DISCRIMINATIVE ACOUSTIC MODEL USING EIGENSPACE MAPPING FOR RAPID SPEAKER ADAPTATION

Bowen Zhou and *John H. L. Hansen*

Robust Speech Processing Group, Center for Spoken Language Research
University of Colorado at Boulder, U.S.A
{zhoub, jhlh}@cslr.colorado.edu   Web: http://cslr.colorado.edu

Abstract

It is widely believed that strong correlations exist across an utterance as a consequence of time-invariant characteristics of speaker and acoustic environments. It is verified in this paper that the first primary eigendirections of the utterance covariance matrix are speaker dependent. Based on this observation, a fast speaker adaptation algorithm entitled Eigenspace Mapping (EigMap) is proposed and described. EigMap rapidly adapts the speaker independent models by constructing discriminative acoustic models in the test speaker's eigenspace. Unsupervised adaptation experiments show that EigMap is effective in improving baseline models using very limited amounts of adaptation data with superior performance to conventional adaptation technique such as block diagonal MLLR. A relative improvement of 18.4% over baseline recognizer is achieved using EigMap with only about 4.5 seconds of adaptation data. It is also demonstrated that EigMap is additive to MLLR by encompassing the speaker dependent discrimination information. A significant relative improvement of 24.6% over baseline is observed by combining MLLR and EigMap techniques.

1. Introduction

With advances in applying speech technology to different tasks, many speech applications require rapid deployment of speech recognition with minimal resources. In such speech systems, it is desirable that the acoustic model can be dynamically improved based only on the baseline model and a very limited amount of adaptation data (i.e., no training data, no speaker dependent models or any speaker clustering information is demanded for adaptation, and the computational and storage overhead of the adaptation process should be inexpensive). In this paper, we describe a rapid speaker adaptation algorithm entitled EigMap to meet these requirements.

It is widely known that correlations exist across an utterance as a consequence of time-invariant characteristics of speaker and/or acoustic environments. Given a sequence of observation feature frames from an utterance, there are at least two types of correlation that exist over the observations: the temporal correlation between feature frames, and the correlation between feature components. However, state-of-the-art speech recognition technologies ignore such correlations. For example, it is usually assumed that observations are independent in both acoustic model training and decoding. The use of dynamic feature components partly captures some correlation between feature frames, but it is limited to neighboring frames. On the other hand, for many practical considerations, such as storage and computation, acoustic models typically assume diagonal covariance. This assumption ignores the correlations between feature components. It is expected that bringing these correlations into consideration should produce more accurate acoustic modeling. For example, linear discriminant analysis (LDA) [4, 7] and maximum likelihood linear transform (MLLT) [2] have been used to improve acoustic model training. The focus of this paper is to introduce a method that dynamically incorporates the correlation at the decoding phase for rapid model adaptation. It is noted that directly modeling of correlation is too expensive and not computationally practical. Alternatively, the proposed method constrains model parameters implicitly based on correlation.

The question of how to capture speaker information from limited amounts of adaptation data, and how to impose the speaker information appropriately into baseline acoustic models are the key problems investigated in this paper. The existence of strong correlation within an utterance has long been noted by researchers in the literature [1]. The motivation for using long distance correlation for rapid speaker adaptation is that the correlation should be speaker dependent. Intuitively, the manner by which speech frames affect each other is highly related to the vocal tract movement and speaking styles, which are largely dependent on the speaker [6]. In Sec. 2, a set of experiments are designed to verify this claim. As one might expect, it is observed from our experiments that the first primary eigendirections of the utterance covariance matrix encodes significant speaker information.

If every component Gaussian distribution in the acoustic model is viewed as a class, then a well-trained baseline model can be assumed to maintain a fair discrimination power between different Gaussians, in the sense of providing a reasonable between-class covariance B_x. B_x can be decomposed into the sum of variances along its different eigendirections. Among them, the variances that belong to the first primary eigendirection reflect the dominant power for discrimination. This paper proposes an algorithm to construct the discriminative acoustic models for the test speaker, by reserving dominant discriminating power from baseline model along the test speaker's first primary eigendirection of the specific speaker's between-class covariance B_y. Other constraints are also imposed on the adapted means to minimize the shift from baseline model due to insufficient observations of adaptation data in rapid adaptation. The adaptation process is performed through a linear transformation in the model space using a method entitled Eigenspace Mapping (EigMap). Based on the similar idea of EigMap, we have previously formulated an algorithm for rapid speaker adaptation

This work was supported by NSF under Cooperative Agreement No. IIS-9817485.

entitled Structural Maximum Likelihood Eigenspace Mapping (SMLEM) [8, 9].

Experimental results from this paper show that EigMap is effective in improving the baseline model using limited amounts of adaptation data with superior performance to MLLR. Moreover, EigMap is highly additive to MLLR by bringing additional discrimination information into the adapted acoustic model that maximizes the adaptation data likelihood.

The remainder of this paper is organized as follows: Sec. 2 investigates the speaker information in utterances, and shows that the first primary eigendirections of the observation covariance matrix encode significant speaker information; Sec. 3 describes the eigenspace mapping algorithm and points out the relationship between EigMap and LDA; Sec. 4 evaluates the EigMap algorithm using multiple experiments; Sec. 5 summarizes the paper contributions.

2. Speaker Information In Utterances

Previous work by other researchers have shown that the covariance of observation frames from a specific speaker encapsulates a range of speaker dependent information. Statistics based on the covariance matrix have been applied successfully in speaker identification and tracking [3].

Typically, dependence in feature observations exist between more than two feature components, and Principal Component Analysis (PCA) can help extract the most important axes of variations. In our study, we claim that the first primary eigendirections encode more significant speaker information than phonemic information. We design experiments to verify our claim. First, we select a set of speakers, $S = \{s_1, s_2, \ldots, s_L\}$, and randomly select an identical set of utterances $U = \{u_1, u_2, \ldots, u_L\}$ produced by each speaker in S. For each utterance u of each speaker, we estimate the covariance matrix B_u of the observation frames in the standard MFCC feature domain. The covariances are estimated independently for the static cepstrum (12 MFCC plus energy), delta, and double delta feature streams. Next, the first p eigendirections $[e_{u1}, e_{u2}, \ldots, e_{up}]$ of B_u are derived using PCA. A well-trained WSJ acoustic model Λ with 100K component Gaussians is used to represent the acoustic space. To measure the relative position of an eigenvector e_{uk} in this space, each Gaussian mean x_i in Λ was projected onto it to obtain an inner product $d_{iuk} = x_i \cdot e_{uk}$. Next, the v_{si}, variance of d_{iuk} across speaker set S, and the v_{ui}, variance across utterance set U, are estimated. The goal is to compare V_s and V_u, the averaged variances of v_{si} and v_{ui} across all component Gaussian projections, respectively. If the claim is correct, and the eigendirections of utterance covariance matrix are more likely to be speaker dependent, one might expect to observe that the former should be higher than the latter.

Part (a) and (b) in Fig. 1 compare the averaged projection variances onto the first and second eigendirections respectively. Clearly, the averaged variance across different speakers with the same utterance, V_s, is higher than the averaged variance across different utterances from the same speakers, V_u. This observation strongly supports the claim that the first primary eigendirections are more likely to be speaker dependent and are less affected by the phoneme contexts in utterances. It is interesting to note that the ratio V_s/V_u are in different ranges for each feature stream (i.e., the ratio is more than five for the static features, above two for the delta stream, while only slightly above one for double delta), as indicated in the lower part of Fig. 1. This again verifies the observation that the static feature stream carries the most significant speaker traits. The experimental results in Fig. 1 also suggest that the feature streams should be treated separately in such eigenspace processing, to assure that we are extracting appropriate speaker information from each stream.

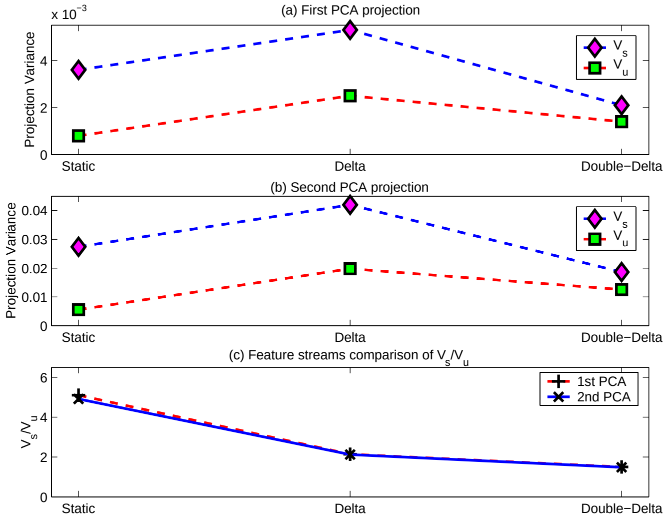

Figure 1: *Comparison of averaged Gaussian mean projection variances across speaker set S and across utterance set U: (a) Variances of projection onto 1st PCA of speaker's eigenspace, (b) Variances of projection onto 2nd PCA of speaker's eigenspace, and (c) comparison of V_s/V_u for different feature streams.*

3. Eigenspace Mapping (EigMap)

For the task of model adaptation, the improved model is achieved by adjusting the baseline model parameters based on adaptation data. From the previous section, it is assumed that the speaker dependent information can be learned from the first primary eigendirections. On the other hand, a well-trained baseline model Λ is assumed to maintain a fair model discrimination between Gaussian means $\{x_i | i = 1, 2, \ldots, N\}$, in the sense of providing a reasonable between-class covariance B_x:

$$B_x = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T - \bar{x}\bar{x}^T \qquad (1)$$

where every component Gaussian is treated as a single class, and $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$. B_x can be decomposed as the sum of variations along its eigendirections $\{e_{x1}, e_{x2}, \ldots, e_{xn}\}$:

$$\log(\det(B_x)) = \sum_{i=1}^{n} \log \lambda_i \simeq \sum_{i=1}^{p} \log \lambda_i \qquad (2)$$

where n is the Gaussian dimension, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ are the rank ordered eigenvalues of the symmetric semi-positive definite matrix B_x. The variance of the ith principal component is λ_i, and in a loose sense, this component "accounts for" a proportion $\lambda_i / \sum_{j=1}^{n} \lambda_j$ of the total variances. It is assumed that $p < n$ is the number of primary eigenvalues that contribute dominant variations, and hence the variations along the corresponding eigendirections $\{e_{x1}, e_{x2}, \ldots, e_{xp}\}$ provide the most significant discrimination power, among any p eigendirections, in the sense of maximizing the Fisher ratio: $F = \frac{\det(B)}{\det(W)}$, where W is the averaged within-class covariance matrix.

3.1. EigMap

The basic idea of EigMap is to maintain the between-class variances (i.e, the discrimination power) of the baseline Gaussian means unchanged along the first primary eigendirections in test speaker's eigenspace. Given the primary eigendirections $\{e_{y1}, e_{y2}, \ldots, e_{yp}\}$ of test speaker's observation covariance matrix B_y, the adapted Gaussian means $\{y_i | i = 1, 2, \ldots, N\}$ should satisfy:

$$\sum_{j=1}^{n} y_{ij} e_{ymj} = \sum_{j=1}^{n} x_{ij} e_{xmj}, m = 1, \ldots, p. \tag{3}$$

For every component Gaussian x_i in the model Λ, all possible adapted means y_i that satisfies Eq. (3) form a $(n-p)$-dimensional subplane Ω in the acoustic space that is given by:

$$\Omega = \{y_i | \sum_{j=1}^{n} y_{ij} e_{ymj} = \sum_{j=1}^{n} x_{ij} e_{xmj}, m = 1, \ldots, p\}. \tag{4}$$

In the task of rapid model adaptation where observation data is sparse, aggressive assumptions based on insufficient adaptation data often tend to be unreliable. Alternatively, a more conservative approach is to minimize the shift from the well-trained baseline model parameters, given the constraint of no loss of discrimination power along the first dominant eigendirections in the test speaker eigenspace:

$$y = \underset{y \in \Omega}{\text{argmin}} \; (x-y)^T (x-y). \tag{5}$$

By substituting Eq. (4) into Eq. (5) and minimizing the objective function using the Lagrange Multiplier method, the adapted mean y can be obtained from x using a linear transformation: $f : \Re^n \rightarrow \Re^n, y = f(x) = \theta x$, with θ an $n \times n$ nonsingular matrix given by:

$$\theta = I_n - \sum_{i=1}^{p} (-1)^{(i-1)} e_{yi}^T (e_{yi} - e_{xi}), \tag{6}$$

where I_n is $n \times n$ identity matrix. Considering the orthogonality between eigenvectors (i.e., $e_{yi} \cdot e_{yj} = 0, \forall i \neq j$), one can show that $\theta = {E_y}^{-1} E_m$, where,

$$E_y^{-1} = [e_{y1}^T, e_{y2}^T, \ldots, e_{yn}^T] \tag{7}$$

and,

$$E_m = [e_{x1}, \ldots, e_{xp}, e_{y(p+1)}, \ldots, e_{yn}]^T. \tag{8}$$

After transforming the baseline model mean x into y using Eq. (6), the discrimination information is assumed to be mostly encapsulated in the first q dimensions, where $p < q < n$, hence the last $n - q$ dimensions of y can be discarded. In the model space, this is equivalent to setting the last $n - q$ rows of θ to zeros:

$$\bar{\theta} = [\theta_{q \times n}, 0_{(n-q) \times n}]^T \tag{9}$$

It should be noted that following the observation from previous section, we treat each feature stream separately in estimating $\bar{\theta}$.

Our previously proposed algorithm, SMLEM, incorporates concepts found in EigMap. In SMLEM, all the component Gaussian means of the well-trained baseline model are clustered into a binary tree. A structural eigenspace mapping approach was employed to allow hierarchical mapping, and at the levels determined by the amount and distribution of adaptation data, the EigMap processing is applied to adapt the component Gaussians within this class. This is motivated by the fact that the discrimination in a smaller class of more similar Gaussian means should be more valuable in pattern recognition than the discrimination in a global space. An eigenspace bias b is also introduced in SMLEM, with the adapted mean y obtained as:

$$y = \theta x + {E_y}^{-1} b \tag{10}$$

where b is derived in a manner that maximizes the adaptation data likelihood $L(O|\Lambda)$. The accumulation equation for estimating the bias b based on the EM algorithm is given in [9].

3.2. Between-Class Variances Estimation

One of the key points in the EigMap scheme is how to estimate the between-class variances B_y for the test speaker, and accordingly, B_x for the baseline model given the adaptation data. One approach is based on Viterbi forced alignment. In this approach, the best state sequence of the adaptation data is found through Viterbi alignment:

$$Q(q_1, \ldots, q_t) = \underset{q_1, \ldots, q_t}{\text{argmax}} \; P(q_1, \ldots, q_t; o_1, \ldots, o_t | \Lambda). \tag{11}$$

B_y is directly computed from observed adaptation speech frames o_i:

$$B_y = B_o = \frac{1}{t} \sum_{i=1}^{t} o_i o_i^T - \bar{o}\bar{o}^T, \tag{12}$$

where t is the number of observed speech frames, and $\bar{o} = \frac{1}{t}\sum_{i=1}^{t} o_i$. At the baseline model side, a "simulated" observation from the perspective of baseline models is, given the best state q_i at each time i:

$$\hat{o_i} = \sum_{q_i : m} w_{q_i m} x_{q_i m}, i = 1, 2, \ldots, t, \tag{13}$$

where $w_{q_i m}$ is the mixture weight of state q_i with the constraint: $\sum_m w_{q_i m} = 1$. Next, B_x is estimated from these "simulated" observations, with $\bar{\hat{o}} = \frac{1}{t}\sum_{i=1}^{t} \hat{o_i}$:

$$B_x = \frac{1}{t} \sum_{i=1}^{t} \hat{o_i} \hat{o_i}^T - \bar{\hat{o}}\bar{\hat{o}}^T. \tag{14}$$

3.3. EigMap and LDA

LDA, or more recently, HDA [4], is used by many researchers to improve acoustic model discrimination before ML model training. The goal of LDA is to find the linear transformation θ in the *feature space* to maximize the following objective function:

$$J(\theta) = \frac{\text{det}\,(\theta B \theta^T)}{det(\theta W \theta^T)}. \tag{15}$$

However, EigMap seeks a linear transformation θ in the *model space* for rapid model adaptation, and therefore no training data is required. If the Gaussian variance is not adapted, then the within-class covariance W is unchanged after EigMap transformation. In this sense, the EigMap transformation θ can be viewed as a solution that maximizes the same objective function in Eq. (15) with the constraint that discrimination is obtained through the speaker's first primary eigendirections.

Table 1: *WER (%) of native speaker (WSJ Spoke4) testing with 4 seconds of adaptation data in average.*

Speaker	4o6	4o7	4o8	4o9	Avg.
Baseline	4.4	3.8	8.0	6.2	5.6
BD-MLLR	5.1	3.7	8.3	5.3	5.6
EigMap	4.0	3.3	7.6	5.9	5.2

4. Experiments and Results

4.1. Experimental Setup

The adaptation experiments reported here are all conducted in an *unsupervised* manner, on the DARPA WSJ Spoke3 and Spoke4 corpus. The baseline system has 6275 context-dependent tied states and 100K diagonal mixture component Gaussians. The baseline system uses a feature of 39 dimensions with 13 static cepstral coefficients plus delta and double-delta. The Spoke4 corpus is collected from 4 native speakers of American English with balanced gender. The Spoke3 data consists of non-native speakers. Each speaker of Spoke3 provides a set of adaptation utterances, and another set of 40 utterances for testing. We select the last 6 speakers from Spoke3 for our experiments [1] (approximately 3900 words in the test set). For Spoke4, all speakers and all the 50 test utterances from group G of each speaker are used in the evaluation (approximately 3300 words in the test set). Since we are primarily interested in rapid adaptation, only a single adaptation utterance is used to improve the baseline model. To account for variability in the small amount of data, and to obtain statistically representable results, 3 randomly selected adaptation utterances are identically used for each test speaker in adaptation. The adaptation data ranges from 3.7 to 5 seconds of speech for different utterances and speakers. All experimental results presented are obtained by averaging all open experiments.

The EigMap algorithm was compared with the block diagonal MLLR (BD-MLLR) scheme, since the amount of adaptation data is very limited, and it is shown from experiments that BD-MLLR achieves better performance than conventional MLLR [9] due to the reduced parameters to be estimated. For the same reason, one global regression class is used for BD-MLLR adaptation. For a fair comparison, EigMap also constructs *only a global eigenspace* for both test speaker and baseline model for the mapping. In our experiments, $n = 13$ for static, delta and double-delta streams, and the p, q are determined automatically for each stream based on the adaptation data distribution.

4.2. Experimental Results

Table 1 shows the performance comparison using Spoke4 corpus with about 4 seconds of adaptation data. In average, BD-MLLR achieves no improvement over baseline, while EigMap obtains consistent improvements for all speakers, with an average of 7% relative improvement from baseline. This observation suggests that even the test data matches the acoustic model well, the discrimination introduced by EigMap is still able to improve the acoustic model for more accurate classification. The experimental results using Spoke3 corpus are summarized in Table 2. It clearly shows that EigMap effectively enhances the baseline by a relative improvement of 18.4% with only about 4.5 seconds of adaptation data, when BD-MLLR achieves a 15.9% relative improvement. Moreover, EigMap is highly additive to MLLR by bringing additional discrimination information into the adapted acoustic model that maximizes the adaptation data likelihood. By applying EigMap to the MLLR adapted model, a significant relative improvement of 24.6% is observed in the experiments.

Table 2: *WER (%) of non-native speaker (WSJ Spoke3) testing with 4.5 seconds of adaptation data in average, where BDM stands for the BD-MLLR.*

Spkr	Baseline	BDM	EigMap	BDM+EigMap
4n5	23.5	20.2	21.4	20.2
4n8	16.4	13.0	13.6	13.3
4n9	21.6	18.9	16.7	15.0
4na	11.9	10.3	8.0	7.5
4nb	32.0	28.3	25.8	25.8
4nc	18.7	13.6	15.9	11.6
Avg	20.7	17.4	16.9	15.6
Rel. Imp	–	15.9%	18.4%	24.6%

5. Conclusions

This paper has introduced the Eigenspace Mapping (EigMap) algorithm for constructing a discriminative acoustic model for rapid speaker adaptation. EigMap preserves the discrimination power of the baseline model in the test speaker's eigenspace with constraints. Unsupervised adaptation experiments show that EigMap can effectively improve the baseline model with very limited amounts of adaptation data. Moreover, EigMap is able to provide additional performance gain to MLLR. Combining MLLR and EigMap, a significant 24.6% relative improvement is achieved with only about 4.5 seconds of adaptation data.

6. References

[1] M. Blomberg, "Speech Recognition Using Long Distance Relation In An Utterance", *From the web*.

[2] R. Gopinath, "Maxmimum Likelihood Modeling with Gaussian Distributions for Classification", *ICASSP '1998*, Seattle, WA., USA

[3] S. Johnson, "Speaker Tracking", *Master Thesis*, Univ. of Cambridge, 1997.

[4] N. Kumar, "Investigation of Silicon-Audiroty Models and Generaliztion of LDA for Improved Speech Recognition", *PhD Thesis*, Johns Hopkisn Univ., 1997.

[5] L. R. Neumeyer et. al., "A Comparative Study of Speaker Adaptation Techniques", *Eurospeech'1995*, 1995.

[6] Edited by J. Perkell and D. Klatt, "Invariance And Variability In Speech Processes", *Law. Erl. Assoc.*, 1986.

[7] G. Saon, et. al., "Maximum Likelihood Discriminant Feature Spaces", *ICASSP'2000*, Istanbul, Turkey, June 2000.

[8] B. Zhou and J. Hansen, "A Novel Algorithm for Rapid Speaker Adaptation Based on Structural Maximum Likelihood Eigenspace Mapping", *Eurospeech'2001*, Aalborg, Denmark, 2001.

[9] B. Zhou and J. Hansen, "Improved Structural Maximum Likelihood Eigenspace Mapping For Speaker Adaptation", *ICSLP'2002*, Denver, CO., 2002.

[1] The first four speakers demonstrate a relatively high Word Error Rate (WER) that is above 65% for the baseline system. We believe this may be in conflict with our assumption for the EigMap algorithm that the SI models are reasonably well-trained for the test speakers. Therefore, we exclude the first 4 speakers.