From the results of Table 4, we see that:

- When we train using a sample of the testing noise we get better performance than when we train on multiple car noises.
- Mapping the full 39-dimensional cepstral vector (cep + first and second order derivatives) seems to perform better at higher SNR's than mapping only the cepstrum and computing the first and second derivatives on the mapped features.
- Condition 1 shows the performance with no compensation, and how the algorithms help more at higher SNR levels.

3.2.2. Evaluation Test Results

We trained many different POF mappings and HMM's, and selected the appropriate mapping at runtime. Using a one-minute sample of noise, we trained gender-dependent POF mappings for many different SNR levels. The gender selection was done using a Bayesian classifier trained with noisy data at a medium SNR level. The SNR was computed using the average of the log spectral SNR computed at the output of the filterbank in the signal processing stage. (This produced SNRs higher than the ones computed in Section 3.1., and is denoted SNR_spec).

To create the compensation models, the one-minute adaptation noise was added to a subset of the WSJ training data consisting of 300 waveforms with a variable scale creating gender and SNR-specific compensation data sets. The 300 waveform compensation sets were used to train both the mapping and the adaptation parameters. At low SNR_spec levels (9-24 dB), we used the combined method (POF + Adaptation), and at high SNR_spec levels (27-33 dB) we used the POF mapping alone. The results of this test are shown in Table 5. For the worst condition (Level 3) the ratio of the clean-speech error to the noisy-speech error was reduced roughly from 5 to 2 after applying the compensation algorithm.

Compensation	Clean	Level 1	Level 2	Level 3
Enabled	-	10.1	8.8	12.5
Disabled	7.1	18.7	11.5	35.0

Table 5. Word error rates for the 1994 ARPA-sponsored evaluation on the Spoke 10 test.

4. SUMMARY

This paper describes how to compensate HMM-based recognizers in the presence of steady additive noise. We compared performance of compensation algorithms that operate in the feature and model domains, and experimentally found that both approaches produced improved results over the baseline condition. A combination of mapping and adaptation, however, yielded the best results at low SNR levels.

ACKNOWLEDGMENTS

This research was partially supported by a grant, NSF IRI-9014829, from the National Science Foundation and by the Advanced Research Projects Agency through Office of Naval Research Contracts ONR N00014-92-C-0154.

REFERENCES

1. 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, pp. I537-I540.
2. 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.
3. 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.
4. Linguistic Data Consortium, "ARPA Spoken Language Systems November 1994 CSR Hub and Spoke Benchmark Test Material," LDC CDROM Disk T8-1.1, file: ./et94spec.doc.
5. L. Neumeyer and M. Weintraub, "Probabilistic Optimum Filtering for Robust Speech Recognition," 1994 IEEE ICASSP, pp. I417-I420.
6. 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.
7. A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," Ph.D. Thesis, Carnegie-Mellon University, September 1990.
8. A. Erell and M. Weintraub, "Filterbank-Energy Estimation Using Mixture and Markov Models for Recognition of Noisy Speech," 1993 IEEE ASSP, vol. 1, no. 1, pp. 68-76.
9. M.J.F. Gales and S.J. Young, "HMM Recognition in Noise using Parallel Model Combination," 1993 Eurospeech, pp. 837-840.
10. J.A. Nolazco Flores and S.J. Young, "Continuous Speech Recognition in Noise using Spectral Subtraction and HMM Adaptation," 1994 IEEE ICASSP, pp. I409-I412.
11. V. Digalakis and L. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods," submitted to 1995 IEEE ICASSP.
12. G. Doddington, "CSR Corpus Development," 1992 DARPA SLS Workshop, pp. 363-366.
13. K.D. Kryter, "The Effects of Noise on Man," 1985 Academic Press.

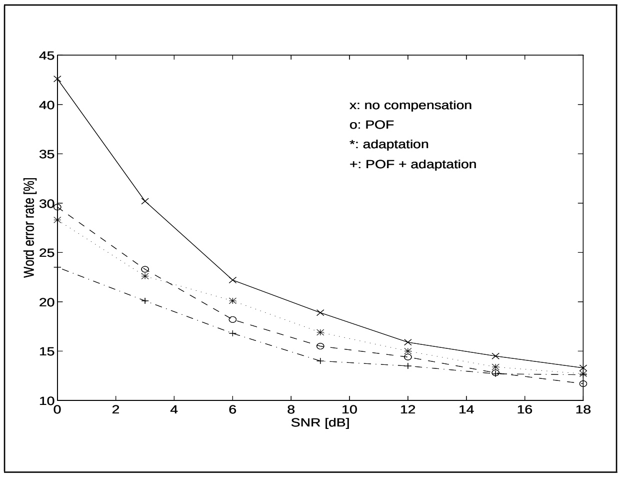

Figure 1: Word error rate vs SNR_wav for various compensation algorithms.

and adaptation at low SNR_wav levels. For SNR_wav levels above 15 dB, the POF-only approach produces the best performance.

All the previous experiments assume prior knowledge of the SNR level of the test data. This is not a serious assumption since we can always estimate the SNR at run-time and select the compensation models trained at a similar SNR. Table 3 shows performance for the combined approach (mapping + adaptation) for the cases in which the testing SNR level may not match the compensation SNR level. This experiment shows that a precise estimate of the SNR is not necessary since performance seems to degrade slowly as the mismatch between the model SNR and the test data SNR increases.

Model SNR_wav [dB]	**Test SNR_wav levels [dB]**							
	0	**3**	**6**	**9**	**12**	**15**	**18**	**inf**
0	23.5	21.4	21.7	24.0	29.2	36.1	47.3	90.6
3	25.1	20.1	17.2	17.2	19.6	22.0	25.7	71.0
6	26.6	20.8	16.8	15.0	15.1	16.5	18.2	45.8
9	28.7	22.0	17.5	14.0	13.2	12.8	14.1	30.1
12	30.7	22.7	18.2	13.9	13.5	12.8	12.6	21.9
15	32.4	23.4	18.7	14.8	13.2	12.7	12.7	17.4
18	36.4	25.1	19.2	16.0	13.6	12.4	12.6	14.7
inf	42.6	30.2	22.2	18.9	15.9	14.5	13.3	11.1

Table 3. Word error rate at various SNR_wav levels. Columns correspond to the test data SNR_wav and rows correspond to the SNR_wav used to compensate the clean models.

In summary, front-end mapping and HMM adaptation can be combined to improve performance in a noisy channel at low SNR_wav levels. These conclusions are applied in the following section.

3.2. ARPA-Sponsored Benchmark Test (Spoke 10)

3.2.1. Development Test Results

This section describes the procedure used for the 1994 ARPA-sponsored CSR evaluation spoke 10 test. The test consisted of WSJ data (5,000-word vocabulary) corrupted with additive noise collected in three different cars. The car noise was recorded in an automobile traveling at 55 m.p.h. with all windows closed and the air-conditioning turned on, with an omnidirectional microphone clipped to the drivers' side sun visor. A one-minute sample of noise, preceding the noise segment added to the speech and scaled to each SNR level, is available for adaptation. Three noisy test sets were created using the same clean utterances and several different noise levels.

The results on the S10 development test set are shown below in Table 4. These experiments used a bigram language model on the male speaker subset (65 sentences) for car #1. The SNR's computed by NIST in the below table use an "A" frequency-weighted filter [13] before computing the SNR. Since car noise contains significant low frequency energies, applying a frequency weighted filter will shift the SNR levels compared to an unweighted SNR computation on the waveform (SNR_wav).

	Experimental Condition			
	1	2	3	4
POF Compensation	disabled	enabled	enabled	enabled
POF Feature		39-D Cep	13-D Cep+C0	13-D Cep+C0
POF # Gaussians		100	300	300
POF # Frames		3	5	5
Training Car Noises		1,2,3	1,2,3	1
Testing Condition (NIST SNR in dB)	**Word Error**	**Word Error**	**Word Error**	**Word Error**
12	80.6	48.9	47.5	43.2
18	53.2	29.8	29.0	26.5
24	29.6	20.7	20.7	18.7
30	19.0	15.9	18.1	15.8
inf	12.8			

Table 4. Word error rates for various conditions on the development test (car 1) set using a bigram language model.

The second line in Table 4 refers to what feature was used by the mapping. The # Gaussians and the # Frames are both parameters of the POF mapping algorithm. The fifth line in Table 4 indicates which car noises the algorithms were trained on: experiments 2 & 3 trained on all 3 car noises (which includes noise from the same car as the development test set), while experiment 4 only trains on a sample of noise collected from the development test set car. The word-error rate's are computed for each condition as a function of the A-weighted SNR.

2.2. Model Adaptation

In the feature-mapping approach clean features are estimated and the HMMs remain unchanged. In model adaptation, however, the opposite occurs: the noisy feature vectors are unchanged and the HMMs are adapted using a sample of the noisy speech data and its orthographic transcription.

Adaptation of the HMMs is implemented using a constrained estimation of the Gaussian mixtures [11]. In this algorithm, we estimate a set of affine transformations that are applied to the Gaussian distributions. The transformations can be either unique for each mixture of Gaussians or shared by different mixtures. The total number of transformations is determined experimentally based on the amount of adaptation data.

As in the mapping approach, the compensation set can be constructed using a variety of speakers and noises. To achieve good performance, however, the characteristics of the noise and the SNR in the adaptation set have to match the test conditions.

2.3. Combination of Mapping and Adaptation

The third approach adapts the HMMs using the mapped feature vectors. In this algorithm, the feature mapping transforms the noisy features to make them look like the clean features. Then, the HMMs are adapted to these mapped noisy features. Finally, at runtime, the POF mapping is applied to the noisy features and these features are recognized with the adapted HMMs.

This approach might be particularly applicable at low SNRs where the mapped features may be significantly distorted, and the adaptation algorithm is not able to compensate the models in the cepstral domain because of the highly nonlinear distortion introduced by the additive noise.

3. EXPERIMENTS

Section 3.1 compares the POF, the HMM adaptation, and the combined approach for various SNR levels. Section 3.2 summarizes the procedure used for the 1994 ARPA-sponsored benchmark tests on noisy channels.

3.1. Comparison of Compensation Techniques

We evaluated the noise compensation algorithms on the large vocabulary Wall Street Journal (WSJ) corpus [12]. The experiments were carried out using SRI's DECIPHER™ speech recognition system [1-3] configured with a six-feature front end: 12 cepstral coefficients, cepstral energy, and their first- and second-order differences. We used genonic HMMs, as described in [1]; for rapid experimentation, we constrained the search using the Progressive Search Technique described in [2]. In the current section (Section 3.1) we used lattices created on the clean test set (before adding the noise) to constraint the recognition search, resulting in optimistic results. In the following section (Section 3.2), we use a full search decoder, resulting in real error rates.

The noisy data were created artificially in the lab by adding the scaled noise to the speech data. Eight minutes of car noise were recorded on a 1985 Honda Civic Station Wagon traveling at a steady speed of 55 m.p.h. with its windows closed. We used the same 8 minute sample of noise for training and testing. To create a noisy sentence (approximately 10 seconds of speech), we selected a continuous block of noise from the long noise recording at random. This block of noise was scaled to achieve a given SNR level and added to the speech data. For these experiments, we computed the SNR on the unfiltered waveform, and designate this as SNR_wav.

Our main goal in this set of experiments was to compare the performance of the three proposed algorithms described in Section 2. However, to have a lower bound in the word error rate under noisy conditions, we also trained the genonic HMM recognizer from scratch using noisy training data at an SNR_wav of 6 dB. Therefore, we have two baseline recognizers, one based on "clean" HMMs and the other with "noisy" HMMs. The training data set consisted of 18,000 WSJ sentences from 170 male speakers. A compensation set was created using a subset of 300 sentences from the training set. The test set consisted of 90 sentences from 4 speakers.

Table 1 compares the performance for these systems. These results show that word error rate degrades from 11.1% for the clean/clean condition to 15.5% for the noisy/noisy condition. These baseline numbers will be used as a reference for the compensation algorithms.

	Test Clean	**Test Noisy**
Train Clean	11.1	22.2
Train Noisy	40.4	15.5

Table 1. Baseline word error rate in percent for clean and noisy conditions. The SNR_wav of the noisy data is 6 dB.

Table 2 compares the performance of the three compensation algorithms described in Section 2 and the baseline results.

Train	**Test**	**Error Rate (%)**
Clean	Clean	11.1
Noisy	Noisy	15.5
Clean	Noisy	22.2
Clean	Noisy+POF	18.2
Clean + Adaptation	Noisy	20.1
Clean + Adaptation	Noisy + POF	16.8

Table 2. Word error for baseline conditions and compensation algorithms. The SNR_wav of the noisy data is 6 dB.

We found that the error rate for mapping is 18.2% and for adaptation is 20.1%. In both cases we optimized each technique to maximize performance. For the combined approach, we found that adapting the HMM's to the mapped features reduced the error rate to 16.8%, only 8.4% ((16.8 - 15.5) / 15.5) worse than the full training in noise condition. Figure 1, which illustrates how the compensation algorithms perform at various SNRs, clearly shows how the combined approach outperforms mapping

ROBUST SPEECH RECOGNITION IN NOISE USING ADAPTATION AND MAPPING TECHNIQUES

Leonardo Neumeyer and Mitchel Weintraub

SRI International
Speech Technology and Research Laboratory
Menlo Park, CA, 94025, USA

ABSTRACT

This paper compares three techniques for recognizing continuous speech in the presence of additive car noise: 1) transforming the noisy acoustic features using a mapping algorithm, 2) adaptation of the Hidden Markov Models (HMMs), and 3) combination of mapping and adaptation. We show that at low signal-to-noise ratio (SNR) levels, compensating in the feature and model domains yields similar performance. We also show that adapting the HMMs with the mapped features produces the best performance. The algorithms were implemented using SRI's DECIPHER™ speech recognition system [1-3] and were tested on the 1994 ARPA-sponsored CSR evaluation test spoke 10 [4].

1. INTRODUCTION

There are several approaches that one can use to recognize speech in the presence of additive background noise. The algorithms that we present here attempt to make each of the major components robust to additive noise: (a) the front-end signal processing and (b) the statistical modeling.

To make the signal processing robust to additive noise, we apply a technique called *Probabilistic Optimum Filtering* (POF) [5]. We have previously showed how this algorithm can be used to recognize narrowband speech recorded over the telephone using wideband HMMs, and how to map speech features obtained from a boom desktop microphone to features generated from a close talking microphone. In summary, our focus in developing POF was the problem of channel mismatches between training and testing conditions.

The class of feature-transformation approaches have been used successfully by other researchers [6,7] to compensate for speech corrupted with additive noise. We extend these techniques by using the POF technique and combine it with the ideas in our earlier noise-robust work [8]. Specifically, we train many different POF filters for different conditions (e.g. different background noise, different SNR levels). At runtime, we automatically select the most appropriate model.

The POF model does not use any assumption about the underlying physical phenomena that corrupted the signal. However, it requires stereo recordings of the clean and noisy speech to estimate its parameters. In the case of additive noise, it is straightforward to build an artificial stereo database when a sample of the noise is available, just by adding the noise to the clean speech.

One approach to make the statistical modeling robust to additive noise is Parallel Model Combination (PMC) [9]. PMC is used to adapt the HMM parameters in a very simple but effective manner and it has also been shown [10] that integrating PMC with a continuous spectral subtraction in the front end is beneficial at low SNRs.

Our approach to robust statistical modeling is to use a model adaptation technique described in [11]. In this case, we apply a set of affine transformations to the Gaussian mixtures of the HMMs. Unlike POF, stereo data are not needed to estimate the adaptation parameters. The clean HMMs are adapted using an orthographically transcribed adaptation set that matches the noisy conditions.

Finally, we investigate how both techniques (mapping and adaptation) perform when they are used together. That is, we enhance the noisy features using POF followed by the adaptation stage. In fact, at low SNRs this technique produces the best performance.

2. COMPENSATION TECHNIQUES

2.1. Feature Mapping

The POF mapping algorithm is designed to estimate a clean feature vector by applying a set of weighted affine transformations to the noisy feature vectors [5]. To estimate the POF transformation parameters, we need a stereo compensation set with simultaneous sequences of the clean and noisy feature vectors. The stereo data is created by adding noise to the clean data to obtain noisy data. The question arises as to what noise to add to the clean speech and how the transformation parameters are affected by the properties of the noise (spectrum and level). Three possible approaches are to (1) add many different types of noise to the training data and train a general mapping that will apply to all types of additive noise, (2) train many different mappings for different noise spectra and SNR's, and select the best model at runtime, and (3) obtain a sample of the actual noise encountered in the field and build a specific mapping for these conditions at runtime.