SPEAKER ADAPTIVE TRAINING: A MAXIMUM LIKELIHOOD APPROACH TO SPEAKER NORMALIZATION

†*Tasos Anastasakos* *John McDonough* *John Makhoul*

BBN Corporation
70 Fawcett Street, Cambridge, MA 02138, USA
†Northeastern University
Boston, MA 02115, USA
E-mail: tasos@lexicus.mot.com

ABSTRACT

This paper describes the speaker adaptive training (SAT) approach for speaker independent (SI) speech recognizers as a method for joint speaker normalization and estimation of the parameters of the SI acoustic models. In SAT, speaker characteristics are modeled explicitly as linear transformations of the SI acoustic parameters. The effect of inter-speaker variability in the training data is reduced, leading to parsimonious acoustic models that represent more accurately the phonetically relevant information of the speech signal. The proposed training method is applied to the Wall Street Journal (WSJ) corpus that consists of multiple training speakers. Experimental results in the context of batch supervised adaptation demonstrate the effectiveness of the proposed method in large vocabulary speech recognition tasks and show that significant reductions in word error rate can be achieved over the common pooled speaker-independent paradigm.

1. INTRODUCTION

Current speaker independent (SI) continuous speech recognition (CSR) systems achieve a certain degree of robustness in recognition performance by estimating their parameters on tens of hours of speech collected from multiple speakers and in various recording environments. An inherent difficulty of this approach is that the resulting statistical models have to contend with a wide range of variation in the speech signal caused not only by phonetically relevant variation sources but also by inter-speaker variability. The spectral distributions often exhibit high variance and hence high overlap among different speech units, which may result in diffused acoustic models with reduced discriminatory capabilities. In addition, a large number of parameters is required for sufficient modeling accuracy of the speech variability.

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 subtraction [3] has been the simplest feature-based normalization method that is used mainly to counteract channel effects. In [8], 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. In [13], an acoustic normalization technique within the framework of mixture density HMMs is applied to normalize the training as well as the test data, and in [11], a maximum likelihood signal bias is estimated jointly with the parameters of a discrete HMM.

This paper considers the *Speaker Adaptive Training* (SAT) algorithm [4] that provides a unified framework for speaker normalization and parameter estimation of HMM-based speech recognizers. In the SAT method, the individual speaker characteristics are modeled by linear transformations of the mean parameters of the acoustic models. Model-based linear transformations offer an effective way for parameter tying and correlation among the different phonetic units. They have been applied successfully to speaker adaptation methods that try to improve the recognition performance of SI systems for a test speaker using little speaker-specific data [7, 9, 12]. In this work, the speaker transformations are integrated in the training phase [4]. The proposed training method is based on a maximum likelihood formulation that *jointly* estimates the HMM acoustic parameters and the speaker transformations. By accounting explicitly for the extraneous speaker-induced variation and reducing its effect in the training data, the resulting acoustic models are truly speaker independent with reduced cross-unit overlap. An approach similar to SAT has been developed contemporaneously in [1].

The SAT algorithm is compared to the common SI training paradigm within the context of supervised adaptation. The proposed acoustic models are shown to adapt to the test speakers more efficiently, thus achieving significant overall word error rate reductions of up to 25% for native speakers of American English and over 50% reduction in word error rate for non-native speakers of American English.

2. DESCRIPTION OF SAT ALGORITHM

The SAT formulation is based on an underlying generative process of speech that consists of two distinct components. The first component represents the variation source of phonetically relevant speech events that is considered independent of any particular speaker. The phonetic variation source is realized by the set of HMM-based acoustic models $\boldsymbol{\lambda}$. In Fig. 1, a 3-state HMM generates the phonetic sequence $\{\boldsymbol{x}_t, \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+2}\}$ that is speaker-independent in the sense that it represents a sequence of phonetic events without speaker-specific characteristics.

The second component of the proposed composite process captures the speaker specific attributes of speech such as variations due to different accents and regional dialects and physiological characteristics such as pitch, vocal tract length, the general anatomy of the vocal cavity, age and gender. This second component is represented as a filter that transforms the phonetic events to speech produced by a particular speaker by assigning the speaker-specific attributes that this filter is capable of modeling. As illustrated in Fig. 1, the sequence of speaker-independent phonetic events is transformed to the speaker-dependent sequence of obser-

vations $\{\boldsymbol{o}_t, \boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}\}$ by passing through the speaker-specific filter.

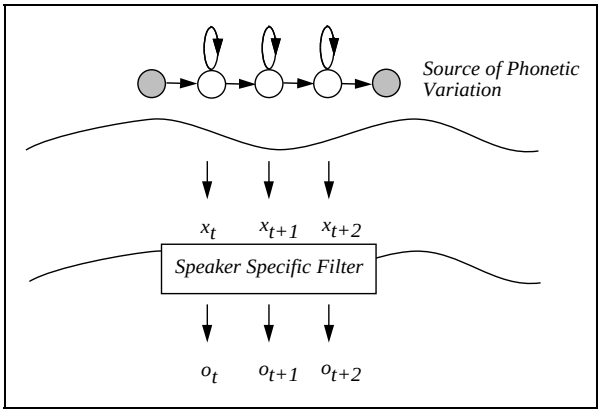

Figure 1. SAT motivated model of speech generation for a particular speaker

The speaker specific characteristics are modeled by linear-regression transformations that map the speaker-independent Gaussian mean vector $\boldsymbol{\mu}_j$ to an estimate of the speaker-dependent mean $\boldsymbol{\mu}_j^{(r)}$ that pertains to speaker r according to

$$\boldsymbol{\mu}_j^{(r)} = \boldsymbol{A}^{(r)}\boldsymbol{\mu}_j + \boldsymbol{\beta}^{(r)} \tag{1}$$

where $\boldsymbol{A}^{(r)}$ is a full matrix and $\boldsymbol{\beta}^{(r)}$ is an additive vector that comprise the speaker specific transformation $\boldsymbol{G}^{(r)}$.

The SAT algorithm begins by partitioning the training data according to the variation source whose effects need to be normalized. In this case, the training data are partitioned according to speaker and a transformation is hypothesized for each speaker to account for the particular speaker individuality. The optimum set of HMM parameters $\widetilde{\boldsymbol{\lambda}}$ and the set of speaker transformations $\widetilde{\boldsymbol{\mathcal{G}}} = (\widetilde{\boldsymbol{G}}^{(1)}, \ldots, \widetilde{\boldsymbol{G}}^{(R)})$ are jointly estimated so as to maximize the likelihood of the training data

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

where $O^{(r)}$ is the observation sequence contributed by speaker r and R is the total number of training speakers.

The proposed formulation allows the employment of a wide range of linear transformations. In this sense, it provides a general framework that extends previous work on normalization by means of cepstrum mean subtraction and maximum likelihood additive bias [1, 8, 12]. In the current work, the transformations are assumed to be "full" regression matrices following the Maximum Likelihood Linear Regression (MLLR) approach [9]. A simplified form is obtained if the matrices are assumed to be diagonal or identity (in the latter case the linear transformation reduces to an additive bias). However, it has been found [9] that full matrices give superior performance and hence all experiments reported in this paper assume the use of full regression matrices.

The SAT parameter estimation is based on the EM algorithm [5, 6] by defining the appropriate auxiliary function. The HMM state transition probabilities and the mixture component weights follow the standard EM estimation formulae. An iterative optimization scheme is employed for the calculation of optimal values for the set of speaker-dependent transformations, the set of speaker-independent Gaussian mean vectors and the set of the corresponding Gaussian variances. Optimal values for one such set of parameters are obtained while the other sets are held constant. The joint optimum is approximated by iterating over every parameter set. Typically, one or two iterations of the outlined optimization scheme are adequate to ensure convergence to an optimal point.

The speaker-dependent transformations are estimated according to the MLLR approach [9]. The estimation of the means of the Gaussian densities conditioned on the speaker-specific linear regression transformations is expressed as

$$\begin{aligned}\widetilde{\boldsymbol{\mu}}_k \;=\; & \left\{\sum_{r,t}^{R,T_r} \gamma_k^{(r)}(t)\, \widetilde{\boldsymbol{A}}^{(r)\,T} \boldsymbol{\Sigma}_k^{-1} \widetilde{\boldsymbol{A}}^{(r)}\right\}^{-1} \times \\ & \left\{\sum_{r,t}^{R,T_r} \gamma_k^{(r)}(t)\, \widetilde{\boldsymbol{A}}^{(r)\,T} \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{o}^{(r)}(t) - \widetilde{\boldsymbol{\beta}}^{(r)}\right)\right\}\end{aligned} \tag{3}$$

where $\boldsymbol{\Sigma}_k$ is the covariance matrix of the k-th Gaussian density, and $\gamma_k^{(r)}(t)$ is the posterior probability that the observation $\boldsymbol{o}_t^{(r)}$, generated by the r-th speaker, was drawn according to the k-th Gaussian density. Similarly, the estimation of the covariance matrices of the Gaussian densities is expressed as

$$\widetilde{\boldsymbol{\Sigma}}_k = \frac{\displaystyle\sum_{r,t}^{R,T_r} \gamma_k^{(r)}(t) \left(\boldsymbol{o}_t^{(r)} - \widetilde{\boldsymbol{\mu}}_k^{(r)}\right)\left(\boldsymbol{o}_t^{(r)} - \widetilde{\boldsymbol{\mu}}_k^{(r)}\right)^T}{\displaystyle\sum_{r,t}^{R,T_r} \gamma_k^{(r)}(t)} \tag{4}$$

It should be noted that the above development assumes the use of a single linear transformation for each speaker. In general, the modeling accuracy is improved by allowing several transformations. In this case, each Gaussian component is assigned to one such transformation and the set of Gaussians that share the same transformation is referred to as a *regression class*. The SAT estimation algorithm and the extension of SAT to multiple regression classes is covered in greater detail in [4, 2].

3. EXPERIMENTAL EVALUATION

The development of the SAT method is predicated upon the fact that acoustic models with reduced cross-unit overlap will adapt to the test conditions more efficiently than the SI acoustic models, which are commonly estimated by *pooling* the training data. Furthermore, the variation that is being modeled by the transformation in the SAT should be compensated during recognition in order to match more accurately the test speaker characteristics. Thus the SAT algorithm is evaluated on the Wall Street Journal (WSJ) corpus, on recognition tests that incorporate speaker adaptation for the test speakers.

3.1. Baseline SI System

The baseline speaker independent system that is used in the experiments is a gender-dependent, cross-word triphone, mixture Gaussian HMM system. 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. Decision tree based state clustering provides a flexible mechanism for tying of the Gaussian densities that leads to two configurations [10]: (i) the *Phonetically Tied Mixture* (PTM) system, where all the allophones of each of the 46 phonemes of the system are modeled by a set of 256 Gaussian densities (a total of 11,776 Gaussian densities), and (ii) the *State Clustered Tied Mixture* (SCTM) system, where each of 3,000 clustered states is modeled by a set of 64 Gaussians (a total of 192,000 Gaussian densities). The SAT estimation uses the pooled SI acoustic models as initial model seed and MLLR transformations with dynamically allocated multiple regression classes.

The acoustic training data consist of 62 hours of speech, collected from 284 speakers (male and female) from the SI-284 portion of the WSJ corpus. Experiments were conducted on three test sets from the development material of the 1994 ARPA CSR evaluation: the H1D94 test that contains sentences of 20K word vocabulary spoken by 20 native speakers of American English, the S0D94 test that contains sentences of 5K word vocabulary spoken by the same 20 native speakers, and the S3D94 test that contains sentences of 5K word vocabulary spoken by 11 non-native speakers of American English.

3.2. Adaptation on the Test

Each test speaker provides 40 enrollment sentences (approximately 3 minutes of speech) for supervised off-line (batch) adaptation. The existing acoustic models are adapted to each test speaker using the available enrollment speech and MLLR adaptation with tree-based regression classes similar to [9].

The performance of the SAT acoustic models is compared to that of the pooled SI models in recognition experiments with and without adaptation to the test speakers. Table 1 shows the comparative results of the two training methods in combination with the use of adaptation on the test. When adaptation of the test speakers

Test Set	Training Cond.	% WER	
		No Adapt.	Adapt.
H1-20K	SI paradigm	12.71	11.4
	SAT paradigm	12.81	10.40
S0-05K	SI paradigm	6.51	5.27
	SAT paradigm	6.47	4.82
S3-05K	SI paradigm	21.45	12.35
	SAT paradigm	24.20	11.73

Table 1. Word Error Rate (% WER) using the PTM configuration for the construction of the pooled-SI and the SAT acoustic models.

is not applied, the average performance of the two paradigms over all test speakers is very similar for the two native speakers tests. This result suggests that the speaker-induced signal variation that is removed from the acoustic models in the SAT paradigm does not contain significant phonetic information. In the S3D94 test however, the training corpus is not representative of the test speakers and the acoustic models need to be smoother than in the native speakers tests. The SAT acoustic models are sharper by construction due to speaker normalization, thus increasing the recognition word error rate in the S3D94 test. When speaker-adaptation is applied on the test, the SAT paradigm achieves a significant reduction of 10% in word error rate over the SI paradigm for the case of native speakers. In all three test sets the proposed training algorithm performs uniformly better than the SI paradigm.

3.3. Parsimonious Modeling

The modeling resolution of the speaker independent acoustic models can be significantly improved at the cost of increasing the parameters of the HMMs. In order to represent accurately the spectral variation in the speech signal due to multiple training speakers and recording conditions, a large number of Gaussian densities is required. The aim of the SAT method is to reduce the extraneous inter-speaker variability from the training data and generate acoustic models that achieve the same modeling efficiency for the relevant phonetic events with fewer parameters than needed to model the unnormalized training data.

Table 2 shows the recognition results of the two different approaches to constructing SI acoustic models. The SI SCTM-64

Acoustic Models	Adapt. on Test	% WER	
		H1D94	*S0D94*
Pooled SI PTM-256	×	12.71	6.51
11,776 Gaussians	√	11.44	5.27
Pooled SI SCTM-64	×	11.52	5.77
192,000 Gaussians	√	10.80	4.92
SAT PTM-256	×	12.81	6.47
11,776 Gaussians	√	10.40	4.82

Table 2. Recognition results of PTM-256, SCTM-64 and SAT-PTM-256. The performance of the three systems when adaptation on the test is applied demonstrates the ability of the SAT formulation to increase recognition performance and to provide an efficient and parsimonious representation of the acoustic space.

system achieves a lower error rate than the SI PTM-256 system both without and with adaptation on the test at the cost of a significant increase in the number of Gaussian densities. When adaptation on the test is employed, which is the condition of interest, the SAT PTM-256 acoustic models demonstrate the advantage of speaker normalization, whereby the recognition performance exceeds that of the SI SCTM-64 system, while maintaining the same number of Gaussians as the SI PTM-256. When adaptation on the test is not applied, the performance of the SAT acoustic models is similar to that of the PTM-256 models as it was also shown in Table 1.

3.4. Acoustic Models with Reduced Variance

To measure the overlap among the Gaussians of a model, we compute the average covariance of the total set of Gaussian densities as the weighted average

$$\overline{\mathbf{\Sigma}} = \left[\overline{\sigma}_d^2\right] = \frac{1}{\sum_{i=1}^{46}\sum_{j=1}^{256} N_{ij}} \sum_{i=1}^{46}\sum_{j=1}^{256} N_{ij}\,\mathbf{\Sigma}_{ij} \qquad (5)$$

where the relative frequencies N_{ij} are the mixture component weights. Eq. (5) computes the average covariance of a PTM system. Based on this measure, we observe the higher overlap of the SI distributions relative to the SAT distributions, as indicated in Fig. 2 by the plot of $\overline{\sigma}_d^2$ with respect to all feature dimensions.

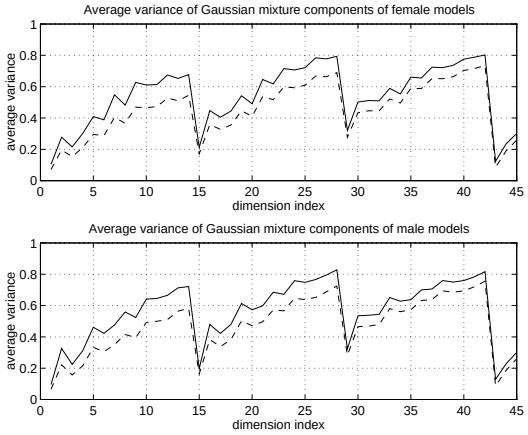

Figure 2. Average per-feature variance of the SI (solid line) and SAT (dashed line) mixture component densities. The feature vector consists of the 14 cepstral, their first and second order time differences, and the log-energy its first and second time difference.

4. CONCLUSIONS

This paper has described a unified approach to speaker normalization and training of speaker independent acoustic models within the maximum likelihood estimation framework. The method allows the use of general linear transformations for modeling of speaker characteristics during training. This formulation reduces the speaker-induced variability that is present in the training data thereby constructing truly speaker-independent acoustic models. Experimental results show that this novel training paradigm formulation can lead to a parsimonious representation of the acoustic space and at the same time achieve significant reductions in word error rate over the pooled speaker-independent training paradigm.

While the focus of this paper was on the normalization of the inter-speaker variability in the training data, the technique can also be used to normalize the effects of recording environment conditions (e.g. ambient noise, microphone, communication channel). We currently focus on applications of the SAT approach to problems of normalization of other extraneous variations and on investigating different parametric forms of transformation within the SAT framework.

REFERENCES

[1] A. Acero and X. Huang, "Speaker and gender normalization for continuous-density hidden Markov models", in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 342–345.

[2] T. Anastasakos, *Speaker normalization methods for speaker independent speech recognition*, PhD thesis, Northeastern University, 1996.

[3] T. Anastasakos, F. Kubala, J. Makhoul, and R. Schwartz, "Adaptation to new microphones using tied-mixture normalization", in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994, pp. 433–436.

[4] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training", in *Proceedings of International Conference in Spoken Language Processing*, 1996.

[5] 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.

[6] 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.

[7] V. Digalakis, D.Rtichev, and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, September 1995.

[8] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization", in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 346–349.

[9] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[10] L. Nguyen, T. Anastasakos, F. Kubala, C. Lapre, J. Makhoul, R. Schwartz, N. Yuan, and G. Zavaliagkos, "The 1994 BBN/BYBLOS speech recognition system", in *Proc. SLS Technology Workshop*. 1995, pp. 77–81, Morgan Kaufmann Publishers.

[11] M. Rahim and B-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 19–30, January 1996.

[12] 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.

[13] Y. Zhao, "An acoustic-phonetic-based speaker adaptation technique improving speaker-independent continuous speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 380–394, July 1994.