Probabilistic Linear Discriminant Analysis for Acoustic Modelling

Liang Lu, *Member, IEEE* and Steve Renals, *Fellow, IEEE*

***Abstract*—In this letter, we propose a new acoustic modelling approach for automatic speech recognition based on probabilistic linear discriminant analysis (PLDA), which is used to model the state density function for the standard hidden Markov models (HMMs). Unlike the conventional Gaussian mixture models (GMMs) where the correlations are weakly modelled by using the diagonal covariance matrices, PLDA captures the correlations of feature vector in subspaces without vastly expanding the model. It also allows the usage of high dimensional feature input, and therefore is more flexible to make use of different type of acoustic features. We performed the preliminary experiments on the Switchboard corpus, and demonstrated the feasibility of this acoustic model.**

***Index Terms*—probabilistic linear discriminant analysis, acoustic modelling, automatic speech recognition**

I. Introduction

Speech recognition systems based on hidden Markov models (HMMs) with Gaussian mixture model (GMM) output distributions defined the state-of-the-art in acoustic modelling for about 20 years [1]–[3]. GMM-based systems have straightforward, easily parallelizable algorithms for training and adaptation, but have a limited ability to capture the correlations in the acoustic space since diagonal covariance matrices are normally used for computational reasons. Although a variety of acoustic feature representations have been investigated [4], [5] high-dimensional features lead to a significant expansion in model size. One way to address these limitations is the use of neural networks (NNs) which can learn the correlations within feature vectors [6]. NNs typically use high-dimensional features covering several frames of acoustic context [7]–[9], and deep neural networks (DNNs) have achieved significant reductions in word error rate (WER) across many datasets [10].

In this letter we propose a new acoustic model, based on the GMM, which is able to capture acoustic correlations and to use high-dimensional features. The approach is based on probabilistic linear discriminant analysis (PLDA), originally proposed for face recognition [11], and now heavily employed for speaker recognition based on i-vectors [12]–[14]. PLDA – which is closely related to joint factor analysis (JFA) [15] used for speaker recognition – is a probabilistic extension of linear discriminant analysis (LDA). In speaker or face recognition, PLDA factorizes the variability of the observations for a specific class (e.g. one speaker) using an *identity variable* – which is shared by all the observations of this class, and a *channel variable* that depends on each observation – which is used to explain the variability caused by the channel noise to each observation. However, the main difference is that JFA operates in the GMM mean supervector domain while the PLDA used in this work is operated directly in the the acoustic feature domain.

We extended PLDA to estimate the HMM state density functions: the PLDA identity variables depend only on the HMM states, while the acoustic observations depend on both the identity variables and channel variables. Since both types of latent variable can be estimated in a low-dimensional subspace, it is feasible for PLDA to be deployed in a high-dimensional feature space. In this letter we present the PLDA-HMM model and a training algorithm, together with experiments on Switchboard [16].

II. Probabilistic Linear Discriminant Analysis

PLDA is formulated by a generative model, where an acoustic frame vector $\mathbf{y}_t$ from the j-th HMM state at time index t can be expressed as

$$\mathbf{y}_t|j = \mathbf{U}\mathbf{x}_{jt} + \mathbf{G}\mathbf{z}_j + \mathbf{b} + \epsilon_{jt}, \quad \epsilon_{jt} \sim \mathcal{N}(\mathbf{0}, \Lambda), \tag{1}$$

where $\mathbf{z}_j$ is the state-dependent identity variable (referred as *state variable* for brevity) shared by the whole set of acoustic frames generated by the j-th state. $\mathbf{x}_{jt}$ is the channel variable which explains the per-frame variance. In this work, we do not consider the correlations between the latent variables to simplify model training. We assume that their prior distributions are both $\mathcal{N}(\mathbf{0}, \mathbf{I})$ for ease of Bayesian inference [11], while more general forms of the PLDA prior distribution have been investigated for speaker recognition [12]. $\mathbf{U}$ and $\mathbf{G}$ are two low rank matrices which span the subspaces to capture the major variations for $\mathbf{x}_{jt}$ and $\mathbf{z}_j$ respectively. They are analogous to the within-class and between-class subspaces in the standard LDA formulation, but are estimated probabilistically. $\mathbf{b}$ denotes the bias and ϵ_{jt} is the residual noise which is Gaussian with a zero mean and diagonal covariance.

A. Mixture of PLDAs

As a single PLDA model can only approximate one Gaussian distribution, we can use a mixture of PLDAs which can be written:

$$\mathbf{y}_t|j, m = \mathbf{U}_m\mathbf{x}_{jmt} + \mathbf{G}_m\mathbf{z}_{jm} + \mathbf{b}_m + \epsilon_{jmt}, \tag{2}$$

where $1 \leq m \leq M$ is the component index. Let c denote the component indicator variable, with prior distribution (weight)

Liang Lu and Steve Renals are with The University of Edinburgh, UK; email: {liang.lu,s.renals}@ed.ac.uk

This work was funded by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

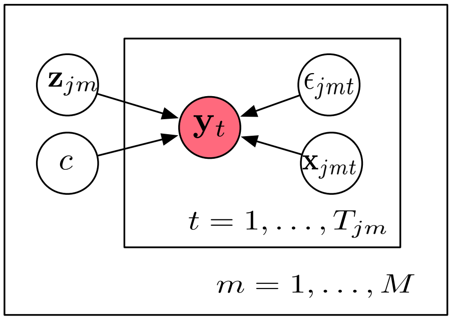

Fig. 1. Graphical model of mixture of PLDAs for the j-th HMM state. T_{jm} is the number of frames generated from the state j and component m.

$P(c = m) = \pi_m$. To avoid clutter we write $P(c = m|\mathbf{y}_t)$ as $P(m|\mathbf{y}_t)$. This model, which is shown in Figure 1, is related to the mixture of factor analysis model used for speaker and speech recognition [17], [18].

B. Covariance modelling

The proposed PLDA model can learn feature correlations, as can be seen from the marginal prior distribution of $\mathbf{y}_t$ given state j and component m under the independence assumption between $\mathbf{x}_{jmt}$ and $\mathbf{z}_{jm}$

$$\begin{aligned} P(\mathbf{y}_t|j,m) &= \int P(\mathbf{y}_t|\mathbf{x}_{jmt}, \mathbf{z}_{jm}, j, m) \\ &\quad \times P(\mathbf{x}_{jmt})P(\mathbf{z}_{jm})d\mathbf{x}_{jmt}d\mathbf{z}_{jm} \\ &= \mathcal{N}\left(\mathbf{y}_t; \mathbf{b}_m, \mathbf{U}_m\mathbf{U}_m^T + \mathbf{G}_m\mathbf{G}_m^T + \Lambda_m\right) \end{aligned} \tag{3}$$

Using low-rank matrices for $\mathbf{U}_m$ and $\mathbf{G}_m$ allows this model to be used in a high-dimensional acoustic feature space, in contrast to other full covariance modelling approaches such as semi-tied covariance matrices [19], diagonal priors [20] or sparsity constrained [21]. Other approaches which use a low-rank matrix approximation include EMLLT [22] and SPAM [23]. The PLDA model is also closely related to the subspace GMM (SGMM) [24], and the factor analysed HMM (FAHMM) [25]. In fact, if we do not consider the softmax function for the GMM weights, the SGMM acoustic model may be represented as

$$\mathbf{y}_t|j,m = \mathbf{G}_m\mathbf{z}_j + \epsilon_{jmt}, \quad \epsilon_{jmt} \sim \mathcal{N}(\mathbf{0}, \tilde{\Lambda}_m) \tag{4}$$

where the state-depended variables $\mathbf{z}_j$ are tied across the Gaussian components, although they can be optionally mixed up to improve model capacity. SGMMs directly use globally shared full covariance matrices without introducing another set of projection matrices to capture the frame level correlations. However, the SGMM is computationally expensive when in the case of high-dimensional features. Similar to SGMMs, the PLDA-based model can also share the state-independent parameters across domains, such as for cross-lingual speech recognition [26].

The FAHMM may be represented as

$$\mathbf{y}_t|j,m = \mathbf{C}_j\mathbf{x}_{jnt} + \epsilon_{jmt}, \quad \epsilon_{jmt} \sim \mathcal{N}(\boldsymbol{\mu}_{jm}, \tilde{\Lambda}_{jm}) \tag{5}$$

$$\mathbf{x}_{jnt} \sim \mathcal{N}(\boldsymbol{\mu}_{jn}, \boldsymbol{\Sigma}_{jn}) \tag{6}$$

where $\mathbf{C}_j$ is analogous to $\mathbf{U}_m$, but is tied to each HMM state rather than being globally shared. ϵ_{jmt} and $\mathbf{x}_{jnt}$ are modelled by two separate GMMs for each HMM state, which makes the inference problem for FAHMM more complex compared to a PLDA model.

III. Model Training

Since the state-dependent and state-independent parameters of a PLDA-based model are correlated, there is no closed form solution to update them in a joint fashion. However, an expectation-maximization (EM) algorithm can be employed.

A. Likelihoods

In order to accumulate the statistics to update the model parameters, we first need to compute the likelihoods of the model given the acoustic data, and the posterior distributions of the latent variables $\mathbf{z}_{jm}$ and $\mathbf{x}_{jmt}$ given the current model estimate. Depending on whether the latent variables are integrated out or not, the likelihood can be estimated as follows.

1) Point estimate: This approach refers to using the maximum a posteriori (MAP) estimate of the latent variables $\mathbf{x}_{jmt}$ and $\mathbf{z}_{jm}$ to compute the likelihood function

$$\begin{aligned} &p(\mathbf{y}_t|\bar{\mathbf{x}}_{jmt}, \bar{\mathbf{z}}_{jm}, j, m) \\ &\quad = \mathcal{N}(\mathbf{y}_t; \mathbf{U}_m\bar{\mathbf{x}}_{jmt} + \mathbf{G}_m\bar{\mathbf{z}}_{jm} + \mathbf{b}_m, \Lambda_m) \end{aligned} \tag{7}$$

where $\bar{\mathbf{x}}_{jmt}$ and $\bar{\mathbf{z}}_{jm}$ denote the means of the posterior distributions of $\mathbf{x}_{jmt}$ and $\mathbf{z}_{jm}$ respectively.

2) Uncertainty estimate: This approach refers to marginalising out the channel variable $\mathbf{x}_{jmt}$ using its prior distribution, in order to compensate for the uncertainties in the estimation of $\mathbf{x}_{jmt}$, resulting in the following likelihood function:

$$\begin{aligned} &p(\mathbf{y}_t|\bar{\mathbf{z}}_{jm}, j, m) \\ &\quad = \int p(\mathbf{y}_t|\mathbf{x}_{jmt}, \bar{\mathbf{z}}_{jm}, j, m)P(\mathbf{x}_{jmt})d\mathbf{x}_{jmt} \qquad (8) \\ &\quad = \mathcal{N}\left(\mathbf{y}_t; \mathbf{G}_m\bar{\mathbf{z}}_{jm} + \mathbf{b}_m, \mathbf{U}_m\mathbf{U}_m^T + \Lambda_m\right). \qquad (9) \end{aligned}$$

This method is similar to the channel integration evaluation method used for JFA based speaker recognition [27], [28]. Note that the likelihood can be efficiently computed without inverting matrices $\mathbf{U}_m\mathbf{U}_m^T + \Lambda_m$ [27], which makes it feasible when $\mathbf{y}_t$ is high dimensional. It is also possible to marginalise out the state variable $\mathbf{z}_{jm}$ alone or jointly with $\mathbf{x}_{jmt}$ similar as the methods used in [28].

B. Posteriors

Given the likelihoods, we then compute the posterior distributions of the latent variables $\mathbf{x}_{jmt}$ and $\mathbf{z}_{jm}$, using conjugate priors. The posterior distribution of $\mathbf{x}_{jmt}$ is given by

$$\begin{aligned} &P(\mathbf{x}_{jmt}|\mathbf{y}_t, \bar{\mathbf{z}}_{jm}, j, m) \\ &\quad = \frac{p(\mathbf{y}_t|\mathbf{x}_{jmt}, \bar{\mathbf{z}}_{jm}, j, m)P(\mathbf{x}_{jmt})}{\int p(\mathbf{y}_t|\mathbf{x}_{jmt}, \bar{\mathbf{z}}_{jm}, j, m)P(\mathbf{x}_{jmt})d\mathbf{x}_{jmt}}. \end{aligned} \tag{10}$$

With some algebraic rearrangement, we can obtain

$$P(\mathbf{x}_{jmt}|\mathbf{y}_t, \bar{\mathbf{z}}_{jm}, j, m) = \mathcal{N}(\mathbf{x}_{jmt}; \mathbf{V}_m^{-1}\mathbf{w}_{jmt}, \mathbf{V}_m^{-1}) \tag{11}$$

$$\mathbf{V}_m = \mathbf{I} + \mathbf{U}_m^T\Lambda_m^{-1}\mathbf{U}_m \tag{12}$$

$$\mathbf{w}_{jmt} = \mathbf{U}_m^T\Lambda_m^{-1}(\mathbf{y}_t - \mathbf{G}_m\bar{\mathbf{z}}_{jm} - \mathbf{b}_m) \tag{13}$$

Similarly, we can derive the posterior distribution of $\mathbf{z}_{jm}$ given all the observations $\mathbf{y}_t$ and the latent variables $\mathbf{x}_{jmt}$ that $\mathbf{z}_{jm}$ depends on (cf. Figure 1):

$$\begin{aligned} &P(\mathbf{z}_{jm}|\mathbf{y}_t, \bar{\mathbf{x}}_{jmt}, j, m, t = 1, \ldots, T_{jm}) \\ &\quad = \mathcal{N}(\mathbf{z}_{jm}; \mathbf{F}_{jm}^{-1}\mathbf{d}_{jm}, \mathbf{F}_{jm}^{-1}) \end{aligned} \tag{14}$$

$$\mathbf{F}_{jm} = \mathbf{I} + \sum_t \gamma_{jmt} \mathbf{G}_m^T \Lambda_m^{-1} \mathbf{G}_m \tag{15}$$

$$\mathbf{d}_{jm} = \mathbf{G}_m^T \Lambda_m^{-1} \sum_t \gamma_{jmt} (\mathbf{y}_t - \mathbf{U}_m \bar{\mathbf{x}}_{jmt} - \mathbf{b}_m). \tag{16}$$

Where γ_{jmt} denotes the per-frame posterior probability, given as (using the uncertainty likelihood estimation (8))

$$\gamma_{jmt} = P(j, m|\mathbf{y}_t) = P(j|\mathbf{y}_t) \frac{\pi_m p(\mathbf{y}_t|\bar{\mathbf{z}}_{jm}, j, m)}{\sum_m \pi_m p(\mathbf{y}_t|\bar{\mathbf{z}}_{jm}, j, m)} \tag{17}$$

where $P(j|\mathbf{y}_t)$ is the HMM state posterior which can be obtained using the forward-backward algorithm[1].

C. Model Update

We may use the EM algorithm to update the model parameters of a PLDA-based acoustic model. For instance, the auxiliary function to update $\mathbf{U}_m$ is

$$\begin{aligned} \mathcal{Q}(\mathbf{U}_m) &= \sum_{jt} \int P(j, m|\mathbf{y}_t) P(\mathbf{x}_{jmt}|\mathbf{y}_t, \bar{\mathbf{z}}_{jm}, j, m) \\ &\qquad \times \log p(\mathbf{y}_t|\mathbf{x}_{jmt}, \bar{\mathbf{z}}_{jm}, j, m) d\mathbf{x}_t \\ &= \sum_{jt} \gamma_{jmt} \mathbb{E}\Bigg[-\frac{1}{2}\mathbf{x}_{jmt}^T \mathbf{U}_m^T \Lambda_m^{-1} \mathbf{U}_m \mathbf{x}_{jmt} \\ &\qquad + \mathbf{x}_{jmt}^T \mathbf{U}_m^T \Lambda_m^{-1} (\mathbf{y}_t - \mathbf{G}_m \bar{\mathbf{z}}_{jm} - \mathbf{b}_m) \Bigg] + k \\ &= \sum_{jt} \gamma_{jmt} \mathrm{Tr} \Bigg(\Lambda_m^{-1} \Big(-\frac{1}{2} \mathbf{U}_m \mathbb{E}[\mathbf{x}_{jmt}\mathbf{x}_{jmt}^T] \mathbf{U}_m^T \\ &\qquad + (\mathbf{y}_t - \mathbf{G}_m \bar{\mathbf{z}}_{jm} - \mathbf{b}_m) \mathbb{E}^T[\mathbf{x}_{jmt}] \mathbf{U}_m^T \Big) \Bigg) + k \end{aligned}$$

where k is a constant value that is independent of $\mathbf{U}_m$, γ_{jmt} denotes the component posterior probability $P(j, m|\mathbf{y}_t)$, and $\mathbb{E}[\cdot]$ is the expectation operation over the posterior distribution of $\mathbf{x}_{jmt}$. By setting $\partial \mathcal{Q}(\mathbf{U}_m)/\partial \mathbf{U}_m = 0$ we obtain

$$\begin{aligned} \mathbf{U}_m = &\left(\sum_{jt} \gamma_{jmt} (\mathbf{y}_t - \mathbf{G}_m \bar{\mathbf{z}}_{jm} - \mathbf{b}_m) \mathbb{E}^T[\mathbf{x}_{jmt}] \right) \\ &\times \left(\sum_{jt} \gamma_{jmt} \mathbb{E}\left[\mathbf{x}_{jmt}\mathbf{x}_{jmt}^T\right] \right)^{-1} \end{aligned} \tag{18}$$

The updates for $\{\mathbf{G}_m, \mathbf{b}_m, \Lambda_m\}$ can be derived similarly.

TABLE I
A TRAINING RECIPE FOR A PLDA-HMM ACOUSTIC MODEL

1. Train a diagonal GMM and initialize $\mathbf{G}_m, \mathbf{U}_m, \mathbf{b}_m$ and Λ_m. Set $\mathbf{z}_{jm}$ and $\mathbf{x}_{jmt}$ to be Gaussian as $\mathcal{N}(\mathbf{0}, \mathbf{I})$.
2. Update $\mathbf{U}_m, \mathbf{b}_m$ and Λ_m using the model as equation (19) for 4 $\sim$ 6 iterations.
3. Select the subset of components for each frame with the highest likelihood using the model from step 2.
4. Update the posterior distribution of $\mathbf{z}_{jm}$ and $\mathbf{x}_{jmt}$ given the current estimate of $\mathbf{G}_m, \mathbf{U}_m, \mathbf{b}_m, \Lambda_m$.
5. Accumulate the statistics to update of $\mathbf{G}_m, \mathbf{U}_m, \mathbf{b}_m, \Lambda_m$.
6. Optionally re-align the training data using the current model.
7. Go to step 4 until convergence.

D. Training Recipe

To obtain appropriate model resolution in the feature space it is necessary to use a relatively large number of components in the PLDA-based model, e.g. $m = 400$ in (2). However, this can results in data sparsity when estimating the state-dependent model parameters $\mathbf{z}_{jm}$. We may tie $\mathbf{z}_{jm}$ across the components, as in the SGMM (4); however, in this work we simply set the distribution of $\mathbf{z}_{jm}$ to be its prior if there is not enough training data. This approach is preferred in order to make it easier to scale the model to a very large training dataset and to avoid sub-state splitting as used in SGMMs [24].

To reduce the computational cost, we do not accumulate statistics over all the PLDA components, but only over those components (typically 10–20) which have higher posterior probabilities for each acoustic frame [24]. This is a reasonable approximation since most of the components have very small posterior probabilities. We selected the subset of components for each frame according to its likelihood to a mixture of factor analysers [29] based global background model

$$\mathbf{y}_t|m = \mathbf{U}_m \mathbf{x}_{mt} + \mathbf{b}_m + \epsilon_{mt} \tag{19}$$

This is analogues to the universal background model used in SGMMs for Gaussian selection, but is trained in the low dimensional features space rather than using full covariance.

As stated before, the state and channel variables are intitialized as Gaussian distributions $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The remaining PLDA parameters may be initialized randomly, however, this can limit reproducibility. In this work, we used a diagonal covariance GMM which has the same number of components as the PLDA to initialize the parameters $\mathbf{G}_m, \mathbf{U}_m, \mathbf{b}_m$, and Λ_m. More precisely, $\mathbf{b}_m$ and Λ_m are initialized by the corresponding mean and covariance of the GMM. We then take the singular value decomposition of all the GMM means and use the principal eigenvectors to initialize the matrices $\mathbf{U}_m$ and $\mathbf{G}_m$. We have observed this to be a good starting point to train the model in our experiments. The full training recipe is summarised in Table I.

IV. EXPERIMENTS

We performed experiments on the Switchboard corpus [16], where the training set contains about 300 hours of conversational telephone speech. The Hub-5 Eval 2000 data [30] is used as the test set, which containss the Switchboard (SWB)

[1] In this work we used Viterbi training, so the value of $P(j|\mathbf{y}_t)$ is binary.

TABLE II
WERs (%) using 33 hours Switchboard training data and maximum-likelihood training criterion.

System	Feature	#Input frames	Dim	Likelihood	CHM	SWB	Avg
GMM-HMM	MFCC_0+Δ+$\Delta\Delta$	1	39	-	54.0	36.6	45.4
GMM-HMM	MFCC_0+LDA_STC	5	40	-	52.4	34.4	43.7
GMM-HMM	MFCC_0+LDA_STC	7	40	-	50.6	33.5	42.2
GMM-HMM	MFCC_0+LDA_STC	9	40	-	50.7	33.3	42.1
GMM-HMM	MFCC_0+LDA_STC	11	40	-	50.9	34.1	42.4
PLDA-HMM	MFCC_0	5	65	Point	63.0	43.7	53.4
PLDA-HMM	MFCC_0	7	91	Point	61.1	43.6	52.4
PLDA-HMM	MFCC_0	9	117	Point	62.3	43.6	53.0
PLDA-HMM	MFCC_0	5	65	Uncertainty	51.4	33.1	42.3
PLDA-HMM	MFCC_0	7	91	Uncertainty	49.5	32.4	41.1
PLDA-HMM	MFCC_0	9	117	Uncertainty	49.3	31.5	40.6
PLDA-HMM	MFCC_0	11	143	Uncertainty	49.7	33.2	41.6
PLDA-HMM	MFCC_0+Δ+$\Delta\Delta$	3	117	Uncertainty	49.9	32.4	41.3
PLDA-HMM	MFCC_0+Δ+$\Delta\Delta$	5	195	Uncertainty	52.2	34.0	43.1
SGMM-HMM	MFCC_0+Δ+$\Delta\Delta$	1	39	-	48.5	31.4	40.1

TABLE III
WERs (%) using 300 hours of training data.

System	CHM	SWB	Avg
GMM+MFCC_0+LDA_STC	42.6	25.6	34.2
PLDA+MFCC_0	41.4	25.2	33.5
SGMM+MFCC_0+Δ+$\Delta\Delta$	39.8	24.4	32.3

and Callhome (CHM) evaluation subsets. We implemented the PLDA-based acoustic model within the Kaldi speech recognition toolkit [31]. We used the pronunciation lexicon that was supplied by the Mississippi State transcriptions [32] which has more than 30,000 words, and a trigram language model was used for decoding.

The GMM and SGMM baseline systems used 39-dimensional mel frequency cepstral coefficients (MFCCs) with first and second derivatives (MFCC_0_Δ_$\Delta\Delta$). To take advantage of longer context information, for the GMM systems we have also performed experiments of using spliced MFCC_0 of different context window size, followed by a global LDA transformation to reduce the feature dimensionality to be 40, and a global semi-tied covariance (STC) matrix transform [19] to de-correlate the features. The PLDA systems directly used the concatenated MFCCs with various size of context window, without de-correlation and dimensionality reduction. Although using features from longer context windows violates the observation independence assumption of HMMs – a well known limitation [33] – we achieved improved accuracy using such features.

Table II shows the results of using a 33 hour subset of the training data. In this case, there are about 2,400 clustered triphone states in the GMM systems, corresponding to about 30,000 Gaussians. The PLDA and SGMM systems have a similar number of clustered triphone states, and a 400-component background model is used for each. The state vector of SGMMs and latent variables of PLDA are all 40-dimensional. We used 20,000 sub-states in the SGMM system, and for PLDA systems, we have also compared the results of using point or uncertainty estimation discussed in section III-A. All of these systems were trained using the maximum likelihood criterion without speaker adaptation.

After estimating the optimal system configurations for different acoustic models, we then performed experiments using the full training set of 300 hours (Table III). We used concatenated static MFCC_0 of 7 input frames, followed by LDA and STC transformation for the GMM system. There were around 8,000 clustered triphone states, with about 200,000 Gaussians. The SGMM system had 120,000 sub-states, and the PLDA system used 9 input frames of static MFCC_0. Again, they have a similar number of clustered triphone states as the GMM system.

V. Discussion and Conclusion

Our experimental results highlight the flexibility of the PLDA acoustic model which can use a variable number of potentially highly correlated input frames without requiring full covariance modelling. Compared to the GMM system using LDA and an STC transform, the PLDA system resulted in a lower WER given the same input features. In terms of likelihood computation, using uncertainty estimation leads to significantly lower WER compared to using point estimation. The PLDA systems obtained higher WER compared with the SGMM systems, however, using both the small and large training sets. As mentioned in Section II-B, we do not tie the state-dependent variables $\mathbf{z}_{jm}$ across all the components as in the SGMM, in order to scale the model easily to large training sets. However, this comes at the cost that we can not balance the size of model according to the amount of training data, in contrast to the sub-state splitting used in SGMMs. We may achieve higher recognition accuracy by using a similar method to tie the state dependent variables, an issue for future work.

In this letter, we have proposed a new acoustic modelling approach based on probabilistic linear discriminant analysis (PLDA). This model is able to use multiple input feature frames by using subspace covariance modelling. We have presented the algorithm and the training recipe to build a PLDA-based acoustic model, and have also shown some preliminary results on the Switchboard corpus which illustrates the potential of this model. In the future, we shall investigate other training approaches such as speaker adaptive training and discriminative training, as well as using different feature representations, for instance, the bottleneck features that are obtained from neural networks.

References

[1] S. Young, “A review of large-vocabulary continuous-speech recognition,” *Signal Processing Magazine, IEEE*, vol. 13, no. 5, p. 45, 1996.

[2] J.-L. Gauvain and L. Lamel, “Large-vocabulary continuous speech recognition: advances and applications,” *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1181–1200, 2000.

[3] M. Gales and S. Young, “The application of hidden markov models in speech recognition,” *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[4] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, “Probabilistic and bottle-neck features for lvcsr of meetings,” in *Proc. ICASSP*, vol. 4. IEEE, 2007.

[5] T. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, “Sparse representation features for speech recognition,” in *Proc. INTERSPEECH*, 2010, pp. 2254–2257.

[6] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[7] N. Morgan and H. Bourlard, “Generalization and parameter estimation in feedforward nets: Some experiments,” in *Proc. NIPS*, 1989, pp. 630–637.

[8] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, “Connectionist probability estimators in hmm speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 161–174, 1994.

[9] N. Morgan and H. A. Bourlard, “Neural networks for statistical recognition of continuous speech,” *Proceedings of the IEEE*, vol. 83, no. 5, pp. 742–772, 1995.

[10] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[11] S. J. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. ICCV*. IEEE, 2007, pp. 1–8.

[12] P. Kenny, “Bayesian speaker verification with heavy tailed priors,” in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[13] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, “Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification,” in *Proc. ICASSP*. IEEE, 2011, pp. 4828–4831.

[15] P. Kenny, “Joint factor analysis of speaker and session variability: theory and algorithms,” CRIM-06/08-13, Tech. Rep., 2005.

[16] J. J. Godfrey, E. C. Holliman, and J. McDaniel, “SWITCHBOARD: Telephone speech corpus for research and development,” in *Proc. ICASSP*. IEEE, 1992, pp. 517–520.

[17] T. Hasan and J. H. L. Hansen, “Acoustic factor analysis for robust speaker verificaiton,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 842–853, 2013.

[18] W.-L. Zhang, B.-C. Li, and W.-Q. Zhang, “Compact acoustic modeling based on acoustic manifold using a mixture of factor analyzers,” in *Proc. ASRU*, 2013.

[19] M. Gales, “Semi-tied covariance matrices for hidden Markov models,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[20] P. Bell and S. King, “Diagonal priors for full covariance speech recognition,” in *Proc. ASRU*. IEEE, 2009, pp. 113–117.

[21] W. Zhang and P. Fung, “Sparse inverse covariance matrices for low resource speech recognition,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, 2013.

[22] J. Huang, V. Goel, R. Gopinath, B. Kingsbury, P. A. Olsen, and K. Visweswariah, “Large vocabulary conversational speech recognition with the extended maximum likelihood linear transformation (EMLLT) model,” in *INTERSPEECH*, 2002.

[23] S. Axelrod, V. Goel, R. A. Gopinath, P. A. Olsen, and K. Visweswariah, “Subspace constrained gaussian mixture models for speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1144–1160, 2005.

[24] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, “The subspace Gaussian mixture model—A structured model for speech recognition,” *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[25] A. Rosti and M. Gales, “Factor analysed hidden markov models for speech recognition,” *Computer Speech & Language*, vol. 18, no. 2, pp. 181–200, 2004.

[26] L. Lu, A. Ghoshal, and S. Renals, “Cross-lingual subspace Gaussian mixture models for low-resouce speech recognition,” *IEEE/ACM Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 1, pp. 17–27, 2014.

[27] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, “Comparison of scoring methods used in speaker recognition with joint factor analysis,” in *Proc. ICASSP*. IEEE, pp. 4057–4060.

[28] X. Zhao and Y. Dong, “Variational bayesian joint factor analysis models for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 20, no. 3, pp. 1032–1042, 2012.

[29] Z. Ghahramani and G. Hinton, “The EM algorithm for mixtures of factor analyzers,” Technical Report CRG-TR-96-1, University of Toronto, Tech. Rep., 1996.

[30] C. Cieri, D. Miller, and K. Walker, “Research methodologies, observations and outcomes in (conversational) speech data collection,” in *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 206–211.

[31] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz, J. Silovský, G. Semmer, and K. Veselý, “The Kaldi speech recognition toolkit,” in *Proc. ASRU*, 2011.

[32] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, “Resegmentation of SWITCHBOARD,” in *Proc. ICSLP*, 1998.

[33] D. Gillick, L. Gillick, and S. Wegmann, “Don’t multiply lightly: Quantifying problems with the acoustic model assumptions in speech recognition,” in *Proc. ASRU*. IEEE, 2011, pp. 71–76.