NOISE-RESISTANT FEATURE EXTRACTION AND MODEL TRAINING FOR ROBUST SPEECH RECOGNITION*

A. Sankar *A. Stolcke* *T. Chung* *L. Neumeyer* *M. Weintraub* *H. Franco* *F. Beaufays*

Speech Technology And Research Laboratory
SRI International
Menlo Park, CA

ABSTRACT

In this paper we report on our recent work on noise-robust feature extraction and model training to alleviate the mismatch caused by different microphones and ambient room noise in the context of the 1995 DARPA-sponsored H3 benchmark test, which used the unlimited-vocabulary North American Business News (NABN) database. We present a novel noise-robust feature extraction algorithm that is a combination of our previously developed minimum mean square error (MMSE) log-energy estimation algorithm and the probabilistic optimum filtering (POF) algorithm. We also studied an approach based on training the automatic speech recognition (ASR) system with previously collected noisy speech. While both the above approaches gave significant improvements, it was found that combining them gave the best results. We also report on a new part-of-speech (POS) language model that makes it possible to train robust POS language models that incorporate longer contexts than is possible with word-based language models. Preliminary results using this approach were encouraging.

1. INTRODUCTION

It is well known that mismatches between the training and testing conditions can severely degrade the performance of automatic speech recognition (ASR) systems. Examples of such mismatches include different noise, microphone, and channel conditions, and different training and testing speakers. In this paper, we present our recent work on noise-robust speech recognition in the context of the 1995 DARPA-sponsored H3 benchmark test, which evaluated the performance of ASR systems under noisy and degraded acoustic conditions. We present a novel noise-robust feature extraction algorithm that is a combination of our previously developed minimum mean square error (MMSE) log-energy estimation algorithm [1] and the probabilistic optimum filtering (POF) algorithm [2]. Our ASR system is based on hidden Markov models (HMM) which were trained using the SI-284 Wall Street Journal (WSJ) Sennheiser clean speech database. Given test speech from a new noisy environment, our noise-robust feature extraction algorithm attempts to estimate clean features so as to bridge the mismatch between the noisy test speech and the clean HMMs. This algorithm was the main noise-robust feature of the SRI system that was evaluated in the 1995 H3 benchmark. This algorithm gave a significant improvement in performance compared to the baseline (nonrobust) Mel-cepstrum-based feature extraction scheme.

After the benchmarks were completed, we studied an approach based on training the HMMs with the alternate microphone data of the WSJ SI-284 database, since this approach was used advantageously by other researchers during the benchmarks. The models trained with this approach are then used to recognize the noisy test speech using the baseline (nonrobust) feature extraction method. The alternate microphones in both the SI-284 database and in the test database generally pick up a higher level of noise than the Sennheiser microphone. If there is a match between the noise observed in the SI-284 training set and the test set, then one would expect this method to work well. Based on our preliminary experiments, we found that for many microphones in the test database, this model-training method performed comparably to the approach used by SRI during the evaluations. However, for one microphone (the B&K sound-level meter), for which the signal to noise ratio (SNR) was significantly lower than the other microphones used for development, the performance of the noise-robust feature extraction scheme was far better than the model-training approach. Finally, it was found that the best results were obtained when the noise-robust feature extraction scheme was combined with the model-training approach.

Acoustic adaptation techniques can also be used to bridge the mismatch between the training and testing acoustic environments, and the differences between individual test speakers and the training population. In previous work, maximum-likelihood and Bayesian adaptation techniques have been developed [3, 4, 5, 6], and have given significant improvements in performance for both nonnative speakers [3] and noisy speech [7]. More recently, we have developed new feature and model-space adaptation techniques and applied them to both nonnative and native speakers [8].

We have also recently developed a novel part-of-speech (POS) tag-based language model. The idea behind this approach is to tag each word by its part-of-speech and then train a statistical language model using these tags as terminal symbols. Since the number of these tags is far less than the number of vocabulary words, it is possible to train robust tag-based language models which incorporate much longer histories (for example, 5- and 6-gram language mod-

*THIS WORK WAS SPONSORED BY DARPA THROUGH NAVAL COMMAND AND CONTROL OCEAN SURVEILLANCE CENTER UNDER CONTRACT N66001-94-C-6048.

els) than is possible with word-based language models. Preliminary experiments using this approach for the 1995 H3 test were encouraging.

The rest of this paper is organized as follows. In Section 2, we briefly describe the 1995 H3 task. In Section 3 we present the new noise-robust feature extraction algorithm. In Section 4 we describe our acoustic adaptation algorithms. In Section 5 we describe the language models we used for the benchmarks. In Section 6 we give the results of the SRI system on the H3 benchmark. In Section 7, we present our initial results by training HMMs using the alternate microphone SI-284 database and compare it to the robust feature-extraction approach used by SRI during the benchmark. Finally we present our conclusions in Section 8

2. THE 1995 H3 TASK

The 1995 H3 task was designed to improve basic speaker-independent (SI) speech recognition performance on unlimited-vocabulary read speech under acoustical conditions that are somewhat more varied and degraded than speech used in previous DARPA evaluations. The evaluation data consisted of 20 speakers each reading 15 North American Business News (NABN) utterances. The utterances were recorded using a close-talking Sennheiser microphone and one other alternate microphone. The alternate microphone was fixed for each speaker but could vary from speaker to speaker. The test comprised two parts: the H3-P0 test measured the performance on the alternate microphone data, and the H3-C0 test measured the performance on the Sennheiser data.

For development, NIST provided data from each speaker recorded through the Sennheiser microphone and simultaneously through seven other microphones. In order to create a development test set from this data for the H3-P0 test, we sampled each of the seven alternate microphones evenly across the 20 speakers to generate about 300 test sentences. For the H3-C0 test data, we used the same utterances recorded through the Sennheiser channel. The acoustic models we used were continuous-density, genonic HMMs [9]. Separate HMMs were trained for males and females using the SI-284 Sennheiser WSJ data. These genonic HMMs have about 1,800 Gaussian mixtures, each with 32 components. For development, we generated a statistical language model using the 60,000 most frequent words in the 1994 NABN text corpus.

In order to facilitate quick experimentation during development, we generated word lattices for the H3-C0 Sennheiser test speech with the genonic HMMs described above using a forward-backward search, a bigram language model, and the word-life algorithm described in [10]. These lattices were then used for experimenting with different algorithms for the noisy H3-P0 data. While this gives overly optimistic results for the H3-P0 data, these results can still be used to qualitatively compare the performance of different algorithms.

3. NOISE-ROBUST FEATURE EXTRACTION

In order to reduce the mismatch caused by ambient room noise, we experimented with methods based on the MMSE log-energy estimation method [1] and the POF algorithm [2], which were previously developed at SRI. We then developed a new method that combined these two approaches.

3.1. MMSE log-energy estimation

Suppose the observed speech y is generated by passing the sum of the original speech x and colored noise n through a microphone channel. If the noise and speech are assumed to be uncorrelated, then we can write

$$P_y(\omega) = P_x(\omega)H_1^2(\omega) + P_{n_w}H_c^2(\omega)H_1^2(\omega), \qquad (1)$$

where H_1 and H_c are the frequency responses of the microphone channel and the noise-coloring filter, respectively, P_y and P_x are the power spectra of y and x, and P_{n_w} is the white-noise spectrum. Estimates of the power spectra are used to compute the log-power L_k in each Mel filter-band, k. This is then followed by an inverse discrete cosine transform (DCT) operation to compute the Mel-cepstrum. The estimates L^y of the log-energy of the noisy speech are distorted because of the noise. The MMSE approach estimates the clean log-energy L^x given the observed log energy L^y by computing the conditional expected value of the clean log-energy given the noisy log-energy [1]. Thus

$$\hat{L^x} = E\left[L^x | L^y\right] \qquad (2)$$

Once the MMSE estimates of the log-energy are computed, we apply the inverse discrete cosine transform to get the Mel cepstrum.

3.2. The POF approach

The POF approach differs from the MMSE approach in that no particular model of signal and noise interaction is assumed. Instead, it is assumed that the original (clean) speech cepstra x_c can be recovered from the noisy speech cepstra y_c by applying a set of linear transformations to the noisy speech cepstra. Each transformation in the set is tied to a separate acoustic region [2]. In order to train the POF filters, it is necessary to use stereo pairs of cepstrum vectors from noisy and clean speech. Ideally, we would like to use stereo pairs where the noisy speech matches that observed in the test environment. Unfortunately, this is not always possible since the test environment changes according to the application. For the research reported in this paper, we created the stereo training pairs by adding white noise to clean WSJ speech. While this is not optimal, we got a significant improvement in performance by using this technique. We trained POF filters at various SNRs from -5 dB to 34 dB at 3 dB intervals. During testing, we computed the SNR for a sentence and used the POF filter with the closest SNR. This approach is similar to our previous work [7]. However, in [7] the noise used to train the POF filters was matched with the test environment noise.

3.3. MMSE/POF combined approach

The above two approaches can be easily combined as follows. First, the noisy speech is processed using the MMSE log-energy estimation algorithm. The cepstrum derived as a result of this approach is then mapped to an estimate of the clean cepstrum using the POF algorithm.

Baseline	MMSE	POF	MMSE+POF
39.7	34.8	31.7	31.0

Table 1. Word Error Rate (percent) Performance

The results of all the above techniques are shown in Table 1. As can be seen from the table, the MMSE and POF approaches individually gave a significant improvement compared to the baseline case. However, the combined approach gave the best results, even though it was only slightly better than the POF approach on this database. However, since these experiments were performed using clean lattices, the difference between the real error rate and the error rate measured with clean lattices tends to be larger as the error rate becomes larger. This is because the constraints imposed by the lattice do not allow as many high-error paths as a full grammar search network. However, for low error rates, the lattice error tends to reflect the true error. This was also verified by our experiments. For this reason, we expect the difference between the error rate for the combined algorithm and the POF algorithm to be larger than what the table shows. We also applied the MMSE+POF algorithm to the Sennheiser test data and found that it slightly improved the performance. Based on this result, we decided to use the MMSE+POF feature extraction scheme for both the alternate microphone data and the Sennheiser data.

4. ACOUSTIC ADAPTATION

Acoustic adaptation can also be used to reduce the mismatch between training and testing acoustic environments. In previous work, maximum-likelihood (ML) and Bayesian adaptation algorithms have been developed [3, 11, 4, 5, 6], and applied to nonnative speech recognition [3] and noisy speech recognition [7]. In particular, maximum-likelihood adaptation has been applied in both the feature-space and model-space [4, 8]. In feature-space schemes, the test speech features are transformed to match the trained models, whereas in model-space schemes, the model parameters are transformed to match the test features.

In our previous ML adaptation algorithms, we used an affine transformation for each feature component [3, 4]. This transform results in a corresponding transformation of the HMM mean and variance vectors. In order to approximate complex functions, a separate transform is used for different Gaussian clusters [3]. The mapping was applied separately to each feature component in order to make the problem mathematically tractable. However, if only the mean vectors are transformed and the variances left untransformed, then a full-matrix affine transformation can be easily estimated [12]. This approach was found to give better results than a component-wise transformation for speaker adaptation because of the modeling of the dependencies between feature components [8].

We have also developed adaptation methods that use transformations of the HMM variances as described in [4, 11, 8]. In one feature-space approach, we assume that the test features are obtained by adding a random bias term to the original speech features [4, 11]. The bias is modeled as a Gaussian random variable with mean μ_b and variance σ_b^2. The speech means and variances are now transformed by adding the mean and variance of the random bias to the HMM means and variances. In a second model-space approach, we have developed a technique for scaling the HMM variances [4, 8]. In this case, each component of the variance vector (in a diagonal covariance matrix) is scaled according to

$$\sigma_y^2 = \alpha \sigma_x^2, \tag{3}$$

where α is a scale factor to be estimated. We have previously reported improvements by using variance transformations for both channel mismatches [11, 4] and speaker adaptation [8]. Because of lack of time, we did not use this approach for the H3 evaluations. However, the variance scaling transformation of Equation 3 was used to advantage in the H3 evaluations by other researchers [13]. We note that the different adaptation techniques described above can be applied in sequence as in [8].

In the H3 test, the speaker session boundaries are assumed to be known, and this information could be used by adaptation algorithms. We derived initial hypotheses for each session by generating N-best lists using the baseline acoustic models and bigram language models. These lists were rescored using trigram language models to generate the hypotheses that were used for acoustic adaptation. The method we used was a block-diagonal [8] matrix affine transformation [12] of the HMM mean vectors. The adapted models were then used to acoustically rescore the N-best lists. We also used adapted crossword models in a similar acoustic rescoring pass, the only difference being that for the crossword models we also used an unseen triphone modeling scheme [14]. In order to adapt the crossword models, we used a context-independent (CI) phone loop as the reference for each sentence. This procedure was used mainly because of the lack of time. However we have previously found that at operating error rates of about 20%, the performance of the CI phone loop approach is comparable to that of an approach using the first-pass hypotheses for adaptation [15]. An advantage of the CI-phone loop approach is that we do not have to do a recognition pass through a large network in order to generate hypotheses before adaptation.

5. LANGUAGE MODELING

The SRI evaluation system made use of four different language models, all based on the 60,000 word vocabulary that had been selected for the official CMU language model. A bigram backoff model (a subset of the trigram model supplied by CMU) was used during decoding. The other three language models were used only as knowledge sources for reordering N-best lists. The rescoring used log probability scores from the language models and combined them by optimized linear weighting (see Section 6). The rescoring models were a standard backoff word trigram language model, a 5-gram part-of-speech model, and a 4th-order POS-based hidden Markov model. We will describe each in turn.

5.1. Word trigram model

The word trigram model used was a standard backoff language model generated using Good-Turing discounting. It was based on the trigram counts supplied by CMU, which

were obtained from 305 million words of NABN sources. We lowered the n-gram cut-offs relative to the official trigram model so as to include all bigrams, and trigrams occurring at least twice, resulting in 16.5 million bigrams and 22.5 million trigrams.

5.2. Part-of-speech 5-gram model

A second knowledge source evaluates hypotheses according to their POS sequences. The rationale behind the POS model is that it can model syntactic dependencies that are outside the scope of the usual word-based N-gram models. Because of the much smaller vocabulary (173 POS labels in our case), a tag-based model of order 5 or higher can be trained and used effectively given the available training data. For POS tagging we use the well-established probabilistic paradigm that models word sequences as outputs from a hidden Markov process whose states are POS N-grams [16].

For a given word sequence $w_1 \dots w_n$ we find the POS tag sequence $t_1 \dots t_n$ with the highest probability $P_{\mathrm{POS}}(t_1 \dots t_n)$, such that each t_i is compatible with the w_i. P_{POS} is given by a 5-gram POS model. To find the possible tags t_i for a word w_i we used a precompiled dictionary where possible. For unknown words, a tree of word suffixes containing POS statistics is consulted, effectively guessing a word's tag based on parts of its morphology. The log probability of the best tag sequence is then used as a knowledge source in rescoring.

The POS 5-gram model was trained on the 1994 CSR NABN corpus. The model is bootstrapped by first training only on POS N-grams that can be unambiguously predicted from an initial dictionary. Following that, the POS model is used to disambiguate tags on the training corpus, and tag sequences thus obtained are used for successive reestimation of the model. The POS model used in the final system contained 1.4 million 4-grams and 2.8 million 5-grams.

5.3. POS-based hidden Markov model

In addition to scoring the POS tags themselves, we can also use the POS model as a language model proper by viewing the POS tags as a hidden state sequence emitting the observed words. We approximate

$$\begin{aligned} P(w_1 \dots w_n) &= \sum_{t_1 \dots t_n} P(w_1 \dots w_n | t_1 \dots t_n) P_{\mathrm{POS}}(t_1 \dots t_n) \\ &\approx \sum t_1 \dots t_n \prod_{i=1}^{n} P(w_i|t_i) P_{\mathrm{POS}}(t_1 \dots t_n) \end{aligned}$$

The probabilities $P(w_i|t_i)$ are computed from statistics obtained in tagging the NABN training corpus. P_{POS} is obtained from the 5-gram POS model described above (the underlying HMM is thus of order 4) Also, the summation above is approximated by considering only the most probable tag sequences $t_1 \dots t_n$ (the 64 best in our implementation).

Although the POS-based language modeling techniques are quite old, they have not been used extensively in large-vocabulary tasks with large amounts of training data. POS-based models typically do not improve on word N-gram models in these tasks since they do not capture local lexical co-occurrence statistics as well (although they can offer good size/performance trade-offs [17]). The three models used here (one word-based, two POS-based) represent complementary aspects of language and should be combined using a principled technique such as maximum entropy [18]. We chose the linear weighting approach as a suboptimal compromise that was less computationally expensive and easy to integrate into our system.

6. H3 BENCHMARK RESULTS

For the 1995 H3 benchmark system, a two-pass recognition approach was used: In the first pass, N-best lists were generated using the baseline genonic HMMs described in Section 2 and the standard 60,000-word bigram language model provided by CMU. The front-end features were the 12-dimensional Mel-cepstrum and normalized energy, along with the corresponding delta and delta-delta features. Since the HMMs are gender-dependent, it is first necessary to determine the gender of the test speaker. Gender identification during recognition is accomplished using a Gaussian mixture model. Separate models were trained for clean speech and noisy speech. The noisy speech model was trained using WSJ data with artificially added noise to give an SNR of 12 dB. For gender selection during recognition, the average SNR for the entire session was computed and compared against a threshold (21 dB). If the SNR was less than the threshold the noisy model was used and otherwise the clean model was used. The 21 dB threshold was determined from SNR histograms of the Sennheiser and alternate microphone data provided in the 1995 development data.

In the second pass, the N-best lists were rescored with six different knowledge sources:

1. Baseline noncrossword genonic HMMs with acoustic adaptation
2. Crossword genonic HMMs with acoustic adaptation
3. The number of words in the N-best hypotheses
4. Trigram language models
5. A 5-gram POS language model
6. A 5th-order HMM using POS tags as the underlying states and words as the observables

The scores from these knowledge sources were linearly combined, with the weights for each knowledge source being computed so as to minimize the error rate on the development test set. This was done using Powell's algorithm [19].

Table 2 gives the results of evaluating our system on the 1995 H3 development and evaluation data. For the evaluation data, we list both the pre-adjudicated error rate as computed at SRI, and the final adjudicated error rate computed by NIST. The first column of the table gives the 1-best error rate of the first-pass N-best generation step. This pass used only noncrossword HMMs and the standard bigram language model. We did not optimize the weight of the language model score in relation to the acoustic model score. The second column gives the result of rescoring the N-best list with the trigram language model and the baseline noncrossword acoustic models. We used the same relative weight between acoustic and language models for the

Non-crossword HMMs (NoCW)	Yes	Yes	Yes		
Cross-word HMMs (CW)			Yes		
Adapted NoCW				Yes	Yes
Adapted CW				Yes	Yes
Bigram	Yes				
Trigram		Yes	Yes	Yes	Yes
Number of words			Yes	Yes	Yes
POS-tag lang. model					Yes
Development test set error rate					
H3-P0	27.7	22.8	21.7	20.7	20.0
H3-C0	16.4	11.9	10.8	10.6	9.9
Evaluation test set error rate					
H3-P0 (adjudicated score)	34.4	28.5	27.8	25.6	25.4 24.6
H3-C0 (adjudicated score)	17.4	12.8	11.8	10.9	10.7 9.7

Table 2. Results on the 1995 H3 benchmark

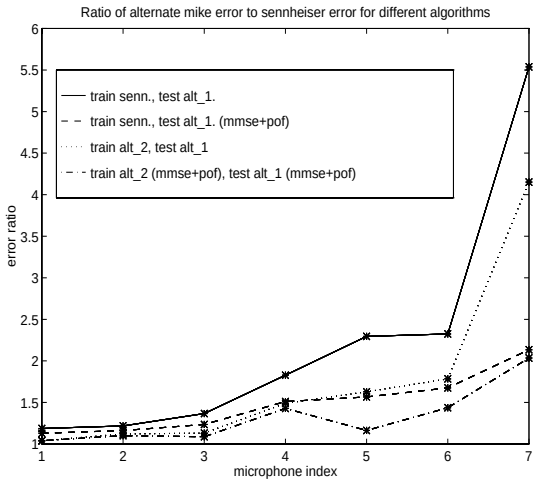

Figure 1. Ratio of alternate microphone error to Sennheiser error for different microphones

first two columns. In the third column, we included cross-word acoustic models and the number of words in the N-best hypothesis. The weights of the different knowledge sources were optimized using Powell's algorithm [19]. The fourth column is identical to the third except that adapted acoustic models are used. Finally, in the fifth column, we added the POS tag-based language model knowledge sources.

7. EXPERIMENTS WITH MODEL TRAINING

After the conclusion of the benchmarks, we studied a method for improving the recognition performance for noisy speech based on training the HMMs using the alternate microphone SI-284 database. Such an approach was used to advantage by some sites during the evaluations (for example see [13]).

We would expect this approach to work well if there is a match between the noise during training and testing. However, if there is a significant departure from the training conditions, then the performance will not be good. We found that for many microphones, the performance of this method was similar to that of the MMSE+POF feature extraction algorithm we used during the evaluations. However, for the B&K sound-level-meter microphone, the approach of training HMMs on the SI-284 alternate microphone database was significantly worse than the MMSE+POF approach. This shows that the method of training on noisy data could fail when there is a significant mismatch between the noise in the training and testing conditions. We then used the models trained using the alternate microphone data, but performed the MMSE+POF robust feature extraction during recognition. The performance of this method was better compared to the previous case and it was also better than the method used by SRI during the evaluations. Finally, we used the MMSE+POF algorithm to process the alternate microphone data prior to training the models, and also used the MMSE+POF scheme during recognition. This approach was found to work the best.

Figure 1 plots the ratio of the error rate for the alternate microphone to the error rate for the Sennheiser microphone for each alternate microphone in the male subset of the H3 development set. We would like this ratio to be close to 1. For severely noisy conditions, this ratio will be large. The figure shows four different plots. The first plot (train Senn., test alt_1) is the baseline case of training on clean data and testing on the alternate microphone data. The second curve (train Senn., test alt_1 (mmse+pof)) reflects the performance of the scheme used by SRI for the evaluations. The models are trained on Sennheiser speech, but during testing the MMSE+POF approach is used. It can be seen that a significant improvement in performance is obtained by using this method. The third plot (train alt_2, test alt_1) shows the performance when the models are trained on the alternate microphone data and the standard (nonrobust) feature extraction scheme is used during training and testing. The figure shows that while this method worked well for most of the microphones, it performed much worse than the MMSE+POF approach for one microphone (the B&K sound-level-meter microphone). Finally, the last curve (train alt_2 (mmse+pof), test alt_1 (mmse+pof)) shows the case of training on the alternate microphone data but after processing with MMSE+POF, and testing on the MMSE+POF processed data. This was seen to give the best performance.

Based on the experiment, we decided to regenerate N-best lists for the evaluation data using the method of training on alternate microphone data processed using MMSE+POF. These lists were then rescored with trigram language models. The results of this experiment are shown in Table 3. From the table we see that both the 1-best error rate from the initial bigram N-best generation pass, and the trigram rescored error are significantly lower than those achieved during the evaluations. We did not rescore the lists with all the other knowledge sources. However, because an additional 3% absolute reduction in the error rate was achieved by using these knowledge sources for the evaluation system

	1-Best	Trigram Rescore
Evaluation System	34.4	28.1
Train alt_2 (mmse+pof), Test alt_1 (mmse+pof)	28.5	22.3

Table 3. Error-rates for SRI evaluation system and model training with alternate microphone data

(see Table 2), we expect the final error rate to drop from the 22.3% in Table 3 to about 19%.

8. CONCLUSION

In this paper we have presented the results of our research on noise-robust recognition, acoustic adaptation, and a new POS language modeling technique. It was found that our previously developed MMSE log-energy estimation techniques and the POF technique were able to significantly lower the error rate for HMMs trained using clean Sennheiser speech and tested on noisy speech collected from alternate microphones. However, the combined MMSE+POF method performed better than either method alone. It was found that acoustic adaptation gave an additional improvement in performance for both the clean and alternate microphone speech. A new POS language model was presented, which can be used to train robust language models that incorporate longer contextual histories than is possible with word-based language models. Our results with this approach were encouraging. Finally, we experimented with an approach for noise-robust recognition based on training the HMMs on alternate microphone data. While it is well known that this approach will perform well if the training noise and testing noise are matched, we found that when there is a significant mismatch, this method has a poor performance (worse than the MMSE+POF approach). However, if combined with the MMSE+POF method, the method of model training with alternate microphone noisy data gave the best results.

REFERENCES

[1] A. Erell and M. Weintraub, "Filterbank-Energy Estimation Using Mixture and Markov Models for Recognition of Noisy Speech," *IEEE-TSAP*, vol. 1, pp. 68–76, January 1993.

[2] L. Neumeyer and M. Weintraub, "Probabilistic Optimum Filtering for Robust Speech Recognition," in *ICASSP*, pp. I–417–I–420, 1994.

[3] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained reestimation of Gaussian mixtures," *IEEE-TSAP*, vol. 3, no. 5, pp. 357–366, 1995.

[4] A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE-TSAP*, May 1996, to appear.

[5] J. Gauvain and C.-H. Lee, "Maximum *a posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE-TSAP*, vol. 2, pp. 291–298, April 1994.

[6] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using combined transformation and bayesian methods," 1994, accepted for publication in *IEEE-TSAP*.

[7] L. Neumeyer and M. Weintraub, "Robust Speech Recognition in Noise using Mapping and Adaptation Techniques," in *ICASSP*, pp. 141–144, 1995.

[8] L. Neumeyer, A. Sankar, and V. Digalakis, "A Comparative Study of Speaker Adaptation Techniques," in *EUROSPEECH*, pp. 1127–1130, 1995.

[9] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized mixture tying in continuous hidden markov model-based speech recognizers," 1994, accepted for publication in *IEEE-TSAP*.

[10] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER(TM) Speech Recognition System: Progressive-Search Techniques," in *ICASSP*, pp. II–319–II–322, 1993.

[11] A. Sankar and C.-H. Lee, "Stochastic Matching for Robust Speech Recognition," *IEEE Sig. Proc. Letts.*, vol. 1, pp. 124–125, August 1994.

[12] C. J. Legetter and P. C. Woodland, "Flexible Speaker Adaptation using Maximum Likelihood Linear Regression," in *Proc. ARPA-SLS Workshop*, pp. 110–115, 1995.

[13] P. C. Woodland, M. J. F. Gales, D. Pye, and V. Valtchev, "The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task," elsewhere in these proceedings.

[14] V. Digalakis, M. Weintraub, A. Sankar, H. Franco, L. Neumeyer, and H. Murveit, "Continuous Speech Dictation on ARPA's North American Business News Domain," in *Proc. ARPA-SLS Workshop*, pp. 88–93, 1995.

[15] A. Sankar, L. Neumeyer, and M. Weintraub, "An Experimental Study of Acoustic Adaptation Algorithms," in *ICASSP*, 1996.

[16] K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted text," in *Second Conference on App. Nat. Lang. Proc.*, pp. 136–143, 1988.

[17] T. R. Niesler and P. C. Woodland, "A variable-length category-based N-gram language model," in *ICASSP*, 1996.

[18] R. Rosenfeld, *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. PhD thesis, School of Comp. Sc., Carnegie Mellon University, Pittsburgh, PA, 1994. Technical Report CMU-CS-94-138.

[19] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipies in C: The Art of Scientific Computing*. Cambridge University Press, 1988.