A Compact Model for Speaker-Adaptive Training

†*Tasos Anastasakos, John McDonough, Richard Schwartz, John Makhoul*

BBN Systems and Technologies
70 Fawcett Street, Cambridge, MA 02138
†Northeastern University, Boston, MA
E-mail: tasos@bbn.com

ABSTRACT

In this work we formulate a novel approach to estimating the parameters of continuous density HMMs for speaker-independent (SI) continuous speech recognition. It is motivated by the fact that variability in SI acoustic models is attributed to both phonetic variation and variation among the speakers of the training population, that is independent of the information content of the speech signal. These two variation sources are decoupled and the proposed method jointly annihilates the inter-speaker variation and estimates the HMM parameters of the SI acoustic models.

We compare the proposed training algorithm to the common SI training paradigm within the context of supervised adaptation. We show that the proposed acoustic models are more efficiently adapted to the test speakers, thus achieving significant overall word error rate reductions of 19% and 25% for 20K and 05K vocabulary tasks respectively.

1. INTRODUCTION

The most common approach to modeling inter-speaker variability for speaker-independent (SI) HMM-based recognizers, is to estimate the parameters of the acoustic models from speech collected from a large population of speakers. While the SI models achieve a low average word error rate for test speakers that are not included in the training data, they are less accurate than adequately trained speaker dependent (SD) acoustic models[1].

An inherent difficulty in modeling SI continuous speech is that spectral variations in each speech unit are caused by inter-speaker variability, in addition to phonetically relevant variation sources. The differences among speakers lie in the anatomy of the vocal tract and the vocal cords, in regional dialects and in speaking idiosyncracies which are all manifested as variations in the speech signal. As a result, the spectral distributions often exhibit higher variance than the corresponding SD distributions and hence higher overlap among different speech units. This may result in diffused acoustic models with reduced discriminatory capabilities.

Previous efforts to generate acoustic models with reduced variation due to speaker- or channel-induced factors focused on normalizing the acoustic space prior to estimating the parameters of the acoustic models. Cepstrum mean removal [1] has been the simplest feature space based normalization method that was used mainly to counteract channel effects. In [2] a parametric model of vocal tract length normalization reduces the inter-speaker variability of the acoustic space by appropriately warping the frequency axis for each training speaker prior to computing the cepstral coefficients. The Metamorphic algorithm [3] estimates a piecewise linear transformation between the spectral space of a prototypical speaker and other reference speakers in order to map the reference speakers onto the prototypical space. In [4], an acoustic normalization technique within the framework of mixture density HMM was applied to normalize the training as well as the test data, and in [5], a maximum likelihood signal bias was jointly estimated with the parameters of a discrete HMM.

In this paper, we propose an approach to HMM training for speaker independent continuous speech recognition that integrates the normalization as part of the continuous density HMM estimation problem. The proposed method is based on a maximum likelihood formulation that aims at separating the two processes, one being the speaker specific variation and the other the phonetically relevant variation of the speech signal. By modeling separately the speaker variation and annihilating its effect in the training data, we are able to reduce the variance and hence the overlap of the acoustic models. We term the resulting HMM acoustic models as *compact models*.

2. GENERAL FRAMEWORK OF THE SPEAKER ADAPTIVE TRAINING PARADIGM

In the common pooled speaker independent training paradigm we estimate the parameters $\boldsymbol{\lambda}$ of the HMM model so that the resulting model maximizes the likelihood of the training observation sequences. Consider a training database that consists of speech collected from R speakers, with each speaker r,

[1] The distinction between speaker independent (SI) and speaker dependent (SD) speech recognizers refers to the collection of the training data only, as both approaches use the same parameter estimation techniques. The parameters of SD acoustic models are estimated from training data from the speaker that will use the system and potentially achieve the lowest error rates. Previous studies have shown that SI recognizers have 2 to 3 times higher word error rate than adequately trained SD systems. However the requirement of large amounts of training data for each test speaker decrease the utility and portability of SD systems.

($r = 1, 2, \ldots, R$), contributing a transcribed observation sequence[2] $O^{(r)} = (o_1^{(r)}, \ldots, o_{T_r}^{(r)})$ of length T_r. The optimal model $\overline{\boldsymbol{\lambda}}$, in the maximum likelihood sense, is derived as:

$$\overline{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \mathcal{L}(\mathcal{O}; \boldsymbol{\lambda}) = \arg\max_{\boldsymbol{\lambda}} \prod_{r=1}^{R} \mathcal{L}(O^{(r)}; \boldsymbol{\lambda}) \qquad (1)$$

where $\mathcal{L}(O^{(r)}; \boldsymbol{\lambda})$ is the likelihood of the observation sequence $O^{(r)}$ given the existing set of models $\boldsymbol{\lambda}$. The underlying assumption of this approach is that all observations are produced from the same source. Hence, speaker characteristics, channel conditions and noise level are considered constant through the entire database.

The proposed *Speaker Adaptive Training* (SAT) paradigm is based on an underlying generative process that addresses explicitly the speaker-induced variations. We hypothesize a model of phonetically relevant variation $\boldsymbol{\lambda}_c$, and use the term *compact model* for $\boldsymbol{\lambda}_c$ to indicate that this model would exhibit less overlap among the speech units. A transformation $\boldsymbol{G}^{(r)}$ for each speaker r in the training population accounts for the particular speaker individuality and maps the compact model to a speaker dependent model in the same way that speaker adaptation methods [6, 7] generate speaker dependent models for each test speaker from a speaker independent seed model. Based on these two components, each training observation sequence $O^{(r)}$, collected from a particular speaker r, is generated by the hypothesized speaker dependent model $\boldsymbol{G}^{(r)}(\boldsymbol{\lambda}_c)$. In the SAT framework, the optimum set of HMM parameters $\overline{\boldsymbol{\lambda}}_c$ and the speaker transformations $\overline{\boldsymbol{\mathcal{G}}} = (\overline{\boldsymbol{G}}^{(1)}, \ldots, \overline{\boldsymbol{G}}^{(R)})$ are estimated jointly from a training database collected from a population of R speakers so as to maximize the likelihood of the training data

$$\left(\overline{\boldsymbol{\lambda}}_c, \overline{\boldsymbol{\mathcal{G}}}\right) = \arg\max_{(\boldsymbol{\lambda}_c, \boldsymbol{\mathcal{G}})} \prod_{r=1}^{R} \mathcal{L}(O^{(r)}; \boldsymbol{G}^{(r)}(\boldsymbol{\lambda}_c)) \qquad (2)$$

The introduction of the speaker transformations aims at reducing the speaker-specific variation in the speech signal allowing the compact model to represent more accurately the phonetically relevant context dependent variation. It is our goal to estimate these transformations from the training data jointly with the compact model HMM parameters.

The choice of transformation is coupled with the SAT paradigm in as much as we would like to apply the same speaker transformation method in the recognition stage. The modeling accuracy of the acoustic models is important in adapting the system to the test speakers efficiently using very little adaptation data. Hence the relative merit of the SAT paradigm would be demonstrated in recognition scenarios that involve speaker adaptation. It intuitively appears that the variation that is being modeled by the transformation in the training should be compensated during recognition in order to match more accurately the test speaker characteristics.

[2]The observation sequence may consist of a number of utterances. We denote the training data of one speaker as a single entity for notation simplicity.

3. SAT PARAMETER ESTIMATION

We assume that $\boldsymbol{\lambda}_c$ is a set of continuous density HMM triphone models with N states that is characterized by the state transition matrix A, the initial probability vector π and a set of state observation probability density functions. The i-th state observation density is assumed to be a mixture of Gaussians given by

$$b_i(o_t) = \sum_{k=1}^{K} c_{ik}\, \mathcal{N}(o_t;\, \mu_{ik},\, \Sigma_{ik}) \qquad (3)$$

where K is the number of mixture components, c_{ik} are the mixture component weights and (μ_{ik}, Σ_{ik}) are the mean vector and the covariance matrix of the d-dimensional multivariate Gaussian k-th component of the i-th state distribution.

The functional form of the transformation depends upon our prior knowledge about the extraneous variation that we wish to compensate. In this work we model the speaker specific characteristics using linear regression matrices, motivated by the Maximum Likelihood Linear Regression (MLLR) method [8, 6] that has recently shown to operate effectively in a variety of scenarios of supervised and unsupervised speaker adaptation. The aim of MLLR is to obtain the linear transformation that maximizes the likelihood of the adaptation data. The transformation $\boldsymbol{G}^{(r)} = \left(W^{(r)}, \beta^{(r)}\right)$ gives a new estimate of the Gaussian means

$$\mu^{(r)} = W^{(r)}\mu + \beta^{(r)} \qquad (4)$$

where $W^{(r)}$ is a $d \times d$ transformation matrix and $\beta^{(r)}$ is an additive bias vector. Hence the i-th state observation density adapted to the specific speaker is given by

$$b_i(o_t^{(r)}) = \sum_{k=1}^{K} c_{ik}\, \mathcal{N}(o_t^{(r)};\, W^{(r)}\mu_{ik} + \beta^{(r)}, \Sigma_{ik}) \qquad (5)$$

In the following development, we shall assume that the speaker specific transformation consists of a single regression matrix for simplicity. It is possible however to define regression classes and associate a regression matrix with each class [6]. The extension of the SAT parameter estimation to multiple regression transformations is straightforward.

The HMM parameter estimation is an incomplete-data problem as available observation sequences $\mathcal{O}$ provide only probabilistic evidence of the underlying *hidden* state sequence $\mathcal{S}$ and mixture component sequence $\mathcal{K}$ that correspond to each observation sequence. The *Expectation-Maximization* (EM) algorithm [9] is an iterative procedure for approximating ML estimates in the context of incomplete-data cases and has been used extensively in the HMM parameter estimation. This procedure consists of maximizing at each iteration the auxiliary Q-function

$$\mathcal{Q}(\overline{\theta}, \theta) = \mathcal{E}\left\{\log \mathcal{L}(\{\mathcal{O}, \mathcal{S}, \mathcal{K}\}; \overline{\theta}) | \mathcal{O}; \theta\right\} \qquad (6)$$

defined as the expectation of the log-likelihood of the complete-data $\{\mathcal{O}, \mathcal{S}, \mathcal{K}\}$ given the incomplete-data $\mathcal{O}$ and the current parameter estimate θ. It can be shown [10] that maximization of the auxiliary function with respect to $\overline{\theta}$ leads to parameter estimates that increase the likelihood of the training data so that $\mathcal{L}(\mathcal{O}; \overline{\theta}) \geq \mathcal{L}(\mathcal{O}; \theta)$.

In the SAT paradigm, the state transition matrix A, the initial probability vector π, and the mixture component weights c_{ik} follow the standard EM re-estimation formulae [10]. Our development focuses on the re-estimation of the Gaussian mixture component parameters and the speaker transformations that deviate from the standard EM re-estimation. The auxiliary function with respect to the Gaussian densities can be written as

$$\mathcal{Q}_{\mathcal{N}}(\overline{\theta},\theta) = \sum_{r,t,i,k}^{R,T_r,N,K} \gamma_{ik}^{(r)}(t)\, \log \mathcal{N}(o_t^{(r)}; \overline{W}^{(r)}\overline{\mu}_{ik} + \overline{\beta}^{(r)}, \overline{\Sigma}_{ik}) \tag{7}$$

where $\gamma_{ik}^{(r)}(t)$ is the posterior probability that the observation $o_t^{(r)}$ was drawn according to the k-th mixture Gaussian component of the i-th state. These probabilities can be computed efficiently using the forward-backward algorithm [10]. The parameter vector θ consists of three sets of parameters: the speaker-specific transformations, the mean vectors and the covariance matrices of the Gaussian densities. Direct maximization of the auxiliary function via the gradient of $\mathcal{Q}_{\mathcal{N}}(\overline{\theta},\theta)$ with respect to the components of the parameter vector $\overline{\theta}$ results in a non-linear system of equations that requires expensive numerical optimization methods for its solution. We have, instead, employed a three-stage iterative scheme to approximate the optimum parameter vector, where at each stage we keep two sets of the parameters fixed and optimize with respect to the third set.

We first maximize the Q-function with respect to the speaker transformations while keeping the Gaussian parameters fixed to their current values. The optimum transformation parameters for each speaker are derived from the solution of

$$\frac{\partial \mathcal{Q}_{\mathcal{N}}(\overline{\theta},\theta)}{\partial \overline{G}^{(r)}} = 0 \tag{8}$$

It is shown in [8] that if the covariance matrices are diagonal the transformation parameters are derived in closed form. A detailed description of the solution is contained in [8].

We then compute the mean vectors of the Gaussians using the updated values of the transformation matrices while keeping the covariance matrices to their current values. The updated mean vectors are derived from the solution of

$$\frac{\partial \mathcal{Q}_{\mathcal{N}}(\overline{\theta},\theta)}{\partial \overline{\mu}_{ik}} = 0 \tag{9}$$

and for $i = 1, \ldots, N$ and $k = 1, \ldots, K$ are given by

$$\begin{aligned} \overline{\mu}_{ik} \quad = \quad & \left\{ \sum_{r,t}^{R,T_r} \gamma_{ik}^{(r)}(t) \overline{W}^{(r)\ T} \Sigma_{ik}^{-1} \overline{W}^{(r)} \right\}^{-1} \times \\ & \left\{ \sum_{r,t}^{R,T_r} \gamma_{ik}^{(r)}(t) \overline{W}^{(r)\ T} \Sigma_{ik}^{-1} \left(o_t^{(r)} - \overline{b}^{(r)} \right) \right\} \end{aligned} \tag{10}$$

Finally the re-estimation of the covariance matrices utilizes the updated values for the speaker-specific transformations and the Gaussian mean vectors in the auxiliary function. The solution of

$$\frac{\partial \mathcal{Q}_{\mathcal{N}}(\overline{\theta},\theta)}{\partial \overline{\Sigma}_{ik}} = 0 \tag{11}$$

gives the updated covariance matrices as

$$\overline{\Sigma}_{ik} = \frac{\sum_{r,t}^{R,T_r} \gamma_{ik}^{(r)}(t) \left(o_t^{(r)} - \overline{\mu}_{ik}^{(r)} \right) \left(o_t^{(r)} - \overline{\mu}_{ik}^{(r)} \right)^T}{\sum_{r,t}^{R,T_r} \gamma_{ik}^{(r)}(t)} \tag{12}$$

where $\overline{\mu}_{ik}^{(r)} = \overline{W}^{(r)}\overline{\mu}_{ik} + \overline{\beta}^{(r)}$ are the Gaussian means adapted to each speaker r using the updated values of the transformation parameters and the Gaussian mean vectors.

It is easily verified that at each stage of the update process the value of the auxiliary function is guaranteed to increase

$$\begin{aligned} \mathcal{Q}_{\mathcal{N}}\left(\boldsymbol{\mathcal{G}}, \{\mu_{ik}, \Sigma_{ik}\}_{i,k}\right) \quad & \leq \quad \mathcal{Q}_{\mathcal{N}}\left(\overline{\boldsymbol{\mathcal{G}}}, \{\mu_{ik}, \Sigma_{ik}\}_{i,k}\right) \\ & \leq \quad \mathcal{Q}_{\mathcal{N}}\left(\overline{\boldsymbol{\mathcal{G}}}, \{\overline{\mu}_{ik}, \Sigma_{ik}\}_{i,k}\right) \\ & \leq \quad \mathcal{Q}_{\mathcal{N}}\left(\overline{\boldsymbol{\mathcal{G}}}, \{\overline{\mu}_{ik}, \overline{\Sigma}_{ik}\}_{i,k}\right) \end{aligned} \tag{13}$$

Hence the likelihood of the training data is also guaranteed to increase, based on the properties of the auxiliary Q-function stated earlier. Typically one or two iterations of the outlined three-stage process are adequate to ensure convergence to an optimal point.

In the experiments that we conducted thus far, we used a sufficiently trained SI model as the initial seed to the SAT re-estimation process. The speaker-specific transformation matrices were initialized to the identity matrix and the additive bias vectors to zero for all training speakers.

4. EXPERIMENTAL RESULTS

In this section, we present the results of several recognition experiments to evaluate the efficacy of the SAT paradigm. These experiments were conducted using BYBLOS, BBN's state-of-the-art large vocabulary speech recognition system [11]. The baseline speaker-independent system is a gender dependent triphone-based continuous density HMM system. All allophone models of each of the 46 phonemes of the system are modeled by a mixture density of 256 Gaussian components in a configuration termed as Phonetically Tied Mixture (PTM) HMM. Speech is parameterized using 14 mel-warped cepstral coefficients, a short-term power coefficient and the first and second order difference of these parameters to give a 45 dimensional feature vector.

The acoustic training data consists of 62 hours of speech, collected from 284 speakers (male and female) from the SI-284 Wall Street Journal (WSJ) corpus. Following the common SI paradigm, we constructed SI gender-dependent acoustic models, to use as initial model seeds for the SAT re-estimation procedure that was outlined in Section 3. We evaluated the efficacy of the new algorithm by comparing the recognition performance of the SAT acoustic models to that of the SI acoustic models with and without adaptation to the test speakers. The testing material was drawn from the 1994 H1 and the 1994 S0 development tasks which are based on 20,000 and 5,000 words vocabulary tasks respectively. Each of the 20 test speakers included in each test provides 40 sentences of transcribed data that

are used for batch supervised adaptation of the acoustic models. The MLLR adaptation paradigm, using dynamically allocated number of regression classes, was used to adapt both the SI and the SAT acoustic models.

Test Set	Training Cond.	Word Error (%)	
		No Adapt.	Adapt.
H1-20K	SI paradigm	12.77	11.41
	SAT paradigm	12.81	10.40
S0-05K	SI paradigm	6.51	5.29
	SAT paradigm	6.51	4.80

Table 1: Word Error Rate (%) Comparisons

The results of the first column of Table 1 show that the un-adapted recognition performance of the SAT derived acoustic models is similar to that of the pooled SI acoustic models. This indicates that the signal variation that is removed from the acoustic models when we apply the SAT paradigm does not have any phonetic information content. Additionally, the results of the second column of Table 1 demonstrate the efficacy of the proposed method as the adapted SAT acoustic models perform consistently better than the corresponding adapted SI acoustic models in both tasks. The SAT acoustic models are able to adapt more accurately to the test speakers using little adaptation data (approximately 3 minutes of speech). Hence they achieve a significant reduction (almost a factor of 2) in word error rate over the adapted SI models and overall achieve 19% and 26% reductions in error rate for the 20k and 05k tasks respectively.

5. CONCLUSIONS

We have presented a novel formulation of the SI training paradigm that aims at reducing the overlap of the SI acoustic models that is caused by variation among the speakers of the training population. The method is based on explicitly accounting for the inter-speaker variation in the HMM parameter estimation process. A modified EM-based algorithm for the continuous density HMM parameter estimation is presented for the case that the speaker-specific characteristics are modeled by a linear transformation.

We have evaluated the proposed training algorithm in large vocabulary continuous speech recognition tasks. We conducted experiments that demonstrate that the SAT acoustic models are more efficiently adapted to the test speakers than common SI trained acoustic models. Experimental results show that SAT acoustic models achieve 10% additional reduction in word error rate relative to the SI acoustic models and result in overall reductions of up to 25% for batch supervised speaker adaptation.

Although the results considered in this paper focus on the extraneous variation caused by the speaker variation in the training population, the process of counteracting variations caused by other sources relies on the same algorithmic development that is described in this work. Specifically recording and channel conditions have been so far assumed stationary within the training data of any particular speaker. Within the proposed framework, we can annihilate variation effects within a smaller time window (such as a single utterance) than that defined by a speaker's activity. Additionally we are currently experimenting with different methods of initializing the SAT re-estimation process, as we believe that the convergence of the algorithm depends strongly on the initial seed.

6. ACKNOWLEDGMENTS

This work was supported by the Advanced Research Projects Agency and monitored by Ft. Huachuca under Contract Nos. DABT63-94-C-0061 and DABT63-94-C-0063. The content of the information does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

7. REFERENCES

1. T. Anastasakos, F. Kubala, J. Makhoul, and R. Schwartz, "Adaptation to new microphones using tied-mixture normalization", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1994, pp. 433–436.
2. H. Eide and H. Gish, "A parametric approach to vocal tract length normalization", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1996, pp. 346–349.
3. J.R. Bellegarda, P.V. de Souza, A. Nádas, D.Nahamoo, M.A. Picheny, and L.R. Bahl, "The Metamorphic Algorithm: A Speaker Mapping Approach to Data Augmentation", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 413–420, July 1994.
4. Y. Zhao, "An Acoustic-phonetic-based Speaker Adaptation Technique Improving Speaker-independent Continuous Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 380–394, July 1994.
5. M. Rahim and B-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 19–30, January 1996.
6. C.J. Leggetter and P.C. Woodland, "Flexible Speaker Adaptation for Large Vocabulary Speech Recognition", in *Proc. Eurospeech*, 1995, pp. 1155–1158.
7. A. Sankar and C-H. Lee, "Stochastic Matching for Robust Speech Recognition", *IEEE Signal Processing Letters*, vol. 1, no. 8, pp. 124–125, August 1994.
8. C.J. Leggetter and P.C. Woodland, "Speaker Adaptation of HMMs Using Linear Regression", Tech. Rep. CUED/F-INFENG/TR.181, Cambridge University Engineering Department, June 1994.
9. A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of Royal Statistical Society*, vol. B 39, pp. 1–38, 1977.
10. L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", *Ann. Math. Stat.*, vol. 41, pp. pp. 164–171, 1970.
11. L. Nguyen et. al, "The 1994 BBN/BYBLOS Speech Recognition System", in *Proc. SLS Technology Workshop*. 1995, pp. 77–81, Morgan Kaufmann Publishers.