PHONETIC SUBSPACE ADAPTATION FOR AUTOMATIC SPEECH RECOGNITION

Sina Hamidi Ghalehjegh[*], *Richard C. Rose*

Electrical and Computer Engineering Department, McGill University, Montreal, Canada

ABSTRACT

An approach is proposed for adapting subspace projection vectors in the subspace Gaussian mixture model (SGMM) [1]. Subword models in the SGMM are composed of states, each of which are parametrized using a small number of subspace projection vectors. It is shown here that these projection vectors provide a compact and well-behaved characterization of phonetic information in speech. A regression based subspace vector adaptation approach is proposed for adapting these parameters. The performance of this approach is evaluated for unsupervised speaker adaptation on two large vocabulary speech corpora.

Index Terms— Speaker Adaptation, Phonetic Subspace

1. INTRODUCTION

This paper presents an approach for linear regression based adaptation of subspace projection vectors in the subspace Gaussian mixture model (SGMM) [1]. The SGMM is differentiated from the continuous density hidden Markov model (CDHMM) in that a large portion of the acoustic parameters are shared amongst all states of the model. The parametrization of the SGMM is summarized in Section 2. SGMM states are characterized by a small number of subspace projection vectors. The remaining shared model parameters are composed of linear subspace matrices and a shared pool of full covariance Gaussians.

Section 3 of the paper argues that these state projection vectors provide a compact and well-behaved characterization of phonetic information in the speech signal. As a result, one might expect that adapting parameters in this space might be efficient. That is, it may provide a good model of phonetic variability with a minimal number of adaptation utterances. To investigate this assertion, a linear regression based subspace vector adaptation (SVA) procedure is proposed for adapting the substate projection vectors. The adaptation procedure and the maximum likelihood based optimization algorithm for parameter estimation is presented in Section 3. An experimental study is performed to evaluate this adaptation procedure on the Spanish CallHome and Resource Management speech corpora in Section 4.

[*]This work was supported by the FQRNT and NSERC.

This work is related to previous work in linear regression based adaptation in both CDHMM and SGMM acoustic models. Applying a linear transformation either in model-space or feature-space has been shown to be a powerful tool for speaker adaptation in the CDHMM framework [2–5]. Some of the techniques include model-space maximum likelihood linear regression (MLLR) [2–4, 6] and feature-space constrained MLLR (CMLLR) [7, 8]. Prior adaptation research in the context of the SGMM was performed by Ghoshal et al. in [9]. That work involved a new estimation method for feature-space MLLR within the SGMM framework. This new estimation technique was shown to provide a relative 3.8% improvement in word error rate (WER) in the best case for the CallHome English corpus.

2. SUBSPACE GAUSSIAN MIXTURE MODEL

This section provides a brief description of the SGMM acoustic model [1]. In this new formalism, HMM states share common parameters. The means and mixture weights are controlled by a global mapping from a vector space, called "state projection vector," to the GMM parameters space and the covariance matrices are shared among all the states. An SGMM state can be represented by one or more state projection vectors. For an SGMM system configured with J states, each having M_j substates, the observation distribution for feature vector $\mathbf{x}_t$ in state j can be written as:

$$b_j(\mathbf{x}_t) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_{jmi}, \boldsymbol{\Sigma}_i) \qquad (1)$$

where m is the substate index. In (1) c_{jm} is the relative weight of substate m in state j. There are I full-covariance Gaussian densities shared between all the states. The mean vector, $\boldsymbol{\mu}_{jmi}$, for substate m in state j is a projection into the i^{th} subspace defined by a $S \times S$ linear subspace matrix $\mathbf{M}_i$,

$$\boldsymbol{\mu}_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}. \qquad (2)$$

The $S \times 1$ state projection vectors, $\mathbf{v}_{jm}$, in (2) for substate m in state j are the state specific parameters in the SGMM. The weights, w_{jmi}, in (1) are obtained from the state projection vector $\mathbf{v}_{jm}$ using a log-linear model:

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}} \qquad (3)$$

where $\mathbf{w}_i$ denotes the weight projection vector. The parameters of the SGMM model are estimated using the Expectation-Maximization (EM) algorithm as described in more detail in [1].

3. SUBSPACE VECTOR ADAPTATION (SVA)

3.1. Motivation

In the SGMM framework each state is associated with a vector-valued quantity that is called a "state projection vector". Figure 1(a) depicts a mapping of the state projection vectors, $\mathbf{v}_{jm}$, onto a two dimensional space[1]. The figure displays a scatter plot of these vectors for states associated with context dependent phoneme models. The centroids of state projection vectors, associated with context dependent models with a given center context phoneme, are displayed as text labels in the figure. There are two important characteristics of this plot. First, it is clear that the state projection vector for states corresponding to particular phonemes form compact clusters. Second, the clusters are naturally arranged in a space that is very similar to the articulatory based vowel triangle. As a result, performing acoustic adaptation by transforming the parameters in this space can be interpreted as adaptation in an articulatory-like space. In this work, the state projection vectors are adapted using an affine transformation of the form $\hat{\mathbf{v}}_{jm} = \mathbf{A}\mathbf{v}_{jm} + \mathbf{b}$. The ultimate goal is to find the transformation that maximizes the likelihood of the adaptation data given the adapted model.

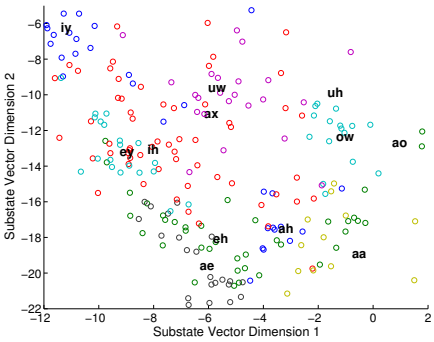
3.2. Defining the Auxiliary Function

Consider all substates have been partitioned into N clusters $\{c_1, \ldots, c_N\}$. To do the clustering, we use k-means algorithm with random initialization and the normalized cosine as a distance measure between state projection vectors: $(\mathbf{v}_i^T \mathbf{v}_j)/(\|\mathbf{v}_i\| \cdot \|\mathbf{v}_j\|)$. Then, all the substate vectors within the same cluster are transformed using a single affine transformation:

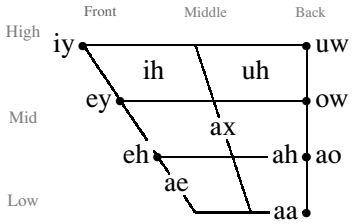
$$\hat{\mathbf{v}}_{jm} = \mathbf{A}^{(c_n)}\mathbf{v}_{jm} + \mathbf{b}^{(c_n)} = \left[\begin{array}{cc} \mathbf{A}^{(c_n)} & \mathbf{b}^{(c_n)} \end{array}\right] \left[\begin{array}{c} \mathbf{v}_{jm} \\ 1 \end{array}\right] = \mathbf{W}^{(c_n)}\mathbf{u}_{jm}$$

in which $\mathbf{A}^{(c_n)}$ is a $S \times S$ matrix and $\mathbf{b}^{(c_n)}$ is the bias vector for cluster c_n. For simplicity in our notations we will drop superscript (c_n), keeping in mind that we are doing adaptation for cluster c_n. The parameters of the affine transformation are found in a maximum likelihood (ML) fashion using EM approach [11]. Writing out the auxiliary function derived using

[1]We re-normalize the state projection vector as explained in Appendix K of [10] to concentrate the most important variation in lower dimensions



(a)



(b)

Fig. 1. (a) Scatter plot of the 1st and 2nd dimension of state projection vectors for RM (b) ARPAbet vowel triangle

Jensens inequality, we have:

$$\mathcal{Q}(\mathbf{W}) = K - \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j \in S^{(c_n)}}\sum_{m=1}^{M_j} \gamma_{jmi}(t) \times$$
$$\left[\frac{1}{2}(\mathbf{x}(t) - \hat{\boldsymbol{\mu}}_{jmi})^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}(t) - \hat{\boldsymbol{\mu}}_{jmi}) - \log \hat{\omega}_{jmi}\right]$$

where $\gamma_{jmi}(t)$ is the probability of being in state j, substate component m and Gaussian mixture component i at time t and the observation sequence $\mathbf{X} = \{\mathbf{x}(1), \ldots, \mathbf{x}(T)\}$ is the sequence of the adaptation data on which the transformation is to be trained. The optimum transformation matrix can be found by maximizing $\mathcal{Q}(\mathbf{W})$ w.r.t. $\mathbf{W}$. The auxiliary function consists of two parts. The mean-related part is very straight-forward to simplify. However, to simplify the weight-related part, we take an approach similar to the one used in [10]. We use the inequality $1 - (x/\overline{x}) \leq -\log(x/\overline{x})$ (which is an equality at $x = \overline{x}$) and also the quadratic approximation to the $\exp(x)$ around $x = x_0$, i.e. $\exp(x) \simeq \exp(x_0)(1 + (x - x_0) + 0.5(x - x_0)^2)$. The final auxiliary function will have the following form:

$$\mathcal{Q}(\mathbf{W}) = \sum_{j,m} \mathbf{f}_{jm}^T \mathbf{W}\mathbf{u}_{jm} - 0.5 \sum_{j,m} \mathbf{u}_{jm}^T \mathbf{W}^T \mathbf{C}_{jm} \mathbf{W}\mathbf{u}_{jm}, \quad (4)$$

where

$$\mathbf{f}_{jm} = \sum_{i'} (\gamma_{jmi'} - \gamma_{jm}\overline{\omega}_{jmi'} + \cdots$$
$$\gamma_{jm}\overline{\omega}_{jmi'}\mathbf{w}_{i'} \cdot \overline{\mathbf{W}}\mathbf{u}_{jm})\mathbf{w}_{i'} + \mathbf{y}_{jm} \quad (5)$$

and

$$\mathbf{C}_{jm} = \sum_{i'} \gamma_{jm} \overline{\omega}_{jmi'} \mathbf{w}_{i'} \mathbf{w}_{i'}^T + \sum_{i} \gamma_{jmi} \mathbf{H}_i \quad (6)$$

are the statistics that we need to accumulate in order to estimate the transformation matrix. The $\overline{\omega}_{jmi'}$ and $\overline{\mathbf{W}}$ correspond to their current values and other parameters in (5) and (6) are defined as follows:

$$\begin{aligned} \mathbf{H}_i &= \mathbf{M}_i^T \mathbf{\Sigma}_i^{-1} \mathbf{M}_i \\ \mathbf{y}_{jm} &= \textstyle\sum_{t,i} \gamma_{jmi}(t) \mathbf{M}_i^T \mathbf{\Sigma}_i^{-1} \mathbf{x}^T(t) \\ \gamma_{jmi} &= \textstyle\sum_t \gamma_{jmi}(t). \end{aligned}$$

Because of using a quadratic approximation while simplifying the weight part, there is no guarantee that increasing the simplified cost function will increase our auxiliary function. To tackle this problem, it is a "safer" option to take $\gamma_{jm}\overline{\omega}_{jmi'}$ in (5) and (6), which is the weighting term in the quadratic part, and replace it with $\max(\gamma_{jm}\overline{\omega}_{jmi'}, \gamma_{jmi'})$ [10].

3.3. Optimizing the Auxiliary Function

To find optimum transformation matrix and bias vector we maximize (4) w.r.t. $\mathbf{A}$ and $\mathbf{b}$. First, to compute $\mathbf{b}$ we take the derivative of (4) w.r.t. $\mathbf{b}$ and equate it to zero:

$$\mathbf{b} = \left(\sum_{j,m} \mathbf{C}_{jm} \right)^{-1} \left(\sum_{j,m} (\mathbf{f}_{jm} - \mathbf{C}_{jm} \mathbf{A} \mathbf{v}_{jm}) \right). \quad (7)$$

Then, to incorporate the new estimated value of $\mathbf{b}$, we recompute the statistics of (5) and (6). After that, we employ a gradient ascent approach to find the optimum transformation matrix. The reason is because finding the direct solution requires inverting a low-rank $S^2 \times S^2$ matrix which would be computationally expensive and cause numerical instabilities. In the iterative method, the transformation matrix in k^{th} iteration can be obtained as:

$$\mathbf{A}^{(k)} = \mathbf{A}^{(k-1)} + \mu^{(k-1)} \left. \frac{\partial \mathcal{Q}(\mathbf{W})}{\partial \mathbf{A}} \right|_{\mathbf{A}^{(k-1)}} \quad (8)$$

where

$$\frac{\partial \mathcal{Q}(\mathbf{W})}{\partial \mathbf{A}} = \sum_{j,m} \mathbf{f}_{jm} \mathbf{v}_{jm}^T - \sum_{j,m} \mathbf{C}_{jm} (\mathbf{A} \mathbf{v}_{jm} + \mathbf{b}) \mathbf{v}_{jm}^T. \quad (9)$$

The iteration terminates when the auxiliary function of $\mathcal{Q}(\mathbf{W})$ stops increasing. We also need to initialize $\mathbf{A}$. If a previous estimate of $\mathbf{A}$ exists (for example, if we are running multiple passes over the adaptation data), it is used as the initial estimate. Otherwise $\mathbf{A}^{(0)} = \mathbf{I}$ is a reasonable starting point. Generally 3-4 passes over the adaptation data will be sufficient to have a good estimate.

4. EXPERIMENTAL STUDY

This section presents an experimental study evaluating the performance of the SVA approach described in Section 3. Performance is reported as the WER obtained after unsupervised speaker adaptation is performed on the Resource Management (RM) and Spanish CallHome speech corpora. After introducing the task domain and describing how the baseline speaker-independent CDHMM and SGMM acoustic models are trained, we will present the speech recognition results using SVA technique. All the HMM training for CDHMM case were done using standard HTK toolkit [12]. For the SGMM, we use an implementation that is an extension to HTK with added libraries [13]. We extended the HTK toolkit to support SVA technique within SGMM framework.

4.1. Resource Management Read Speech Corpus

In the DARPA RM speech corpus, the degradation in ASR performance is mainly due to intrinsic sources of variability in speech. The environment and channel variability has relatively minor effect on the ASR performance. This is not the case for other corpora such as conversational telephone speech domain. As a result, one can attribute reductions in WER to the impact of adaptation techniques and the fact that how good they can model intrinsic sources of variability in the target speaker.

The RM corpus consists of 3990 utterances from 109 speakers taken from the standard RM SI-109 training set. The speech is parametrized using 12 MFCCs, normalized energy and the first and second differences of these parameters to give a 39 dimensional acoustic vector. The baseline system was based on three-state left-to-right HMM triphone models. Decision tree clustering was used to obtain a system with 1704 states, each having 6 mixtures of Gaussians. Also, the SGMM system was trained using the same training data set with $I = 256$ Gaussian mixtures shared between 1704 states with joint posterior initialization (JPI) [13]. No speaker adaptive training was used during training the baseline models. The ASR WER for CDHMM and SGMM baseline systems are 4.91% and 4.52% respectively. The ASR was evaluated using 1200 utterances from 12 speakers taken from the RM speaker dependent evaluation (SDE) set. Also a 991 word bi-gram language model was used.

Speaker adaptation is performed in an unsupervised mode with an average duration of 5.33 minutes of speech data per speaker. Figure 2 depicts the WER versus different number of clusters for SVA adaptation technique and standard CMLLR adaptation technique. The SGMM adaptation gives a relative 25% WER improvement with respect to SGMM baseline for the best case.

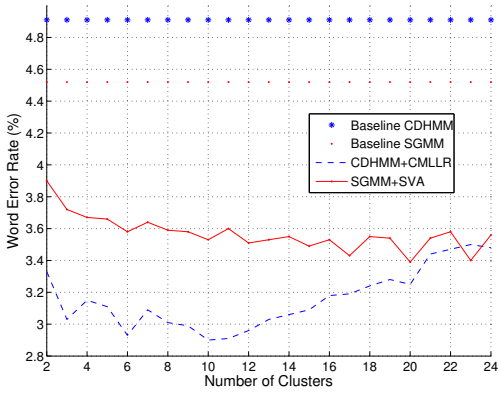


Fig. 2. The ASR word error rate for different number of clusters for RM task domain

4.2. Spanish CallHome Conversational Speech Corpus

The Spanish CallHome corpus is known to be a unique challenge for speech recognition [14]. Apart from the small size of the corpus, the speech data consists of inherent disfluencies. We used 16.5 hours of conversational speech data for training, and our test data consisted of 2.0 hours of conversational speech data collected from 46 speakers. The baseline system was based on three state left-to-right HMM triphone models, with a total of 1604 states. We used 16 Gaussian mixtures per state. A set of 13 PLP features along with their first and second differences were used as feature vectors. A trigram LM was used with a vocabulary of 45k words. The same 16.5 hours training set was used for training the SGMM system. This system has $I = 400$ shared full covariance Gaussians shared between 1604 states. The system was initialized with Gaussians obtained from a UBM obtained from speech-only segments from all speakers in the training corpus. Also, the flat start initialization approach was utilized [13]. No speaker adaptive training was used during training the baseline models. The baseline WERs for the CDHMM and SGMM systems are displayed in the first two rows of Table 1.

Speaker adaptation is performed in an unsupervised mode with an average duration of 2.62 minutes of speech data per speaker. We used only 2 clusters while doing the speaker adaptation experiment. The third and fourth rows of Table 1 display the WERs for the CMLLR adapted CDHMM system and the SVA adapted SGMM system, respectively.

4.3. Discussion

These two experiments show that the proposed unsupervised SGMM adaptation technique provides substantial improvement over an unadapted SGMM baseline system. It is clear that the relative performance improvement is less than that obtained by applying CMLLR to the CDHMM acoustic model. One possibility can be the numerical issues and the way we find the optimum solution. As discussed in earlier, due to ill-conditionality, we take the gradient ascent approach to optimize our auxiliary function. There are so many ways of optimizing a cost function in an iterative manner. One can use the simple gradient ascent algorithm and update the entire elements of the matrix simultaneously at each iteration (as discussed in Section 3.3). An alternative method can be a row-by-row approach, in which rather than the entire matrix, the rows are updated at each iteration using the same approach as in [15]. Among the methods we tried, the simple gradient ascent algorithm gives the best performance. Another possibility for higher WER of SGMM adaptation can be the choice of auxiliary function. In our proposed method we use the maximum likelihood approach to find the best solution. Therefore, there is no guarantee that the separability of the phonemic elements will be preserved. So one can use the interpolation of maximum likelihood criterion and discriminative objection function, as a final auxiliary function.

Table 1. WER for Spanish CallHome system. SAT indicates that the speaker adaptive training was used in training.

System	WER [%]
Baseline CDHMM	68.61
Baseline SGMM	67.29
CDHMM+SAT+CMLLR	65.55
SGMM+SVA	65.91

5. SUMMARY AND CONCLUSION

A new speaker adaptation technique was proposed. We presented SVA approach for adapting substate projection vectors in the SGMM framework. The SVA technique was motivated by the observation that the substate projection vectors are distributed in compact articulatory-like space. We then presented the experimental results. We obtained a 25% relative reduction in WER on the RM task domain and a 1.38% absolute reduction in WER on the Spanish CallHome task domain. These improvements are consistent with those obtained by Ghoshal et al in [9] using feature-space MLLR over the SGMM baseline system. The paper concluded with a discussion about the possibilities for having higher WER for SVA adaptation compared to CMLLR adaptation.

The future work will involve combining our SVA technique with the SGMM-based feature-space MLLR adaptation technique. We would like to investigate if we can use them as two complimentary techniques for speaker adaptation within the SGMM framework.

6. REFERENCES

[1] D. Povey et al., "The subspace gaussian mixture model–a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[2] A. Sankar and C. H. Lee, "Robust speech recognition based on stochastic matching," in *ICASSP*, 1995, pp. 121–124.

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

[4] 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. 171–186, 1995.

[5] L. R. Neumeyer, A Sankar, and V. V. Digalakis, "A comprehensive study of speaker adaptation techniques," in *Eurospeech*, 1995, pp. 1127–1130.

[6] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.

[7] C. J. Leggetter, *Improved Acoustic modelling for HMMs using linear transformations*, Ph.D. thesis, Cambridge University, 1995.

[8] J. Neto et al., "Unsupervised speaker-adaptation for hybrid HMM-MLP continuous speech recognition system," 1995, pp. 187–190, Eurospeech.

[9] A. Ghoshal et al., "A novel estimation of feature-space MLLR for full-covariance models," in *ICASSP*, 2010, pp. 4310–4313.

[10] D. Povey, "A tutorial-style introduction to subspace Gaussian mixture models for speech recognition," *Microsoft Research, Redmond, WA*, 2009.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[12] S. Young et al., "The HTK book (for HTK version 3.4)," 2006.

[13] R. Rose, S. C. Yin, and Y. Tang, "An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition," in *ICASSP*, 2011, pp. 4508–4511.

[14] G. Zavaliagkos, M. Siu, M. Colthurst, and J. Billa, "Using untranscribed training data to improve performance," in *ICSLP*, 1998.

[15] K. C. Sim and M. J. F. Gales, "Adaptation of precision matrix models on large vocabulary continuous speech recognition," in *Proc. ICASSP*, 2005, vol. 1, pp. 97–100.