ADAPTATION OF HIDDEN MARKOV MODELS USING MULTIPLE STOCHASTIC TRANSFORMATIONS

V. Diakoloukas and V. Digalakis
Dept. of Electrical & Computer Engineering
Technical University of Crete
73100 Chania, Crete, GREECE
{vdiak,vas}@telecom.tuc.gr

ABSTRACT

The recognition accuracy in recent large vocabulary Automatic Speech Recognition (ASR) systems is highly related to the existing mismatch between the training and test sets. For example, dialect differences across the training and testing speakers result to a significant degradation in recognition performance. Some popular adaptation approaches improve the recognition performance of speech recognizers based on hidden Markov models with continuous mixture densities by using linear transforms to adapt the means, and possibly the covariances of the mixture Gaussians. In this paper, we propose a novel adaptation technique that adapts the means and, optionally, the covariances of the mixture Gaussians by using multiple stochastic transformations. We perform both speaker and dialect adaptation experiments, and we show that our method significantly improves the recognition accuracy and the robustness of our system. The experiments are carried out with SRI's DECIPHER^{TM} speech recognition system.

1. INTRODUCTION

The mismatch that frequently occurs between the training and testing conditions of an automatic speech recognizer (ASR) can be efficiently reduced by adapting the parameters of the recognizer to the testing conditions. Recently, a family of adaptation algorithms for large-vocabulary continuous-density hidden-Markov-model (HMM) based speech recognizers have appeared that are based on constrained reestimation of the distribution parameters [3, 6, 9]. In these approaches, all the Gaussians in a single mixture, or a group of mixtures, if there is tying of transformations, are transformed using the same linear transformation.

The linear assumption, however, may be too restrictive and inadequate in modeling the characteristics of the testing conditions. In this paper, we introduce a new adaptation method for continuous-density HMMs that is based on a more complex transformation of the Gaussians, which consists of a collection of piecewise-linear transformations that are shared among all the Gaussians in each mixture. The transformation for each Gaussian is selected probabilistically, based on weight probabilities that are trained from the adaptation data. We evaluate our new method using SRI's DECIPHER^{TM} speech recognition system on dialect and speaker adaptation experiments.

2. ADAPTATION USING A LINEAR TRANSFORMATION

Recently developed fast adaptation algorithms [3, 6, 9] for continuous mixture density HMMs are based on constrained reestimation of the mixture Gaussians. Maximum likelihood reestimation of the Gaussians in all these adaptation schemes is performed using the expectation-maximization (EM) algorithm.

The observation densities of the speaker-independent (SI) speech recognition system in continuous mixture-density HMMs have the following form:

$$P_{SI}(y_t|s_t) = \sum_{i=1}^{N_\omega} \mathrm{p}(\omega_i|s_t) N(y_t; m_{s_t i}, S_{s_t i}), \qquad (1)$$

where $N(y_t; m_{s_t i}, S_{s_t i})$ denotes the multivariate normal density with mean vector $m_{s_t i}$ and covariance matrix $S_{s_t i}$.

These models need large amounts of training data for robust estimation of their parameters. However, given a small amount of training data that match the testing conditions, the initial SI system can be adapted to new conditions, like the speaker, channel, or the dialect. In [3], it is assumed that the vector process $[x_t]$ of the mismatched testing condition can be obtained through a sequence of linear transformations (A_{s_t}, b_{s_t}) from a corresponding process $[y_t]$ that matches the training population:

$$x_t = A_{s_t} y_t + b_{s_t}. \qquad (2)$$

The transformation used at each time t depends on the underlying HMM state s_t, in which case the observation densities of the adapted models can be written:

$$P_A(x_t|s_t) = \sum_{i=1}^{N_\omega} \mathrm{p}(\omega_i|s_t) N(x_t; A_{s_t} m_{s_t i} + b_{s_t}, A_{s_t} S_{s_t i} A_{s_t}^T), \qquad (3)$$

where A^T denotes the transpose of a matrix. In [6], the linear constraint is only applied to the means of the adapted observation densities, which become

$$P_A(x_t|s_t) = \sum_{i=1}^{N_\omega} \mathrm{p}(\omega_i|s_t) N(x_t; A_{s_t} m_{s_t i} + b_{s_t}, S_{s_t i}). \qquad (4)$$

Closed-form solutions for the reestimation formulae of method (3) can be derived in the case of diagonal transformation and covariance matrices. The reestimation formulae for (4) are simpler, can be used for full transformation matrices, but do not adapt the covariances of the Gaussians. A comparative study of these two approaches was done in [7].

3. ADAPTATION USING MULTIPLE STOCHASTIC TRANSFORMATIONS

The linear assumption may not adequately model the dependency between the training and testing conditions. For example, it may be too simplistic to assume that the mapping between the observation spaces of a new speaker and the speakers in the training population is linear, even when we are looking only at the observations of a particular group of HMM states. An alternate to the deterministic linear transformation described in equation (2), is to use for an observation drawn from the i-th Gaussian of a particular HMM state s_t a probabilistic, piecewise linear transformation of the form:

$$x_t = \begin{cases} A_{s_t 1} y_t + b_{s_t 1}, \text{with probability } \mathrm{p}(\lambda_1 | s_t, \omega_i) = l_1 \\ A_{s_t 2} y_t + b_{s_t 2}, \text{with probability } \mathrm{p}(\lambda_2 | s_t, \omega_i) = l_2 \\ \quad \vdots \\ A_{s_t N_\lambda} y_t + b_{s_t N_\lambda}, \\ \qquad \text{with probability } \mathrm{p}(\lambda_{N_\lambda} | s_t, \omega_i) = l_{N_\lambda} \end{cases} \quad (5)$$

where N_λ is the number of component transformations used by each HMM state, $l_1 + l_2 + \ldots + l_{N_\lambda} = 1$, and $l_j \geq 0, j = 1, \ldots, N_\lambda$. The random variable λ_j is the index of the transformation that is used at each time, and the component transformations $A_{s_t j}, b_{s_t j}$ for $j = 1, \ldots, N_\lambda$ are shared by all the Gaussians used by state s_t. The probabilities $\mathrm{p}(\lambda_j | s_t, \omega_i)$ that select the j-th transformation at time t for the i-th Gaussian of the HMM state s_t, however, are specific to each Gaussian in the mixture.

Let us consider adaptation using a complex transformation consisting of N_λ component transformations. The adapted observation densities of the HMM-based speech recognizer will then have the following form:

$$P_A(x_t | s_t) = \sum_{i=1}^{N_\omega} \sum_{j=1}^{N_\lambda} \mathrm{p}(\lambda_j | s_t, \omega_i)\ \mathrm{p}(\omega_i | s_t) \cdot N(x_t; A_{s_t j} m_{s_t i} + b_{s_t j}, A_{s_t j} S_{s_t i} A_{s_t j}^T). \quad (6)$$

Alternatively, if we choose to apply the transformations only to the means of the Gaussians, as in equation (4), then the adapted observation densities will be:

$$P_A(x_t | s_t) = \sum_{i=1}^{N_\omega} \sum_{j=1}^{N_\lambda} \mathrm{p}(\lambda_j | s_t, \omega_i)\ \mathrm{p}(\omega_i | s_t) \cdot N(x_t; A_{s_t j} m_{s_t i} + b_{s_t j}, S_{s_t i}). \quad (7)$$

In either case, the parameters that must be estimated from the adaptation data for each HMM state include the transformation parameters $A_{s_t j}, b_{s_t j}, j = 1, \ldots, N_\lambda$ and the transformation probabilities, $\mathrm{p}(\lambda_j | s_t, \omega_i), j = 1, \ldots, N_\lambda$, for each Gaussian in the mixture, $i = 1, \ldots, N_\omega$. In [1] we show that these parameters can be estimated using the EM algorithm. The proof is based on the maximization of the following auxiliary function at each EM iteration:

$$\theta_n = \mathrm{argmax}_\theta E\{\log f(\mathcal{X}, \mathcal{Z} | \theta) | \mathcal{X}, \theta_o\}$$

where θ_o are the model parameters of the previous iteration, $\mathcal{X}$ is the set of the observation data samples x_k and $\mathcal{Z}$ denotes the set of the corresponding unobserved data z_k which consists of the HMM states, the set of mixture indices $\omega_i \epsilon \Omega$ and the set of the component transforms' indices $\lambda_j \epsilon \Lambda$.

The training procedure can then be summarized as follows:

- **Initialization:** Initialize all transformation parameters $A_{s_t j}$, $b_{s_t j}$ and $\mathrm{p}(\lambda_j | s_t, \omega_i)$. For our experiments, we set $A_{s_t j} = I$, where I is the identity matrix, and $b_{s_t j} = h_j \otimes s_{s_t}$ where $\otimes$ represents the element-wise product of two vectors, s_{s_t} is a vector with elements the standard deviations of the observation vector for state s_t, h_j is the j-th column of a $d \times d$ Hadamard matrix, and d is the dimension of the offset vector $b_{s_t j}$. Finally, we initialize the weight probabilities with $\mathrm{p}(\lambda_j | s_t, \omega_i) = 1/N_\lambda$.
- **E-step:** Perform one iteration of the forward-backward algorithm on the speech data, using the adapted Gaussians with the current value of the transformation parameters $\theta_{s_t}^{(k)} = [A_{s_t j}^{(k)}, b_{s_t j}^{(k)}, \mathrm{p}^{(k)}(\lambda_j | s_t, \omega_i), \forall j = 1, \ldots, N_\lambda, \forall i = 1, \ldots, N_\omega]$ where k is the iteration index. Collect the sufficient statistics as defined below:

$$n_{s_t ij} = \sum_t \rho(s_t) \phi_i(s_t) \psi_{ij}(s_t) \quad (8)$$

$$\bar{\mu}_{s_t ij} = \frac{1}{n_{s_t ij}} \sum_t \rho(s_t) \phi_i(s_t) \psi_{ij}(s_t) x_t \quad (9)$$

$$\bar{\Sigma}_{s_t ij} = \frac{1}{n_{s_t ij}} \sum_t \rho(s_t) \phi_i(s_t) \psi_{ij}(s_t) \cdot (x_t - \bar{\mu}_{s_t ij})(x_t - \bar{\mu}_{s_t ij})^T \quad (10)$$

where $\rho(s_t) = \mathrm{p}(s_t | \mathcal{X}, \zeta^{(k)})$ is the probability of being at state s_t at time t given $\mathcal{X}$ and the current HMM parameters $\zeta^{(k)}$ and is computed by the forward-backward recursions. The posterior probabilities

$$\phi_i(s_t) = \mathrm{p}(\omega_i | x_t, s_t, \theta_{s_t}^{(k)}) \quad (11)$$

$$\psi_{ij}(s_t) = \mathrm{p}(\lambda_j | \omega_i, x_t, s_t, \theta_{s_t}^{(k)}) \quad (12)$$

can be computed using Bayes rule.

When the transformation is applied only to the means of the Gaussians, then only the first order statistics given in equations (8) and (9) have to be computed, since the covariance remains the same through the iterations.

- **M-step:** Compute the new transformation parameters and component transformation probabilities $[A_{s_t j}^{(k+1)}, b_{s_t j}^{(k+1)}, \mathrm{p}^{(k+1)}(\lambda_j | s_t, \omega_i)]$. The component transformation probabilities are calculated from the quantity:

$$\mathrm{p}^{(k+1)}(\lambda_j | s_t, \omega_i) = \frac{n_{s_t ij}}{\sum_{j=1}^{N_\lambda} n_{s_t ij}}. \quad (13)$$

For diagonal covariances $S_{s_t i}$ and transformation matrices $A_{s_t j}$,

$$\begin{aligned} S_{s_t i} &= \mathrm{diag}(s_{s_t i1}^2, s_{s_t i2}^2, \ldots, s_{s_t id}^2) \\ A_{s_t j} &= \mathrm{diag}(a_{s_t j1}, a_{s_t j2}, \ldots, a_{s_t jd}) \\ b_{s_t j} &= [b_{s_t j1}, b_{s_t j2}, \ldots, b_{s_t jd}]^T \\ m_{s_t i} &= [m_{s_t i1}, m_{s_t i2}, \ldots, m_{s_t id}]^T \\ \bar{\Sigma}_{s_t ij} &= \mathrm{diag}(\bar{\sigma}_{s_t ij1}^2, \bar{\sigma}_{s_t ij2}^2, \ldots, \bar{\sigma}_{s_t ijd}^2) \\ \bar{\mu}_{s_t ij} &= [\bar{\mu}_{s_t ij1}, \bar{\mu}_{s_t ij2}, \ldots, \bar{\mu}_{s_t ijd}]^T, \end{aligned}$$

where d is the dimension of the observation vectors, the maximization step is equivalent to solving the following set of equations $\forall h = 1, \ldots, d$, in addition to reestimating the transformation probabilities from (13),

$$\left(\sum_{i=1}^{N_\omega} n_{s_t ij}\right) a^2_{s_t jh} - \left(\sum_{i=1}^{N_\omega} \frac{n_{s_t ij}}{s^2_{s_t ih}}\right) b^2_{s_t jh} - \left(\sum_{i=1}^{N_\omega} \frac{n_{s_t ij} m_{s_t ih}}{s^2_{s_t ih}}\right) a_{s_t jh} b_{s_t jh} + \left(\sum_{i=1}^{N_\omega} \frac{n_{s_t ij} \bar{\mu}_{s_t ijh} m_{s_t ih}}{s^2_{s_t ih}}\right) a_{s_t jh} + \left(2 \sum_{i=1}^{N_\omega} \frac{n_{s_t ij} \bar{\mu}_{s_t ijh}}{s^2_{s_t ih}}\right) b_{s_t jh} - \left(\sum_{i=1}^{N_\omega} n_{s_t ij} \frac{\bar{\mu}^2_{s_t ijh} + \bar{\sigma}^2_{s_t ijh}}{s^2_{s_t ih}}\right) = 0 \quad (14)$$

where the offset $b_{s_t jh}$ is given by:

$$b_{s_t jh} = \frac{\sum_{i=1}^{N_\omega} \frac{n_{s_t ij}(\bar{\mu}_{s_t ijh} - m_{s_t ih} a_{s_t jh})}{s^2_{s_t ih}}}{\sum_{i=1}^{N_\omega} \frac{n_{s_t ij}}{s^2_{s_t ih}}}. \quad (15)$$

In the general case, when the covariances and transformations are full matrices, we can use iterative schemes to solve a system of second order equations [3].

When the transformation is applied only to the means of the Gaussians, then the maximization step involves the computation of the component transformation probabilities $\mathrm{p}(\lambda_j|\omega_i, \theta)$ from equation (13) and the transformation parameters which is now equivalent to solving the following system of equations [1]:

$$\sum_{i=1}^{N_\omega} n_{s_t ij} S^{-1}_{s_t i} \left[A_{s_t j} m_{s_t i} + (b_{s_t j} - \bar{\mu}_{s_t ij})\right] m^T_{s_t i} = 0 \quad (16)$$

$$b_{s_t j} = \left(\sum_{i=1}^{N_\omega} n_{s_t ij} S^{-1}_{s_t i}\right)^{-1} \cdot \left(\sum_{i=1}^{N_\omega} n_{s_t ij} S^{-1}_{s_t i} \left[\bar{\mu}_{s_t ij} - A_{s_t j} m_i\right]\right) \quad (17)$$

- if another iteration go to E-step.

The adapted mixture densities in equations (6) and (7) using the multiple stochastic transformations consist of N_λ-times as many Gaussians as the original, SI observation densities, which means that the adapted system will require additional computation during recognition. This can be avoided if we constrain the transformation probabilities:

$$\mathrm{p}(\lambda_j|s_t, \omega_i) = \begin{cases} 1 & \text{for the transformation with} \\ & \quad \text{the highest probability,} \\ 0 & \text{elsewhere,} \end{cases} \quad (18)$$

which means that we only apply the transform with the highest probability to each Gaussian.

4. EXPERIMENTS

We have tested our new algorithm in dialect adaptation experiments, trying to develop a multi-dialect SI speech recognition system for the Swedish language which will require only a small amount of dialect-dependent data. We use the Swedish language corpus collected by Telia, and the recognizer used in a bidirectional speech translation system between English and Swedish that has been developed under the SRI-Telia Research Spoken Language Translator project [8]. We have also evaluated our algorithm in speaker adaptation experiments based on the "spoke 3" task of the large-vocabulary Wall Street Journal (WSJ) corpus [5]. The goal of this task is to improve recognition performance for nonnative speakers of American English.

4.1. Dialect Adaptation Experiments

For our dialect adaptation experiments we used data from the Stockholm and Scanian dialects, that were, respectively, the seed and target dialects. There is a total of 40 speakers from the Scanian dialect, both male and female, and each of them recorded more than 40 sentences. We selected 8 of the speakers (half of them male) to serve as testing data and the rest composed the adaptation/training data with a total of 3814 sentences. Experiments were carried out using SRI's DECIPHER^{TM} system [4]. The system's front-end was configured to output 12 cepstral coefficients, cepstral energy and their first and second derivatives. The cepstral features are computed with a fast Fourier transform (FFT) filterbank and subsequent cepstral-mean normalization on a sentence basis is performed.

The SI continuous HMM system which served as seed models for our adaptation scheme, was a phonetically-tied mixture (PTM) system [4] trained on approximately 21,000 sentences of Stockholm dialect. The system's recognition performance on an air travel information task similar to the English ATIS one was benchmarked at a 8.9% word error rate using a bigram language model when tested on Stockholm speakers. On the other hand, its performance degraded significantly when tested on the Scanian-dialect testing set, reaching a word error rate of 25.08%.

In previous work [2], we adapted the Stockholm-dialect system using equation (3) with diagonal transformations (method I) and equation (4) with structured transformations (method II). The transformation matrices in method II are block diagonal matrices, with three blocks that perform a separate transformation to every basic feature vector (cepstrum, and its first and second derivatives). The results are summarized in Table 1 for 198 and 520 adaptation sentences, and we found that method II outperformed method I because of the more complex transformations that allowed rotation, in addition to scaling and shifting. These results were consistent with similar findings on speaker-adaptation experiments reported in [7].

The results of our new method are also summarized in Table 1 for different numbers of component transformations. We used multiple diagonal transformations applied to both the means and covariances, as in (6). We see that even with as few as two component transformations, we get a performance improvement over methods I and II. When more component transformations are used, the

	Parameters per transform	*Word Error Rate %*	
		198 train. sentences	520 train. sentences
Method I	78	15.5	13.3
Method II	546	13.7	12.6
Stochastic			
$N_\lambda = 2$	156	13.3	12.1
$N_\lambda = 3$	234	13.0	10.8
$N_\lambda = 4$	312	13.2	10.2
$N_\lambda = 5$	390	12.3	10.0
$N_\lambda = 6$	468	12.8	10.2

Table 1. Number of adaptation parameters per mixture and dialect-adapted word-error rates for linear transformations (methods I and II) and multiple stochastic transformations with 2-6 component transformations.

new multiple stochastic transform method gives significantly better results than the previous approaches, with the best performance achieved for five transformations. The word error rate for 198 adaptation sentences is reduced by 21% and 10% over methods I and II, respectively. For 520 adaptation sentences, the word error rate is reduced by 25% and 21% over methods I and II, respectively, although the number of adaptation parameters is smaller than those used in method II.

4.2. Speaker Adaptation Experiments

For the speaker adaptation experiments we used the DECIPHER^{TM} system on the "spoke 3" task of the large-vocabulary Wall Street Journal (WSJ) corpus [5]. The speaker-independent, continuous HMM systems that were used as seed models for adaptation were gender dependent, trained on 140 speakers and 17,000 sentences for each gender. Each of the two systems was phonetically tied, having 12,000 context-dependent phonetic models that shared 100 Gaussians specific to each center phone. We used the 5,000-word closed-vocabulary bigram language model provided by the MIT Lincoln Laboratory, and the 1994 development set that consists of six female and 5 male speakers, each one of them speaking 40 phonetically rich adaptation sentences. The test set consisted of 11 speakers and 20 sentences per speaker.

The speaker-independent word-error rate for this test set is 29.06%. We evaluated our new method for 10, 20 and 40 stochastic transformations. Each of the stochastic transformations was used to adapt the Gaussians of all allophone states clustered in each of 10, 20 and 40 groups, respectively, that corresponded to one of the stochastic transformations. We used diagonal component transformations applied to both the means and the covariances (6), and the number of component transformations in each stochastic transformation varied from 1 (in which case our new method simply reduces to method I) to 8. The results are summarized in Table 2. We see that with as few as 2 component transformations there is a significant improvement in recognition performance, compared to method I. The improvement becomes more obvious as we use more component transformations. The best performance is achieved for 8, 6 and 6 component transformations reducing the speaker-independent word-error rate by 38.8%, 40.8% and 42.2% when 10, 20 and 40 transformations are used, respectively. The improvement in performance over method I is 23.3%, 16.9% and 13.8% for 10, 20 and 40 stochastic transformations, respectively.

Component Transforms	*Word Error Rate %*		
	10 trans.	20 trans.	40 trans.
1 (method I)	23.2	20.7	19.5
2	21.4	18.8	18.5
3	19.9	19.0	17.9
4	19.1	17.7	17.7
5	18.4	18.1	17.5
6	17.9	17.2	16.8
7	18.2	17.5	17.1
8	17.8	17.4	17.3

Table 2. Speaker-adapted word-error rates for several numbers of stochastic transformations consisting of 1 (method I) up to 8 components.

ACKNOWLEDGMENTS

The work we have described was accomplished under contract to Telia Research.

REFERENCES

[1] V.Diakoloukas and V.Digalakis, "Adaptation of Hidden Markov Models Using Multiple Stochastic Transformations", *Technical Report, Technical University of Crete*, Chania, Greece, 1997.

[2] V.Diakoloukas, V.Digalakis, L.Neumeyer, J.Kaja, "Development of Dialect-Specific Speech Recognizers Using Adaptation Methods," *Proc. Int'l. Conf. on Acoust., Speech and Signal Processing,* Munich, 1997.

[3] V. Digalakis, D. Rtischev and L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures," *IEEE Transactions Speech and Audio Processing,* pp. 357–366, September 1995.

[4] V. Digalakis, P. Monaco and H. Murveit, "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers," *IEEE Transactions Speech and Audio Processing,* June 1996.

[5] F. Kubala *et al.*, "The Hub and Spoke Paradigm for CSR Evaluation," *Proc. ARPA Workshop on Human Language Technology,* March 1994.

[6] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language,* pp. 171–185, 1995.

[7] L.Neumeyer, A.Sankar and V.Digalakis, "A Comparative Study of Speaker Adaptation Techniques", *Proceedings of European Conference on Speech Communication and Technology,* pp. 1127–1130, Madrid, Spain, 1995.

[8] M. Rayner,I. Bretan,D. Carter, M. Collins, V. Digalakis, B. Gambäck, J. Kaja, J. Karlgren, B. Lyberg, P. Price, S. Pulman and C. Samuelsson, "Spoken Language Translation with Mid-90's Technology: A Case Study," *Proc. Eurospeech '93*, Berlin, 1993.

[9] A. Sankar and C.-H. Lee, "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Transactions Speech and Audio Processing,* pp. 190–202, May 1996.