ADAPTATION TO NEW MICROPHONES USING TIED-MIXTURE NORMALIZATION

Anastasios Anastasakos†, Francis Kubala, John Makhoul, Richard Schwartz

BBN Systems and Technologies
Cambridge MA 02138
†Northeastern University
Boston MA 02115

ABSTRACT

In this paper, we present several approaches designed to increase the robustness of BYBLOS, the BBN continuous speech recognition system. We address the problem of increased degradation in performance when there is mismatch in the characteristics of the training and the test microphones. We introduce a new supervised adaptation algorithm that computes a transformation from the training microphone codebook to that of a new microphone, given some information about the new microphone. Results are reported for the development and evaluation test sets of the 1993 ARPA CSR Spoke 6 WSJ task, which consist of speech recorded with two alternate microphones, a stand-mount and a telephone microphone. The proposed algorithm improves the performance of the system when tested with the stand-mount microphone by reducing the difference in error rate between the high quality training microphone and the alternate stand-mount microphone recordings by a factor of 2. Several results are presented for the telephone speech leading to important conclusions: a) the performance on telephone speech is dramatically improved by simply retraining the system on the high-quality training data after they have been bandlimited in the telephone bandwith; and b) additional training data recorded with the high quality microphone give further substantial improvement in performance.

1. INTRODUCTION

Interactive speech recognition systems are usually trained on substantial amounts of speech data collected with a high quality close-talking microphone. During recognition, these systems require the same type of microphone to be used in order to achieve their standard accuracy. This is a highly restricting condition for practical applications of speech recognition systems. One can imagine a situation, where it would be desirable to use a different microphone for recognition than the one with which the training speech was collected. For example, some users may not want to wear a head-mounted microphone. Others may not want to pay for a high quality microphone. Additionally, many applications involve recognition of speech over telephone lines and telephone sets with high variability in quality and characteristics. However, we know that even highly accurate speech recognition systems perform very poorly when they are tested with microphones with different characteristics than the ones that they were trained on [1].

There is a wide range of approaches in order to compensate for this degradation in performance including:

- Retrain the HMMs with data collected with the new microphone encountered during the recognition stage, a rather expensive approach for real applications, or by training on a large number of microphones in the hope that the system will obtain the necessary robustness.
- Use robust signal processing algorithms.
- Develop a feature transformation that maps the alternate microphone data to training microphone data.
- Use statistical methods in order to adapt the parameters of the acoustic models.

In previous work we had discussed the use of Cepstrum Mean Subtraction and the RASTA algorithm as two simple signal processing algorithms to compensate the degradation caused by an alternate channel [7]. In this paper, we present an approach towards feature mapping by modeling the difference between the test and the training microphone, prior to recognition.

We have developed the Tied-Mixture Normalization Algorithm, a technique for adaptation to a new microphone based on modifying the continuous densities in a tied-mixture HMM system, using a relatively small amount of stereo training speech. This method is presented in detail in Section 2. In Section 3 we describe several experiments on a known microphone task and the effect of the adaptation method in the performance of the recognition system.

2. TIED MIXTURE NORMALIZATION

In a *Tied-Mixture Hidden Markov Model (TM-HMM)* system [2, 6], speech is represented using an ensemble of Gaussian mixture densities. Every frame of speech is represented as a Gaussian mixture model. Specifically the probability density function for an observation conditioned on the HMM state is expressed as:

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1994		**00-00-1994 to 00-00-1994**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Adaptation to New Microphones Using Tied-Mixture Normalization	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **BBN Technologies,10 Moulton Street,Cambridge,MA,02238**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**5**	

$$b_{s_t} = Pr(x_t|s_t) = \sum_{k=1}^{C} c_k N(x_t; \mu_k, \Sigma_k)$$

where $x_t, s_t, C, c_k, \mu_k, \Sigma_k$ are the observed speech frame at time t, the HMM state at time t, the number of clusters of the codebook, and for k-th mixture density, the mixture weight, the mean and the covariance matrix respectively.

The vector quantization (VQ) codebook which consists of these mean vectors and covariance matrices, has been derived from a subset of the training data, therefore it is mostly characteristic of the location and distribution of the training data and the training microphone in the acoustic space. However if the codebook was created with data collected with some other microphone, due to the additive and convolutional effect on speech specific to this new microphone, the data would be distributed differently in the acoustic space and the ensemble of means and covariances of the codebook would reflect the characteristics of the new microphone. This is the case of the mismatch in training and testing microphone. Without any compensation, we quantize the test data, recorded with the new microphone, using the mixture codebook generated from recordings with the training microphone. This inevitably results in a degradation in performance, since the codebook does not model the test data.

We introduce a new algorithm, called *Tied Mixture Normalization (TMN)* to compute the codebook transformation from the training microphone to the new test microphone. The TMN algorithm requires a relatively small amount of stereo speech adaptation data, recorded with the microphone used for training (primary microphone) and the new microphone (alternate microphone). Then using the stereo data, we can adapt the existing HMM model to work well on the new test condition despite the mismatch with the training.

Figure 1 provides a schematic description of the TMN algorithm. We assume that we have a tied-mixture densities codebook (set of Gaussians distributions), derived from a subset of the training data that was recorded with the primary microphone. We quantize the adaptation data from the primary channel and label each frame of speech with the index of the most likely Gaussian distribution in the tied-mixture codebook. Since there is an one-to-one correspondence between data of the primary and alternate channel we use the VQ indices of the frames of the data of the primary channel to label the corresponding frames of the data of the alternate channel. Then for each of the VQ clusters, from all the frames of the alternate microphone data with the same VQ label, we compute the sample mean and the sample covariance of the cepstrum vectors that represent a possible shift and scaling of this cluster in the acoustic

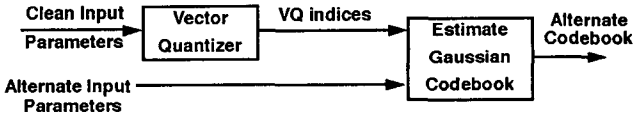

Figure 1: Estimation of alternate microphone Gaussian mixture densities codebook

space (Fig. 2). These are the new means and covariances of the Gaussian distributions of the new normalized codebook.

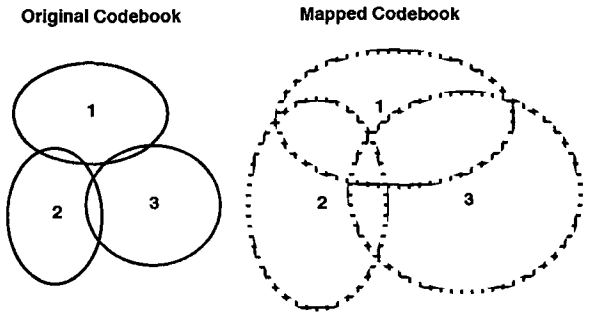

Figure 2: The mapped Gaussian codebook is a shifted and scaled version of the original codebook

The new Gaussian densities are used in conjunction with the mixture weights c_k (sometimes called the discrete probabilities) of the original model to compute the observation probability density function as expressed previously.

One of the possible weaknesses of the TMN algorithm is that each cluster of the original codebook is transformed independently of all the others. This assumption goes against our intuition that a codebook transformation, due to different microphone characteristics, should maintain continuity between adjacent codebook clusters and shift all the clusters in the same general direction. Additionally, a potential problem may arise when a particular cluster does not have enough samples to compute its statistics. Hence, we may not estimate the correct transformation due to insufficient or distorted data by modeling each codebook cluster independently. To alleviate this problem we use the following approach, originally suggested for speaker adaptation [4]: when the centroid of the ith codebook cluster is denoted by m_i and that of the transformed alternate microphone by μ_i,

the deviation vector between these two centroids is

$$d_i = \mu_i - m_i \quad i = 1, 2, ..., C \tag{1}$$

where C is the size of the codebook. For each cluster centroid c_i, the deviation vectors of all clusters $\{d_i\}$ are summed with weighting factors $\{w_{ik}\}$ to produce the shift vector Δ_i:

$$\Delta_i = (\sum_{k=1}^{C} w_{ik} d_i)/(\sum_{k=1}^{C} w_{ik}) \tag{2}$$

The weighting factor w_{ik} is the probability $\{P(m_k|m_i)\}^\alpha$ of centroid m_k of the original codebook to belong to the ith cluster raised to the α power. This weight is a measure of vicinity among clusters and the exponentiation controls the amount of smoothing between the clusters. Finally, the centroid c_i' of the ith cluster of the transformed codebook is:

$$c_i' = c_i + \Delta_i \tag{3}$$

Similarly the covariances of the clusters of the new codebook are computed as the averaged summations over all sample covariances computed in the first implementation of TMN.

3. DESCRIPTION OF EXPERIMENTS

In this section we describe the results we obtained applying the TMN algorithm for the Spoke 6 of the *Wall Street Journal* (WSJ) speech corpus. This is the known alternate microphone 5000-word closed recognition vocabulary, speaker independent speech recognition task. It addresses two different alternate microphones, the Audio-Technica 853a, a high quality directional, stand-mount microphone, and a standard telephone handset (the AT&T 720 speaker phone). The adaptation and test database includes simultaneous recordings of high quality speech using the primary microphone (Sennheiser HMD-414 head-mounted microphone with noise canceling element) and speech recorded with each of the two alternate microphones.

All of the experiments that will be described were performed using the BBN BYBLOS speech recognition system [3]. The front end of the system computes steady-state, first- and second-order derivative Mel-frequency cepstral coefficients (MFCC) and energy features over an analysis range of 80 to 6000 Hz. Cepstrum mean subtraction is a standard feature of the system used to compensate for the unknown channel transfer function. In cepstrum mean subtraction we compute the sample mean of the cepstrum vector over the utterance, and then subtract this mean from the cepstrum vector at each frame. No distinction is made between speech and non-speech frames. The acoustic models are trained on 62 hours of speech (37000 sentences) from the WSJ0 and WSJ1 corpora, collected from 37 speakers, with the Sennheiser high quality close-talking microphone. The recognition is done using trigram language models. The test data comes from the development and evaluation sets of Spoke 6 of the WSJ1 corpus and consists of stereo recordings with the Sennheiser microphone and the Audio-Technica microphone or a telephone handset over external telephone lines. Adaptation data was supplied separately consisting of a total of 800 stereo recorded utterances from 10 speakers; 400 sentences recorded simultaneously with the Sennheiser and the Audio-Technica and 400 sentences recorded with the Sennheiser and the telephone handset.

We evaluated the TMN algorithm for each of the two new microphones and we present the results on the development and the 1993 ARPA WSJ official evaluation test sets.

3.1. Audio-Technica (AT) Microphone

We applied the TMN algorithm, as described in Section 2, on the 400 adaptation sentences simultaneously recorded with the Sennheiser and the Audio-Technica (AT) microphones to compute the codebook transformation for the alternate microphone. For the evaluation of the system, the comparative experiments include:

- Recognition on the Sennheiser recorded portion of the test data to access the lower bound on the error rate, that the baseline system can achieve with matched training and testing microphone.
- Recognition on the Audio-Technica recorded portion of the test data to access the degradation in the performance of the baseline system for the mismatch condition when no adaptation is used, other than the standard cepstrum mean subtraction.
- Recognition on the Audio-Technica recorded portion of the test data, using the proposed adaptation scheme to determine the improvement on the system performance due to the adaptation algorithm.

In Table 1, we list the word error rates for these experiments. The mismatch between the Audio-Technica and the

System Configuration	Dev. Test	Eval. Test
Sennheiser	8.3%	7.9%
AT with no adaptation	10.5%	10.6%
AT with TMN adaptation	9.0%	9.6%

Table 1: Comparison of word error rate (%) for microphone adaptation using the Sennheiser or the Audio-Technica microphone

Sennheiser microphone does not cause a serious degradation, even when no adaptation is used to account for the

channel mismatch. The TMN adaptation reduces the additional degradation due to the channel mismatch by about a factor of 2 in both test sets.

3.2. Telephone Speech

The telephone handset (TH) differs radically from the other two microphones, having the main characteristic of allowing a much narrower band of frequencies than the others. Therefore, prior to applying any adaptation scheme, we chose to bandlimit the Sennheiser training data between 300-3300 Hz, to create new bandlimited phonetic word models. This was accomplished by retaining the DFT coefficients of the feature analysis in the range 300-3300 Hz to compute the MFCC coefficients. We bandlimited the stereo adaptation and test data in the same way. We applied the TMN algorithm on the bandlimited adaptation data to compute the codebook transformation for the telephone speech. During testing, the data is bandlimited as described, and quantized using the normalized telephone codebook. In evaluating the adaptation algorithm for the telephone speech we performed the same series of experiments as with the Audio-Technica microphone. We consider using full bandwidth phonetic models as the baseline system and the generation of bandlimited phonetic models as part of the scheme for adaptation to the telephone speech. In Table 2 we list the word error rates for these experiments. The degradation in performance due to

System Configuration	Dev. test	Eval. test
Sennheiser	8.9%	8.7%
TH with no adaptation	–	29.5%
TH with Bandlimiting and TMN	12.7%	12.8%

Table 2: Comparison of word error rate (%) for microphone adaptation using the Sennheiser or the Telephone handset microphone

the mismatch between the Sennheiser recorded speech and the telephone speech is severe (the error rate goes from 8.9% to 29.5%). The combined effect of bandlimiting the data and the TMN adaptation reduces the error rate by a factor of 2.3 bringing the error rate of recognition of telephone speech close to that of high quality microphone recordings.

Since the telephone speech is radically different from speech collected with the primary microphone, we conducted some more experiments to access the contribution of the bandlimiting process, the adaptation algorithm and the amount of training separately in the performance of the system. Specifically we tested the following conditions:

- *Amount of training data:* All training data is collected with the primary microphone and comprise the WSJ0 and WSJ1 corpora with 12 and 50 hours of recorded speech respectively. We trained two sets of phonetic models using the WSJ0 corpus and the combined WSJ0+WSJ1 training data to determine the impact of additional training data collected with the primary microphone.
- *Bandlimited phonetic models:* Determine the effect of bandlimiting separately from and in combination with the TMN algorithm.
- *TMN Adaptation:* Determine the effect the TMN algorithm separately from and in combination with of bandlimiting.

The results are shown in Tables 3 and Tables 4. We have no clear explanation for the surprising result that additional training speech recorded with a high quality microphone improves the performance of the system on telephone speech. However the error rate reduces by a factor of 2 for some conditions by adding 50 hours of training high quality recorded speech. Furthermore bandlimiting is essential for the good performance of the system for telephone speech, as in all conditions reduces the error rate by a factor of 2. As a contrast, we also computed the error rate of the WSJ0+WSJ1 bandlimited system on the bandlimited Sennheiser recorded data portion of the test and found that to be 11.0%. The latter result compared with 8.9% (Table 2) which is the error rate of the full bandwidth system on the same speech implies that most of the loss in performance between recognizing high-quality Sennheiser recordings and telephone speech is due to the loss of information outside the telephone bandwidth. Using the telephone bandwith, switching from the high-quality Sennheiser microphone to the telephone handset increases the error rate only by a small factor, from 11.0% to 13.9%. Finally the effect of the TMN algorithm is much more significant when telephone bandwidth is not used.

WSJ0-12 hours	Without TMN	With TMN
No bandlimiting	41.8%	36.3%
With bandlimiting	26.8%	24.0%

Table 3: Comparative experiments using 12 hours of training speech recorded with the primary microphone tested on WSJ Spoke 6 development test set telephone recordings.

4. CONCLUSIONS

We have presented a supervised adaptation algorithm that improves the recognition accuracy of the BYBLOS speech recognition system when there is a microphone mismatch between training and testing conditions.

WSJ0+WSJ1-62 hours	Without TMN	With TMN
No bandlimiting	31.8%	22.9%
With bandlimiting	13.9%	12.7%

Table 4: Comparative experiments using 62 hours of training speech recorded with the primary microphone tested on WSJ Spoke 6 development test set telephone recordings.

We tested the algorithm on two different alternate microphones, a high-quality stand-mount microphone and a telephone handset. TMN adaptation reduces the degradation due to mismatch between the Sennheiser and the Audio-Technica microphone by a factor of 2. The results on the telephone handset were more dramatic as the error rate reduced from 29.3% to 12.5% using bandlimited phonetic models and TMN adaptation. We showed that bandlimited phonetic models are essential, as most of the degradation is due to the loss of information outside the narrow bandwidth of the telephone. The 12.5% word error rate is close to the error rate achieved using the primary microphone, which is considered the best performance the system can achieve for a microphone. However the overall good performance of the system of telephone speech may also be an artifact of the data collection procedure, as the speech was only sent over a local loop, there was no long distance calling for example, and the telephone handset did not vary, as the case would be in a conventional application.

5. ACKNOWLEDGMENT

This work was supported by the Defense Advanced Research Projects Agency and monitored by the Office of Naval Research under Contract Nos. N00014-91-C-0115, and N00014-92-C-0035.

References

1. A. Acero, and R.M. Stern, "Environmental Robustness in Automatic Speech Recognition", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, April 1987, pp. 849-852.
2. J. Bellegard, D. Nahamoo, "Tied Mixture Continuous Parameter Modeling for Speech Recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Dec. 1990, vol. 38, No. 12.
3. Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J. Makhoul, P. J. Price, S. Roucos, and R. M. Schwartz, "BYBLOS: The BBN Continuous Speech Recognition System", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, April 1987, pp. 89-93.
4. S. Furui, "Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, May 1989, pp. 286-289.
5. H. Hermansky, and N. Morgan, "Towards Handling the Acoustic Environment in Spoken Language Processing", *Proc. International Conference in Spoken Language Processing*, 1992, pp. 85-88.
6. X. Huang, K. Lee H. Hon, "On Semi-Continuous Hidden Markov Modeling", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, April 1990, paper S13.3.
7. R. Schwartz, Anastasakos T., F. Kubala, J. Makhoul, L. Nguyen, and G. Zavaliagkos, "Comparative Experiments on Large Vocabulary Speech Recognition", *Proc. ARPA Human Language Technology Workshop*, March 1993.