ROBUST SPEECH RECOGNITION BASED ON STOCHASTIC MATCHING

*Ananth Sankar** *Chin-Hui Lee*
AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

ABSTRACT

We present a maximum likelihood (ML) stochastic matching approach to decrease the acoustic mismatch between a test utterance Y and a given set of speech hidden Markov models Λ_X so as to reduce the recognition performance degradation caused by possible distortions in the test utterance. This mismatch may be reduced in two ways: 1) by an inverse distortion function $F_\nu(.)$ that maps Y into an utterance X which matches better with the models Λ_X, and 2) by a model transformation function $G_\eta(.)$ that maps Λ_X to the transformed model Λ_Y which matches better with the utterance Y. The functional form of the transformations depends upon our prior knowledge about the mismatch, and the parameters are estimated along with the recognized string in a maximum likelihood manner using the EM algorithm. Experimental results verify the efficacy of the approach in improving the performance of a continuous speech recognition system in the presence of mismatch due to different transducers and transmission channels.

1. INTRODUCTION

While progress in automatic speech recognition (ASR) has been encouraging, it has become increasingly clear that ASR systems must be robust to changing speaking environments and speaking styles in order to maintain a reasonable performance across a wide range of variable acoustic conditions (e.g. [1, 2]). ASR systems trained in one environment often perform poorly in new environments due to mismatches between the training and testing conditions. These mismatches could be due to different transducers and transmission channels, changing speaking styles and accents, the presence of varying ambient and channel noise, or modeling and estimation errors caused by incomplete characterization of the speech signal and insufficient training data.

In this paper, we present a maximum likelihood (ML) approach to stochastic matching for robust speech recognition. The speech features are assumed to be modeled by a set of subword hidden Markov models (HMM) Λ_X. Due to the possible mismatch between the test utterance Y and the models Λ_X, there is often a degradation in recognition performance. The mismatch may be reduced in two ways. First, we may map the distorted features, Y, to an estimate of the original features, $X = F_\nu(Y)$, so that the given models Λ_X can be used for recognition. Secondly, we can map the given models, Λ_X, to the transformed models, $\Lambda_Y = G_\eta(\Lambda_X)$, which better match the observed utterance Y. The first mapping operates in the *feature space*, whereas the second operates in the *model space*. We present an approach in which the functional form of these mappings is assumed. The unknown parameters, ν or η, are iteratively estimated, using the expectation-maximization (EM) algorithm [3], so as to maximize the likelihood of the observed speech Y given the models Λ_X, thus decreasing the mismatch due to the distortion. The estimation of ν or η requires only the test utterance Y, and the models, Λ_X, and *does not make use of any training data.*

In other related work on speaker adaptation [4, 5], a fixed bias is estimated that transforms each individual speaker to a reference speaker and then the estimated bias is subtracted from every frame of speech. A similar approach has been used for estimating channel mismatch in speech recognition [2, 6] where the speech is modeled by a vector quantization (VQ) codebook. More recently, an approach for speaker adaptation has been presented where the mismatch is modeled by an affine transformation, and the parameters of the transformation are estimated separately for different clusters of Gaussian densities [7, 8]. In all the above approaches, the mismatch is treated as a deterministic feature transformation, whereas in the stochastic matching algorithm, we view the mismatch in both the feature and model spaces. As mentioned above, the method makes use only of the test data and the stored HMMs. This is in contrast to some approaches that make use of a stereo database from the mismatched environments before processing as in [2, 9].

2. GENERAL FRAMEWORK

We are interested in the following problem. Given a set of trained HMMs Λ_X, where the subscript X denotes the fact that the models are based on a given set of training data $\{X\}$, and a test utterance $Y = \{y_1, y_2, \cdots, y_T\}$, we want to recognize the sequence of words $W = \{W_1, W_2, \cdots, W_L\}$ embedded in Y. If there exists a mismatch between the training data $\{X\}$, and the test utterance Y, then this results in errors in the recognized word sequence W. We are interested in reducing this mismatch and hence improving the recognition performance.

The mismatch may be viewed either in the feature-space or in the model space [11, 12]. In the feature-space, let the distortion function map the original utter-

*Ananth Sankar is now with Speech Technology and Research Laboratory, SRI International, Menlo Park, CA.

0-7803-2431-5/95 $4.00

ance $X = \{x_1, x_2, \cdots, x_T\}$ into the sequence of observations $Y = \{y_1, y_2, \cdots, y_T\}$. If this distortion is invertible, then we may map Y back to the original speech X with an inverse function F_ν, so that

$$X = F_\nu(Y), \tag{1}$$

where ν are the parameters of the inverse distortion function. Alternately, in the model-space consider the transformation G_η with parameters η that maps Λ_X into the transformed models Λ_Y so that

$$\Lambda_Y = G_\eta(\Lambda_X). \tag{2}$$

One approach to decreasing the mismatch between Y and Λ_X is to find the parameters ν or η, and the word sequence W that maximize the joint likelihood $p(Y, W|\Lambda_X,)$. Thus, in the feature-space, we need to find ν' such that

$$(\nu', W') = \underset{(\nu, W)}{\text{argmax}}\, p(Y, W|\nu, \Lambda_X), \tag{3}$$

and in the model space we need to find η' such that

$$(\eta', W') = \underset{(\eta, W)}{\text{argmax}}\, p(Y, W|\eta, \Lambda_X). \tag{4}$$

This joint maximization over the variables ν and W in Equation 3 or over η and W in Equation 4 may be done iteratively by keeping ν or η fixed and maximizing over W, and then keeping W fixed and maximizing over ν or η. The process of finding W is just the usual continuous speech decoding problem and has been studied by many researchers. In this paper, we are interested in the problem of finding the parameters ν and η. To simplify expressions, we remove the dependence on W, and write the maximum likelihood estimation problem corresponding to Equations 3 and 4 as

$$\nu' = \underset{\nu}{\text{argmax}}\, p(Y|\nu, \Lambda_X), \tag{5}$$

and

$$\eta' = \underset{\eta}{\text{argmax}}\, p(Y|\eta, \Lambda_X). \tag{6}$$

For this study, we assume that Λ_X is a set of left to right continuous density subword HMMs [13]. The observation density $p_x(x|i)$ for state i is assumed to be a mixture of Gaussians, given by

$$p_x(x|i) = \sum_{j=1}^{M} w_{i,j} N(x; \mu_{i,j}, C_{i,j}), \tag{7}$$

where M is the number of mixtures, $w_{i,j}$ is the probability of mixture j in state i, and N is the normal distribution. $C_{i,j}$ and $\mu_{i,j}$ are the covariance matrix and mean vector corresponding to mixture j in state i.

Let $S = \{s_1, s_2, \cdots, s_T\}$ be the set of all possible state sequences for the set of models Λ_X and $C = \{c_1, c_2, \cdots, c_T\}$ be the set of all mixture component sequences. Then Equation 5 can be written as

$$\nu' = \underset{\nu}{\text{argmax}} \sum_S \sum_C p(Y, S, C|\nu, \Lambda_X). \tag{8}$$

Similarly, we may write Equation 6 as

$$\eta' = \underset{\eta}{\text{argmax}} \sum_S \sum_C p(Y, S, C|\eta, \Lambda_X). \tag{9}$$

In general, it is not easy to estimate ν' or η' directly. However, for some F_ν and G_η, we can use the EM algorithm [3] to iteratively improve on a current estimate and obtain a new estimate such that the likelihoods in Equations 8 and 9 increase at each iteration. In the next two sections, we discuss the application of the EM algorithm to find the estimates of the parameters ν of the feature-space transformation F_ν, and the parameters η of the model-space transformation G_η respectively.

3. FEATURE SPACE MATCHING

In this section we use the EM algorithm to find the estimates ν' of Equation 8. The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step (E step), we compute the auxiliary function given by

$$Q(\nu'|\nu) = E\left(\log p(Z|\nu')|Y, \nu\right), \tag{10}$$

where Y is the *incomplete* observed data, and Z is the *complete data* [3]. Different choices of Z lead to different EM algorithms. Under the choice of complete data $Z = \{Y, S, C\}$, the auxiliary function is

$$Q(\nu'|\nu) = E\{\log p(Y, S, C|\nu', \Lambda_X)|Y, \nu, \Lambda_X\}, \tag{11}$$

which can be re-written as

$$Q(\nu'|\nu) = \sum_{S,C} p(Y, S, C|\nu, \Lambda_X) \log p(Y, S, C|\nu', \Lambda_X). \tag{12}$$

In the second step, called the maximization step (M step), we find the value of ν' that maximizes $Q(\nu'|\nu)$, i.e.

$$\nu' = \underset{\nu'}{\text{argmax}}\, Q(\nu'|\nu) \tag{13}$$

It can be shown [3] that if $Q(\nu'|\nu) \geq Q(\nu|\nu)$ then $p(Y|\nu', \Lambda_X) \geq p(Y|\nu, \Lambda_X)$. Thus iteratively applying the E and M steps of Equations 10 and 13 guarantees that the likelihood is nondecreasing. The iterations are continued until the increase in the likelihood is less than some predetermined threshold.

In general, the function $F_\nu(.)$ of Equation 1 can map a block of Y into a block of X of different size. However, for simplicity, we assume that the function is such that it maps each frame of Y onto the corresponding frame of X, so that $x_t = f_\nu(y_t)$. We further assume that $f_\nu(y_t)$ operates separately on each component, i.e., $x_{t,i} = f_{\nu,i}(y_{t,i})$, and that the covariance matrices $C_{n,m}$ are diagonal, i.e., $C_{n,m} = \text{diag}(\sigma^2_{n,m})$. In what follows, for ease of the expressions, we drop the reference to the subscript i denoting the ith component of the vectors. We consider functions of the form

$$f_\nu(y_t) = ag(y_t) + b, \tag{14}$$

where $g(y_t)$ is a known (possibly non-linear) differentiable function of y_t, and $\nu = \{a, b\}$ is the set of unknown parameters. The auxiliary function of Equation 12 can now be written as [10, 12]

$$Q(a', b'|a, b) = \sum_{t,n,m}^{T,N,M} \gamma_t(n, m) \quad \mathrm{X}$$
$$\left[-\frac{1}{2} \frac{(a' g(y_t) + b' - \mu_{n,m})^2}{\sigma^2_{n,m}} + \log a' \right], \quad (15)$$

where $\gamma_t(n, m)$ is the joint likelihood of observing Y and the mth mixture component of the nth state producing the observation y_t given the current estimate of the parameters a and b. $\gamma_t(n, m)$ can be computed using the forward-backward algorithm (e.g. [10]). In order to compute the parameters that maximize the auxiliary function, we take the derivative of Equation 15 with respect to a' and b' respectively and set them to zero, getting

$$\sum_{t,n,m}^{T,N,M} \gamma_t(n, m) \left[\frac{1}{a'} - \frac{(a' g(y_t) + b' - \mu_{n,m})\, g(y_t)}{\sigma^2_{n,m}} \right] = 0, \quad (16)$$

and

$$\sum_{t,n,m}^{T,N,M} \gamma_t(n, m) \left[\left(a' g(y_t) + b' - \mu_{n,m} \right) / \sigma^2_{n,m} \right] = 0. \quad (17)$$

We can solve equation 16 and 17 explicitly for the estimates a' and b'.

In particular, consider an additive bias distortion $y_t = x_t + b_t$. This results by setting each component $g(y_t) = y_t$, and $a = 1$, in the above functional form. Thus a is known and only the parameter b has to be estimated. The iterative estimation formula for the ith component of a fixed bias b for a speech segment can be shown to be

$$b'_i = \frac{\sum_{t,n,m}^{T,N,M} \gamma_t(n, m)(y_{t,i} - \mu_{n,m,i})/\sigma^2_{n,m,i}}{\sum_{t,n,m}^{T,N,M} \gamma_t(n, m)/\sigma^2_{n,m,i}}. \quad (18)$$

4. MATCHING IN THE MODEL SPACE

In this section, we consider the case of probabilistic distortion functions which corresponds to viewing the problem in the model space. We estimate the parameters η of the model transformation $\Lambda_Y = G_\eta(\Lambda_X)$ by maximizing the likelihood $p(Y|W, \eta, \Lambda_X)$. In particular, consider the case of a *random* additive bias sequence $B = \{b_1, \cdots, b_T\}$. For simplicity, we assume that b_t is independent of the speech, and is i.i.d. Gaussian with mean μ_b and diagonal covariance $diag(\sigma^2_b)$, and $\Lambda_B = \{\mu_b, \sigma^2_b\}$. Under these conditions, Λ_Y is obtained by adding the mean μ_b and the variance σ^2_b to the means and variances of each mixture component of Λ_X. μ_b can be estimated as in Equation 18, except that $\sigma^2_{n,m,i}$ is replaced by $\sigma^2_{n,m,i} + \sigma^2_{b,i}$. However, to obtain a closed form expression for σ^2_b, we need to derive a new EM algorithm by considering the complete data $Z = (X, B, S, C)$ where the observed incomplete data is given by $Y = X + B$ [14, 12]. σ^2_b is now estimated by

$$\sigma^{2\,\prime}_{b,i} = \frac{\sum_{t,n,m}^{T,N,M} \gamma_t(n, m) E\left(b^2_{t,i} | y_{t,i}, n, m, \Lambda_X, \Lambda_B\right)}{\sum_{t,n,m}^{T,N,M} \gamma_t(n, m)} - \mu'^{\,2}_{b,i}. \quad (19)$$

$E\left(b^2_{t,i} | y_{t,i}, n, m, \Lambda_X, \Lambda_B\right)$ is the conditional expected value of $b^2_{t,i}$ given the tth observation $y_{t,i}$ and that this observation is generated from the mth mixture component of the nth state. This conditional expectation is easily evaluated from our knowledge of Λ_X and the current estimate of $\Lambda_B = \{\mu_b, \sigma^2_b\}$.

Table 1. Word Error Rate (%) Comparisons

	MIS	FS1	FS2	MS2	MAT	MS1
A-MIC	14.1	4.7	4.6	4.1	2.1	2.2
A-TEL	24.3	14.6	10.7	7.1	2.7	2.5
B-MIC	25.8	7.7	7.4	6.3	6.3	5.5
B-TEL	24.1	13.7	10.8	7.4	7.0	6.5

5. EXPERIMENTAL RESULTS

We tested the proposed approach on the 991-word DARPA resource management (RM) task using the RM word-pair grammar with a perplexity of 60. New simultaneous recordings of two non-native speakers were collected through two channels: 1) a close-talking microphone (MIC), and 2) a telephone handset over a dial-up line (TEL). The problem of speaker mismatch is not dealt with here. Therefore speaker adaptive models for 1769 context dependent units were created with 300 adaptation sentences and a set of speaker independent seed models using a MAP algorithm [15]. The test set consisted of an additional 75 utterances for each speaker (A and B) and each channel (MIC and TEL).

For this problem, we assume that the mismatch between the recordings can be modeled as an additive bias b_t in the cepstral observation domain. A fixed additive bias model in the *feature space* has previously been used for speaker adaptation [4]. In our approach, however, the bias may be viewed in the *feature space* or the *model space*. The bias parameters may also be estimated separately for different signal segments. For example, when part of the mismatch is due to noise, the additive bias model for channel mismatch is inaccurate, especially in regions of the utterance where the noise dominates. We are thus motivated to experiment with separate bias parameters for speech and silence frames, both in the signal and model space. In our experiments, the bias estimation was performed on a per-utterance basis. Table 1 gives the percentage word-error rates for speaker A and B for test data recorded using either MIC or TEL channels under mismatched conditions (MIS), and after processing with three sets of bias estimation approaches, namely: 1) a single bias in feature space (FS1); 2) a separate speech and silence bias in feature space (FS2); and 3) a separate speech and silence random bias in model space (MS2). The table also shows the word-error rates in matched conditions with 1) no processing (MAT), and 2) stochastic matching by estimating a single random bias in the model space (MS1).

Table 2. Word Error Rate (%) After CMS Processing

	MIS	FS1	FS2	MS2	MAT	MS1
A-MIC	5.0	5.2	3.7	3.7	3.0	3.1
A-TEL	12.0	12.6	8.3	6.1	3.1	3.1
B-MIC	9.9	9.7	8.4	7.9	5.0	5.3
B-TEL	8.9	9.1	7.0	7.7	6.3	5.0

We see that the proposed stochastic matching algorithm consistently reduces the word-error rate by about 70% for all the speaker/channel combinations (see MS2 results in the table). It is seen that estimating separate speech and silence feature-space bias parameters (FS2) is superior to single parameter estimates (FS1). This is also true for the model-space estimates [12]. Furthermore, the table also shows that model space parameter estimation gives better performance than feature space estimation. For speaker B, the performance approaches that of the matched conditions. In addition, the proposed approach maintains the performance even in matched conditions (see the last two columns of Table 1).

We also compared the stochastic matching approach with the popular cepstral mean subtraction (CMS) approach where the average cepstrum over the entire utterance is subtracted from each frame. Table 2 gives the results for the different stochastic matching approaches of Table 1, except that now the training and testing utterances are first processed by CMS.

Comparisons between the single feature space bias estimate (FS1 in Table 1) and CMS (MIS in Table 2) do not clearly show the superiority of one over the other. However, the CMS results (MIS column in Table 2) were not as good compared to the stochastic matching algorithm results shown in Table 1 for the cases of two feature space bias estimates (FS2) and the model space approaches (MS2).

The stochastic matching algorithm can also be applied *after CMS processing* as shown in Table 2. It can be seen that a single feature space bias estimate (FS1) results in similar performance to the mismatched case (MIS). This is not surprising, as the CMS processing has caused both the training and testing utterances to be zero mean. However, a separate speech and silence bias vector estimate (FS2) results in an additional performance improvement. Furthermore, the results show that the model space bias parameter estimation procedures (MS2) also decrease the word error rate. When compared with the model space approaches without CMS processing (Table 1), there is no clear improvement shown in the MS2 results listed in Table 2. Finally, under matched conditions (the last two columns of Table 2), we see that the stochastic matching algorithm maintains the performance well.

6. SUMMARY

We have presented a maximum likelihood stochastic matching approach to decrease the acoustic mismatch between a test utterance and a given set of speech HMMs so as to reduce the recognition performance degradation caused by possible distortions in the test utterance. In contrast to most approaches that use stereo data to estimate the mismatch model before recognition, the stochastic matching approach estimates the mismatch and performs recognition at the same time using the proposed EM algorithm. We found the approach mathematically attractive and gave improved recognition performance in mismatch training and testing conditions while maintaining recognition accuracy if training and testing conditions are acoustically similar.

REFERENCES

[1] B.-H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, vol. 5, pp. 275–294, 1991.

[2] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers, 1992.

[3] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[4] S. Cox and J. Bridle, "Unsupervised Speaker Adaptation by Probabilistic Spectrum Fitting," in *Proc. ICASSP*, pp. 294–297, 1989.

[5] Y. Zhao, "A New Speaker Adaptation Technique Using Very Short Calibration Speech," in *Proc. ICASSP*, pp. II–562–II–565, 1993.

[6] M. Rahim and B.-H. Juang, "Signal Bias Removal for Robust Telephone Speech Recognition in Adverse Environments," in *Proc. ICASSP*, 1994.

[7] V. Digalakis, D. Rtischev, and L. Neumeyer, "Fast speaker adaptation using constrained reestimation of Gaussian mixtures," 1994, submitted to *IEEE T-SAP*.

[8] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of continuous density HMMs using multivariate linear regression," in *Proc. ICSLP*, pp. 451-454, 1994.

[9] L. Neumeyer and M. Weintraub, "Probabilistic optimum filtering for robust speech recognition," in *Proc. ICASSP*, pp. I–417–I–420, 1994.

[10] B.-H. Juang, "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," *AT&T Technical Journal*, vol. 64, no. 6, pp. 1235–1249, 1985.

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

[12] A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," 1994, submitted to *IEEE T-SAP*.

[13] C.-H. Lee, L. R. Rabiner, R. Pieraccini, and J. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, vol. 4, pp. 127–165, April 1990.

[14] R. Rose, "Integrated Models of Speech and Background with Application to Speaker Identification in Noise," *IEEE T-SAP*, vol. 2, pp. 245–257, April 1994.

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