EIGENSPACE-BASED MLLR WITH SPEAKER ADAPTIVE TRAINING IN LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION

Vlasios Doumpiotis,Yonggang Deng

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218, USA
{*vlasios,dengyg*}*@jhu.edu*

ABSTRACT

In this paper, Speaker Adaptive Training(SAT) which reduces inter-speaker variability and Eigenspace-based Maximum Likelihood Linear Regression (EigenMLLR) adaptation, which takes advantage of prior knowledge about the test speaker's linear transforms, are combined and developed. During training, SAT generates a set of speaker independent (SI) Gaussian parameters, along with matched speaker dependent transforms for all the speakers in the training set. Then a set of regression class dependent Eigen transforms are derived by doing Singular Value Decomposition (SVD). Normally during recognition the test speaker's linear transforms are obtained with MLLR. In this work, the test speaker's linear transforms are assumed to be linear combination of the decomposed Eigen transforms. Experimental results conducted on large vocabulary conversational speech(LVCSR) material from the *Switchboard Corpus* show that this strategy has better performance than ML-SAT and significantly reduces the number of parameters needed(an 87% reduction is achieved), while still effectively capturing the essential variation between speakers.

1. INTRODUCTION

Although typical state-of-the-art large vocabulary conversational speech recognition (LVCSR) systems achieve high performance, these systems can be improved upon by adapting the models to the characteristics of a particular speaker using a small amount of adaptation or enrollment data. Adaptation is very important to compensate for the differences between the speech on which an ASR system was trained and the speech which it has to recognize. The most popular model-based adaptation techniques can be grouped into three families depending on the application[1]: Maximum a posteriori (MAP) family, linear transformation family including MLLR [2], and speaker clustering based family including CAT[3] and eigenvoice [4, 5].

In MLLR, a transform is applied to the Gaussian model parameters in the estimation of the state independent observation distributions in order to match the specific conditions of interest and has been shown to be effective in improving the performance of speaker independent (SI) LVCSR systems by adapting the system to the test set. Adaptation can also be applied to the speakers in the training set to produce matched conditions with the test set, and this is termed Maximum Likelihood (ML) Speaker Adaptive Training (SAT) [6]. The goal of SAT is to reduce inter-speaker variability within the training set. SAT is an iterative estimation procedure that generates a set of speaker independent (SI) Gaussian parameters along with matched speaker dependent transforms for all the speakers in the training set using MLLR.

Basically there are two forms of adaptation: supervised and unsupervised. In supervised adaptation the true transcription of the data is known and in unsupervised adaptation no reference transcription is provided but it is hypothesized. This initial hypothesis may contain errors which makes it difficult to reliably estimate large number of parameters. Moreover, supervised techniques usually perform better.

Thus in speaker adaptation(MLLR, ML-SAT) although we can estimate a large number of transforms for any of the training speakers since we have the correct transcription and adequate amount of enrollment data, this is not the case for the test speakers (unsupervised adaptation with few data). Therefore the number of transforms that can be reliably estimated is limited(usually no more than 2 transformation matrices). Furthermore the prior knowledge about speaker variants from the training set, typically doesn't assist in the testing stages. To alleviate the problem of reliably estimating model parameters when there is only a small amount of adaptation data available, Kunh et al[4] proposed an 'Eigenvoice' approach which incorporates prior knowledge and requires a set of speaker dependent models. Chen et al[5] introduced a rapid speaker adaptation scheme, termed Eigenspace-based Maximum Likelihood Linear Regression(Eigen-MLLR).

In this work, we investigated the Eigenspace based MLLR adaptation framework along with SAT. Given that SAT is an estimation procedure, aiming at reducing the inter-speaker variability within the training set and Eigen-MLLR incorporates prior information about the transforms from the training set we expect the integration of these two techniques to yield improved performance since they capture different acoustic phenomena. Thus with SAT we generate a set of speaker independent (SI) Gaussian parameters along with matched speaker dependent transforms and improve performance through successive iterations of parameter estimation. We then incorporate EigenMLLR adaptation after SAT training. Using SVD, a set of 'canonical' transforms, termed Eigen transforms were obtained. During testing, new speaker's linear transforms are a linear combination of those Eigen transforms, rather than obtained with MLLR. Since the number of linear coefficients is much less than the number of parameters in transform matrices, the approach only requires a small amount of adaptation data for a robust estimation. Furthermore we can use more than 2 transforms for each test speaker.

This work was supported by the National Science Foundation under grants No. #IIS-9982329 and No. #IIS-0122466.

2. SPEAKER ADAPTIVE TRAINING

Speaker Adaptive Training (SAT) [6] has been shown to be effective in improving the performance of speaker independent (SI) LVCSR systems. For each speaker, a transformation matrix is usually applied to the mean vector of each Gaussian, because they define major characteristics of the distributions. Covariance adaptation is less commonly used and its effects are less profound than the mean adaptation[7].

Under this model the emission density of state s is reparametrized for each speaker $k = 1, 2, ..., N$ as

$$q(o_\tau|s,k;\theta) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_s|}} e^{-\frac{1}{2}\left(o_\tau - T_r^{(k)}\xi_s\right)^T \Sigma_s^{-1}\left(o_\tau - T_r^{(k)}\xi_s\right)}$$

and we have N speakers in the training set.

To avoid introducing more parameters than can be reliably estimated, transformations are tied across sets of states. Here, $T_r^{(k)}$ is the extended speaker dependent transformation matrix $[b_r^{(k)}\ A_r^{(k)}]$ associated with a group of states $S_r = \{s|\mathcal{R}(s) = r\}$ for classes $r = 1, \ldots, R$ and ξ_s is the extended mean vector $[1\ \mu_s^T]^T$. The function S_r gives a set of mixtures belonging to the same regression class r.

Since the training data are collected from a population of N speakers, all utterances are partitioned according to speaker identity. To incorporate information about the speaker identities, we denote by $\{\tau : \hat{k}_\tau = k\}$, the sequence of feature vectors o_τ belonging to speaker k. The augmented state dependent parameter set is defined as $\theta = (T_r^{(k)}, \mu_s, \Sigma_s)$, for all speakers k. Our objective is to compute the speaker dependent transforms and speaker independent Gaussian parameters of the state dependent distributions under the ML criterion. This is done by maximizing the following auxiliary function:

$$Q(\bar{\theta}|\theta) = -\frac{1}{2}\sum_{k,r}\sum_{s\in S(r)}\sum_{\tau:\hat{k}_\tau=k}\gamma_s(\tau;\theta)\big[log|\Sigma_s| + (o_\tau - \bar{\mu}_s^{(k)})^T\Sigma_s^{-1}(o_\tau - \bar{\mu}_s^{(k)})\big] + C \quad (1)$$

where C is constant independent on θ coefficients and o_τ is the adaptation data. The parameter update equation is:

$$\bar{\theta} : \sum_{k,r}\sum_{s\in S_r}\sum_{\tau:\hat{k}_\tau=k}\gamma_s(\tau;\theta)\nabla_\theta \log q(o_\tau|s,k;\ \bar{\theta}) = 0\ . \quad (2)$$

where we define $\gamma_s(\tau;\theta) = q_{s_\tau}(s|\hat{w}_1^n, o_1^l, k;\theta)$ is the conditional occupancy probability of state s at time τ given the training acoustics and the reference transcription $\hat{w}_1^n$.

2.1. Estimation of SAT Transforms

With the HMM parameters fixed, the parameter update relationship of equation (2) can be expressed as:

$$\bar{T}_r^{(k)} : \sum_{s\in S_r}\sum_{\tau:\hat{k}_\tau=k}\gamma_s(\tau;\theta)\cdot\nabla_{T_r^{(k)}}\log q(o_\tau|s,k;\bar{T}_r^{(k)},\mu_s,\Sigma_s) = 0\ . \quad (3)$$

The gradient of logarithm of the emission density q with respect to $T_r^{(k)}$ can be found as

$$\nabla_{T_r^{(k)}}\log q(o_\tau|s,k;\theta) = \Sigma_s^{-1}\left(o_\tau - T_r^{(k)}\xi_s\right)\xi_s^T$$

Substituting this into equation (3) it follows that the new transform estimates $\bar{T}_r^{(k)}$ should satisfy:

$$\sum_{s\in S_r}\Sigma_s^{-1}\sum_{\tau:\hat{k}_\tau=k}\gamma_s(\tau;\theta)o_\tau\xi_s^T = \sum_{s\in S_r}\sum_{\tau:\hat{k}_\tau=k}\gamma_s(\tau;\theta)\Sigma_s^{-1}\bar{T}_r^{(k)}\xi_s\xi_s^T \quad (4)$$

Here, the state occupancies $\gamma_s(\tau;\theta)$ are found via counts accumulated for each speaker under the initial parameters $(T_r^{(k)}, \mu_s, \Sigma_s)$.

2.2. Gaussian Parameter Estimation

The state independent Gaussian mean and variance parameters for ML-SAT are estimated under the ML criterion (2), using the updated values of the speaker dependent affine transforms $\bar{T}_r^{(k)}$ (4). The parameter set is $\tilde{\theta} = (\bar{T}_r^{(k)}, \mu_s, \Sigma_s)$. The derivation of the update formulas involves the gradient of the reparametrized emission density with respect to μ_s and Σ_s^{-1}. Subsequently we solve for $\bar{\mu}_s$ and $\bar{\Sigma}_s$.

For brevity we provide the final update equations where the speaker independent means are reestimated as

$$\bar{\mu}_s = \left(\sum_k\sum_{\tau:\hat{k}_\tau=k}\gamma_s(\tau;\tilde{\theta})\bar{A}_r^{(k)T}\Sigma_s^{-1}\bar{A}_r^{(k)}\right)^{-1}\times \sum_k\bar{A}_r^{(k)T}\Sigma_s^{-1}\sum_{\tau:\hat{k}_\tau=k}\gamma_s(\tau;\tilde{\theta})\left(o_\tau - \bar{b}_r^{(k)}\right). \quad (5)$$

The speaker independent variances are reestimated as

$$\bar{\Sigma}_s = \frac{\sum_k\sum_{\tau:\hat{k}_\tau=k}\gamma_s(\tau;\tilde{\theta})(o_\tau^2 - 2o_\tau\bar{\mu}_s^{(k)} + \bar{\mu}_s^{(k)2})}{\sum_k\sum_{\tau:\hat{k}_\tau=k}\gamma_s(\tau;\tilde{\theta})} \quad (6)$$

where $\bar{\mu}_s^{(k)} = \bar{A}_r^{(k)}\bar{\mu}_s + \bar{b}_r^{(k)}$, are the new speaker dependent means.

This derivation describes a two-stage, iterative procedure. Initially, speaker dependent transforms are estimated via equation (4), after which speaker independent Gaussian parameters are found via equation (5) and equation (6).

3. INTEGRATION OF SAT AND EIGENSPACE DECOMPOSITION

This section describes how we combine the SAT and EigenMLLR procedures. During SAT, we find each speaker's linear transforms $\bar{T}_r^{(k)}$ from the SI model using MLLR (4) and these transforms are D dimensional. For each regression class r, a supervector S_i for each speaker is composed and a large $D \times N$ matrix Ξ is formed, with columns corresponding to speakers, and rows to the parameters in the speaker's transforms.

The next step is to reduce this very large matrix into a compressed matrix using singular value decomposition(SVD). By doing SVD according to:

$$\Xi = U \cdot S \cdot V^T, \quad (7)$$

N Eigen transforms ordered by eigenvalue are obtained from the basis $U \cdot S^T$. Typically the first few K Eigen transforms capture most of the variation in the data. Decreasing K, the number of dimensions retained, reduces the accuracy with which Ξ can be

recreated from its component matrices(usually, $K < N < D$).

Suppose that K Eigen transforms for each regression class r have been obtained in training stage: we denote them by $W_k^{(r)}, k = 1, 2, \ldots, K$. The basic problem then becomes how to exploit the adequate amount of data in the training set, in order to obtain robust estimates for the test speakers' transforms. To achieve this we employ prior knowledge of what the estimates might be. Thus during recognition the adapted transform for a new test speaker is constrained to be located in the space spanned by those K Eigen transforms according to:

$$\hat{W}^{(r)} = \sum_{k=1}^{K} \lambda_k^{(r)} W_k^{(r)} \ . \tag{8}$$

Finally the transformed test speaker mean is given by:

$$\hat{\mu}_s = \hat{W}^{(r)} \xi_s = \sum_{k=1}^{K} \lambda_k^{(r)} W_k^{(r)} \xi_s \ . \tag{9}$$

The maximum likelihood estimation for the parameters $\lambda_k^{(r)}$ is done by using the auxiliary function (1). Since the Eigen transforms are orthogonal to each other, it is assumed that the λ coefficients to be calculated are independent to each other. By substituting equation (9) into the auxiliary function (1) and setting the partial derivatives w.r.t. $\lambda_k^{(r)}$ to zero, the following equations are obtained, ignoring all terms independent of $\lambda_k^{(r)}$:

$$Z_k^{(r)} = \sum_{j=1}^{K} \lambda_j^{(r)} Y_{k,j}^{(r)}, \qquad k = 1, 2, \ldots, K \tag{10}$$

where the accumulators $Z_k^{(r)}$ and $Y_{k,j}^{(r)}$ are given by

$$Z_k^{(r)} = \sum_{s \in S(r)} \sum_{\tau=1}^{T} \gamma_s(\tau; \theta) o_\tau^T \Sigma_s^{-1} (W_k^{(r)} \xi_s) \tag{11}$$

$$Y_{k,j}^{(r)} = \sum_{s \in S(r)} \sum_{\tau=1}^{T} \gamma_s(\tau; \theta) (W_k^{(r)} \xi_m)^T \Sigma_m^{-1} (W_j^{(r)} \xi_m) \ . \tag{12}$$

It's obvious that matrix $Y^{(r)} = (Y_{k,j}^{(r)})$ is symmetric. And the K transform weights are obtained by solving K linear equations $Z^{(r)} = Y^{(r)} \lambda^{(r)}$. The computational requirement is similar to standard MLLR, plus the overhead in estimating (10).

4. EXPERIMENTAL RESULTS

4.1. System Description

The experimental results in this section are conducted on material from the *Switchboard Corpus* which is a database of spontaneous dialogue, with no pre-selected topic, among English speakers. The system is a speaker independent continuous mixture density, tied state, cross-word, gender-independent, triphone HMM system with $N = 209$ speakers. The baseline acoustic models used as seed models for our experiments were built using HTK [8] from 16.4 hours of Switchboard-1 and 0.5 hour of Callhome English data. This collection defined the development training set for the 2001 JHU LVCSR system [9]. We have used only 17 hours of data from 209 speakers for computational reasons, we expect similar behavior with bigger systems. The speech was parameterized into 39-dimensional PLP cepstral coefficients with delta and acceleration components. Cepstral mean and variance normalization was performed over each conversation side.

The acoustic models used cross-word triphones with decision tree clustered states [8], where questions about phonetic context as well as word boundaries were used for clustering. There were 4000 unique triphone states with 6 Gaussian components per state. To define the regression classes and assign the Gaussians, we employed the HTK regression class tree implementation [8].

Lattice rescoring experiments were performed using the AT&T Large Vocabulary Decoder [10], with a 33k-word trigram language model provided by SRI [11]. The recognition tests were carried out on a subset of the 2000 Hub-5 Switchboard-1 evaluation set (SWBD1) and the 1998 Hub-5 Switchboard-2 evaluation set (SWBD2). The SWBD1 test set was composed of 866 utterances consisting of 10260 words from 22 conversation sides, and the SWBD2 test set was composed of 913 utterances consisting of 10643 words from 20 conversation sides. The total test set was 2 hours of speech.

Our system was seeded from a well trained MMIE model. Our approach is based on the MMI training procedure developed by Woodland and Povey [12], but we used triphone lattices on the training data.

4.2. Eigen-MLLR performance without SVD

Initially we conducted a series of experiments to compare MLLR and Eigen-MLLR with different number of transforms. These experiments are shown in Table 1. MLLR and Eigen-MLLR were performed with a MMIE trained model (39.9%,49.7%). We also keep all the 'eigen-values' thus the size of vector $\lambda^{(r)} = [\lambda_k^{(r)}]$ is $K = N = 209$ parameters for Eigen-MLLR for each regression class r. The purpose of this experiment is to state the baseline and compare Eigen-MLLR and standard MLLR for different number of transforms without applying SVD.

Using multiple regression classes with MLLR, resulted in suboptimal performance which is not surprising given the unsupervised nature of the adaptation, the high word error rate and the large number of parameters that have to be estimated given our limited adaptation data. On the other hand in Eigen-MLLR, the speaker's linear transform is a linear combination of the decomposed Eigen transforms computed in the training set. As the number of transforms increases the WER does not increase, which is a consequence of integrating prior knowledge about the transforms from the training set(enough data, low error rate, supervised mode). We get our best result with 4 transforms (35.9%,46.0%). In parentheses the number of parameters estimated for each test speaker is shown under both methods. Since we are interested in the number of parameters used, we present the number of transforms rather than thresholds for regression classes trees

	MLLR		Eigen-MLLR K=209	
#TRANS	SWBD1	SWBD2	SWBD1	SWBD2
2	36.1(2*1600)	46.8	36.6(2*209)	46.6
4	37.0(4*1600)	47.9	35.9(4*209)	46.0
6	38.0(6*1600)	49.1	35.7(6*209)	46.4

Table 1. Word Error Rate (%) of systems with MLLR, Eigen-MLLR evaluated on Swbd1 and Swbd2 test sets.

4.3. Eigen-SAT performance

Here we investigate the SAT and EigenMLLR procedures as described in the previous section. Table 2 shows the performance of the ML-SAT and Eigen-SAT estimation procedures initialized with a MMIE trained model. In this implementation of ML-SAT, the transformation parameters and the Gaussian mean and variance parameters, are estimated at each iteration via Baum-Welch procedure, over the transcribed training data. We have selected 2 and 4 regression classes based on the results in Table 1. We have done 5 iterations of Speaker Adaptive Training. Again as in MLLR adaptation, ML-SAT with 4 regression classes yields worse results. The experimental results show that Eigen-SAT (34.4%,44.7%,#4) gives better performance than ML-SAT (34.8%,45.1%,#2). Furthermore Eigen-SAT uses only 25% of the parameters that ML-SAT is using(4*209 vs 2*1600 parameters).

	ML-SAT		Eigen-SAT K=209	
#TRANS	SWBD1	SWBD2	SWBD1	SWBD2
2	34.8(2*1600)	45.1	34.4(2*209)	45.0
4	35.2(4*1600)	46.3	34.4(4*209)	44.7

Table 2. Word Error Rate (%) of systems with ML-SAT, Eigen-SAT evaluated on Swbd1 and Swbd2 test sets.

Finally we carried out a series of experiments by using the first $K < N$ Eigen transforms that capture most of the variation in the data. Results are shown in Table 3. We see that we get similar performance with Eigen-SAT($K = 209$) from Table 2, by using even fewer parameters, e.g Eigen-SAT($K = 100$, #4). These results show that SVD has been able to find canonical transforms since only K(significantly smaller than N) bases are required in the proposed approach. We achieve even bigger dimensionality reduction(4*100 vs 2*1600 parameters) and make the whole procedure more robust and efficient.

	Eigen-SAT K=100		Eigen-SAT K=75	
#TRANS	SWBD1	SWBD2	SWBD1	SWBD2
2	34.7(2*100)	45.4	34.9(2*75)	45.8
4	34.4(4*100)	44.8	35.0(4*75)	45.7

Table 3. Word Error Rate(%) of systems with Eigen-SAT, evaluated on Swbd1 and Swbd2 test sets by varying K(most significant eigenvalues).

5. CONCLUSIONS & FUTURE WORK

The main concern of this work is rapid adaptation in LVCSR, where a limited amount of adaptation data is available. We proposed the integration of SAT and Eigen-space based adaptation which: i) reduces inter-speaker variability within the training set and ii) utilizes the training set transforms during testing via Eigen-space decomposition. Since the number of linear coefficients is much less than the number of parameters used in conventional MLLR, the approach iii) only requires a small amount of adaptation data for a robust estimation and iv) makes estimation of more than two regression classes feasible.

The experimental results confirm the effectiveness of Eigen-MLLR/Eigen-SAT in achieving word error rates superior to those obtained with other currently popular MLLR/ML-SAT adaptation techniques. The Eigen-matrices can effectively capture the inter-speaker variation and achieve better performance by using only 12% of the parameters (Table 3, $K = 100$, #4), that conventional ML-SAT is using.

Our goal is to incorporate Eigen-space transforms into Segmental Minimum Bayes Risk (SMBR) estimation [13]. We note that due to the great diversity of ASR errors in large vocabulary tasks, we expect the primary challenge to be robust estimation from sparse data. Eigen-space transforms described in this work are an ideal solution for tackling the sparsity problem in speaker independent systems because they incorporate prior knowledge from the training set.

6. REFERENCES

[1] P. C. Woodland, "Speaker adaptation: Techniques and challenges," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2000, pp. 85–90.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," in *Computer Speech and Language*, 1985, vol. 9, pp. 171–185.

[3] M. Gales, "Cluster adaptive training for speech recognition," in *Proc. ICSLP'98*, Sydney, Australia, 1998, pp. 1783–1786.

[4] R. Kuhn et. al, "Eigenvoices for speaker adaptation," in *Proc. ICSLP'98*, Sydney, Australia, 1998, pp. 1771–1774.

[5] K. Chen et al, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proc. ICSLP*, Beijing, Oct. 2000.

[6] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *International Conference on Spoken Language Processing*, 1996, pp. 1137–1140.

[7] M. Gales and P.C. Woodland, "Variance compensation within the mllr framework," in *Technical Report CUED/F-INFENT/TR242*. University of Cambridge Engineering Department, Cambridge, UK, February 1996.

[8] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book, Version 3.0*, July 2000.

[9] W. Byrne, "The JHU March 2001 Hub-5 Conversational Speech Transcription System," in *Proceedings of the NIST LVCSR Workshop*. NIST, 2001.

[10] M. Mohri and M. Riley, "Integrated context-dependent networks in very large vocabulary speech recognition," in *European Conference on Speech Communication and Technology*, 1999.

[11] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. S onmez, F. Weng, and J. Zheng, "The SRI march 200 Hub-5 conversational speech transcription system," in *Proceeding of the Speech Transcription Workshop*. NIST, 2000.

[12] P. C. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *Proceedings of the Tutorial and Research Workshop on Automatic Speech Recognition*. ISCA, 2000.

[13] V. Doumpiotis, S. Tsakalidis, and W. Byrne, "Discriminative training for segmental minimum Bayes-risk decoding," in *IEEE Conference on Acoustics, Speech and Signal Processing*. IEEE, 2003.