsymbol{z}_n) \quad (6)$$

is a probabilistic cross-correlation matrix.

The algorithm can be completely trained without supervision and requires no additional information other than the simultaneous waveforms.

The run-time estimate of the clean feature vector can be computed by integrating the outputs of all the filters as follows:

$$\hat{\boldsymbol{x}}_n = \sum_{i=0}^{I-1} \boldsymbol{W}_i^T \boldsymbol{Y}_n p(g_i|\boldsymbol{z}_n) = \left\{ \sum_{i=0}^{I-1} \boldsymbol{W}_i^T p(g_i|\boldsymbol{z}_n) \right\} \boldsymbol{Y}_n \quad (7)$$

3. EXPERIMENTS

A series of experiments show how the mapping algorithm can be used in a continuous speech recognizer across acoustic environments. In all of the experiments the recognizer models are trained with data recorded with high-quality microphones and digitally sampled at 16,000 Hz. The analysis frame rate is 100 Hz.

The tables below show three types of performance indicators:

- *Relative distortion measure*. For a given component of a feature vector we define the relative distortion between the clean and noisy data as follows:

$$d = \sqrt{\frac{E[(x-y)^2]}{\mathrm{var}(x)}} \quad (8)$$

- *Word recognition error.*

- *Error ratio.* The error ratio is given by E_n / E_c where E_n is the word recognition error for the test-noisy/train-clean condition, and E_c is the word recognition error of the test-clean/train-clean condition.

3.1. Single Microphone

To test the POF algorithm on a single target acoustic environment we used the DARPA Wall Street Journal database [15] on SRI's DECIPHER™ phonetically tied-mixture speech recognition system [2]. The signal processing consisted of a filterbank-based front end that generated six feature streams: cepstrum *(c1-c12)*, cepstral energy *(c0)*, and their first- and second-order derivatives. Cepstral-mean normalization [16] was used to equalize the channel. We used simultaneous recordings of high-quality speech (Sennheiser 414 head-mounted microphone with a noise-canceling element) along with speech recorded by a standard speaker phone (AT&T 720) and transmitted over local telephone lines. We will refer to this stereo data as *clean* and *noisy* speech, respectively. The models of the recognizer were trained using 42 male WSJ0 training talkers (3500 sentences) recorded with a Sennheiser microphone. The models of the mapping algorithm were trained using 240 development training sentences recorded by three speakers. The test set consisted of 100 sentences (not included in the training set) recorded by the same three speakers.

In this experiment we mapped two of the six features: the cepstrum *(c1-c12)* and the cepstral energy *(c0)* separately. The derivatives were computed from the mapped vectors of the cepstral features. For the conditioning feature we used a 13-dimensional cepstral vector *(c0-c12)* modeled with 512 Gaussians with diagonal covariance matrices. The results are shown in Table 2.

Filter Coefficients	Average Distortion	Recognition Error (%)	Error Ratio
No mapping	0.72	27.6	2.46
$\mathbf{A}_{i,0}$=**I**, $\mathbf{b}_i$	0.62	18.1	1.62
$\mathbf{A}_{i,0}$, $\mathbf{b}_i$	0.57	17.0	1.52
$\boldsymbol{A}_{i,-1}$,.., $\boldsymbol{A}_{i,-1}$, $\boldsymbol{b}_i$	0.51	17.3	1.54
$\boldsymbol{A}_{i,-2}$,.., $\boldsymbol{A}_{i,-2}$, $\boldsymbol{b}_i$	0.50	16.4	1.46
$\boldsymbol{A}_{i,-3}$,.., $\boldsymbol{A}_{i,-3}$, $\boldsymbol{b}_i$	0.49	15.9	1.42
$\boldsymbol{A}_{i,-4}$,.., $\boldsymbol{A}_{i,-4}$, $\boldsymbol{b}_i$	0.49	16.1	1.44

Table 2: Performance of the POF algorithm for different number of filter coefficients. The number of Gaussian distributions is 512 per feature and the conditioning feature is a 13-dimensional cepstral vector.

The baseline experiment produced a word error rate of 27.6% on the noisy test set, that is, 2.46 times the error obtained when using the clean data channel. A 34% improvement in recognition performance was obtained when using only the additive filter coefficient $\boldsymbol{b}_i$. (Recognition error goes down to 18.1%.) The best result (15.9% recognition error) was obtained for the condition $p=3$, in which six neighboring noisy frames are being used to estimate the feature vector for the current frame. The correlation between the average relative distortion between the six clean and noisy features and the recognition error is 0.9.

3.2. ATIS Simultaneous Corpus

To test the performance of the POF algorithm on multiple microphones we used SRI's stereo-ATIS database. (See [1] for details.) A corpus of both training and testing speech was collected using simultaneous recordings made from subjects wearing a Sennheiser HMD 414 microphone and holding a telephone handset. The speech from the telephone handset was transmitted over local telephone lines during data collection. Ten different telephone handsets were used. Ten male speakers were designated as training speakers, and three male speakers were designated as the test set. The training set consisted of 3,000 simultaneous recordings of Sennheiser microphone and telephone speech. The test set consisted of 400 simultaneous recordings of Sennheiser and telephone speech. The results obtained with this pilot corpus are shown in Table 3.

- The mapping algorithms described in this paper are able to incorporate many pieces of information that have been traditionally difficult to incorporate into HMM models and into adaptation algorithms. These include observations that span across several frames and the correlation of the state features with global characteristics of the speech waveform.

These two techniques are not mutually exclusive and can be used together to achieve robust speech recognition performance. The boundary between these two techniques can be blurred when the mapping algorithm is dependent on the speech recognizer's hypothesis.

2. THE POF ALGORITHM

The mapping algorithm is based on a probabilistic piecewise-nonlinear transformation of the acoustic space that we call *Probabilistic Optimum Filtering* (POF). Let us assume that the recognizer is trained with data recorded with a high-quality close-talking microphone (clean speech), and the test data is acquired in a different acoustic environment (noisy speech). Our goal is to estimate a clean feature vector $\hat{\boldsymbol{x}}_n$ given its corresponding noisy feature $\boldsymbol{y}_n$ where n is the frame index. (A list of symbols is shown in Table 1.) To estimate the clean vector we vector-quantize the clean feature space in I regions using the generalized Lloyd algorithm [14]. Each VQ region is assigned a multidimensional transversal filter (see Figure 1). The error between the clean vector and the

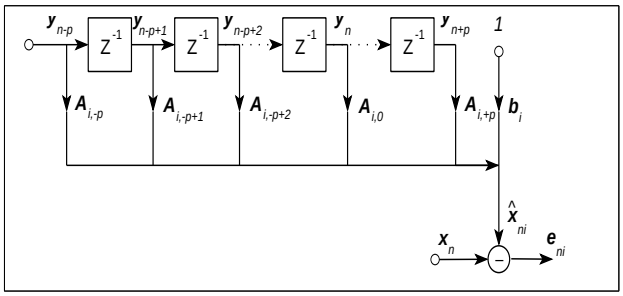

Figure 1: Multi-dimensional transversal filter for cluster *i*.

estimated vectors produced by the *i*-th filter is given by

$$\boldsymbol{e}_{ni} = \boldsymbol{x}_n - \hat{\boldsymbol{x}}_{ni} = \boldsymbol{x}_n - \boldsymbol{W}_i^T \boldsymbol{Y}_n \tag{1}$$

where $\boldsymbol{e}_{ni}$ is the error associated with region i, $\boldsymbol{W}_i$ is the filter coefficient matrix, and $\boldsymbol{Y}_n$ is the tapped-delay line of the noisy vectors. Expanding these matrices we get

$$\boldsymbol{W}_i^T = \left[\boldsymbol{A}_{i,-p} \cdots \boldsymbol{A}_{i,-1}\ \boldsymbol{A}_{i,0}\ \boldsymbol{A}_{i,1} \cdots \boldsymbol{A}_{i,p}\ \boldsymbol{b}_i\right] \tag{2}$$

$$\boldsymbol{Y}_n^T = \left[\boldsymbol{y}_{n-p}^T \cdots \boldsymbol{y}_{n-1}^T\ \boldsymbol{y}_n^T\ \boldsymbol{y}_{n+1}^T \cdots \boldsymbol{y}_{n+p}^T\ 1\right] \tag{3}$$

The conditional error in each region is defined as

$$E_i = \sum_{n=p}^{N-1-p} \left\|\boldsymbol{e}_{ni}\right\|^2 p(g_i|\boldsymbol{z}_n) \tag{4}$$

where $p(g_i|\boldsymbol{z}_n)$ is the probability that the clean vector $\boldsymbol{x}_i$ belongs to region g_i given an arbitrary conditional noisy feature vector $\boldsymbol{z}_n$. Note that the conditioning noisy feature can be any acoustic vector generated from the noisy speech frame. For example, it may include an estimate of the SNR, energy, cepstral energy, cepstrum, and so forth.

The conditional probability density function $p(\boldsymbol{z}_n|g_i)$ is modeled as a mixture of I Gaussian distributions. Each Gaussian distribution models a VQ region. The parameters of the distributions (mean vectors and covariance matrices) are estimated using the corresponding $\boldsymbol{z}_n$ vectors associated with that region. The posterior probabilities $p(g_i|\boldsymbol{z}_n)$ are computed using Bayes' theorem and the mixture weights $p(g_i)$ are estimated using the relative number of training clean vectors that are assigned to a given VQ region.

Symbol	Dimension	Description
n	*1*	frame index
i	*1*	region index
L	*1*	feature vector size
M	*1*	conditioning feature vector size
N	*1*	number of training frames
I	*1*	number of VQ regions
p	*1*	maximum filter delay
$\boldsymbol{e}_{ni}$	$L \times 1$	estimation error vector
$\boldsymbol{x}_n$	$L \times 1$	clean feature vector
$\hat{\boldsymbol{x}}_n$	$L \times 1$	estimate of clean feature vector
$\boldsymbol{y}_n$	$L \times 1$	noisy feature vector
$\boldsymbol{z}_n$	$M \times 1$	conditioning noisy feature vector
μ_i	$M \times 1$	mean vector of gaussian i
Σ_i	$M \times M$	covariance matrix of gaussian i
$\boldsymbol{W}_i$	$(2p+1)L+1 \times L$	transversal filter coefficient matrix
$\boldsymbol{Y}_n$	$(2p+1)L+1 \times 1$	tap input vector
$\boldsymbol{A}_{ik}$	$L \times L$	multiplicative tap matrix
$\boldsymbol{b}_i$	$L \times 1$	additive tap matrix
$\boldsymbol{R}_i$	$(2p+1)L+1 \times (2p+1)L+1$	auto-correlation matrix
$\boldsymbol{r}_i$	$(2p+1)L+1 \times L$	cross-correlation matrix

Table 1: List of symbols

To compute the optimum filters in the mean-squared error sense, we minimize the conditional error in each VQ region. The minimum mean-squared error vector is obtained by taking the gradient of E_i defined in Eq. (4) with respect to the filter coefficient matrix and equating all the elements of the gradient matrix to zero. As a result, the optimum filter coefficient matrix has the form, $\boldsymbol{W}_i = \boldsymbol{R}_i^{-1}\boldsymbol{r}_i$ where

$$\boldsymbol{R}_i = \sum_{n=p}^{N-1-p} \boldsymbol{Y}_n \boldsymbol{Y}_n^T p(g_i|\boldsymbol{z}_n) \tag{5}$$

MICROPHONE-INDEPENDENT ROBUST SIGNAL PROCESSING USING PROBABILISTIC OPTIMUM FILTERING

Leonardo Neumeyer and Mitchel Weintraub

SRI International
Speech Technology and Research Laboratory
333 Ravenswood Avenue
Menlo Park, CA 94025

ABSTRACT

A new mapping algorithm for speech recognition relates the features of simultaneous recordings of clean and noisy speech. The model is a piecewise nonlinear transformation applied to the noisy speech feature. The transformation is a set of multidimensional linear least-squares filters whose outputs are combined using a conditional Gaussian model. The algorithm was tested using SRI's DECIPHER™ speech recognition system [1-5]. Experimental results show how the mapping is used to reduce recognition errors when the training and testing acoustic environments do not match.

1. INTRODUCTION

In many practical situations an automatic speech recognizer has to operate in several different but well-defined acoustic environments. For example, the same recognition task may be implemented using different microphones or transmission channels. In this situation it may not be practical to recollect a speech corpus to train the acoustic models of the recognizer. To alleviate this problem, we propose an algorithm that maps speech features between two acoustic spaces. The models of the mapping algorithm are trained using a small database recorded simultaneously in both environments.

In the case of steady-state additive homogenous noise, we can derive a MMSE estimate of the clean speech filterbank-log energy features using a model for how the features change in the presence of this noise [6-7]. In these algorithms, the estimated speech spectrum is a function of the global spectral signal-to-noise ratio (SNR), the instantaneous spectral SNR, and the overall spectral shape of the speech signal. However, after studying simultaneous recordings made with two microphones, we believe that the relationship between the two simultaneous features is nonlinear. We therefore propose to use a piecewise-nonlinear model to relate the two feature spaces.

1.1. Related Work on Feature Mapping

Several algorithms in the literature have focused on experimentally training a mapping between the noisy features and the clean features [8-13]. The proposed algorithm differs from previous algorithms in several ways:

- The MMSE estimate of the clean speech features in noise is trained experimentally rather than with a model as in [6, 7].
- Several frames are joined together similar to [13].
- The conditional PDF is based on a generic noisy feature not necessarily related to the feature that we are trying to estimate. For example, we could condition the estimate of the cepstral energy on the instantaneous spectral SNR vector.
- Multidimensional least-squares filters are used for the mapping transformation. This exploits the correlation of the features over time and among components of the spectral features at the same time.
- Linear transformations are combined together without hard decisions.
- All delta parameters are computed after mapping the cepstrum and cepstral energy.
- The mapping parameters are trained using stereo recordings with two different microphones. Once trained, the mapping parameters are fixed.
- The algorithm can either map noisy speech features to clean features during training, or clean features to noisy features during recognition.

1.2. Related Work on Adaptation

The algorithm used to map the incoming features into a more robust representation has some similarities to work on model adaptation. Some of the high-level differences between hidden Markov model (HMM) adaptation and the mapping algorithms proposed in this paper are:

- The mapping algorithm works by primarily correcting shifts in the mean of the feature set that are correlated with observable information. Adapting HMM model parameters has certain degrees of freedom that the mapping algorithm does not have- for example the ability to change state variances, and mixture weights.
- Two HMM states that have identical probability distributions and are not tied can have different distributions after adaptation. These distributions cannot be differentiated by mapping features.