Probabilistic Linear Discriminant Analysis with Bottleneck Features for Speech Recognition

Probabilistic Linear Discriminant Analysis with Bottleneck Features for Speech Recognition

Liang Lu, Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK

{liang.lu, s.renals}@ed.ac.uk

Abstract

We have recently proposed a new acoustic model based on probabilistic linear discriminant analysis (PLDA) which enjoys the flexibility of using higher dimensional acoustic features, and is more capable to capture the intra-frame feature correlations. In this paper, we investigate the use of bottleneck features obtained from a deep neural network (DNN) for the PLDA-based acoustic model. Experiments were performed on the Switchboard dataset — a large vocabulary conversational telephone speech corpus. We observe significant word error reduction by using the bottleneck features. In addition, we have also compared the PLDA-based acoustic model to three others using Gaussian mixture models (GMMs), subspace GMMs and hybrid deep neural networks (DNNs), and PLDA can achieve comparable or slightly higher recognition accuracy from our experiments.

Index Terms: speech recognition, bottleneck features, probabilistic linear discriminant analysis

1. Introduction

Deep neural network (DNN) approaches have recently produced significant increases in the accuracy of acoustic modelling for speech recognition, across a range of application domains and evaluation datasets [1, 2]. Compared to the hybrid neural network / hidden Markov model (HMM) architecture studied in the early 1990s [3, 4], DNNs typically use more hidden layers and a wider output layer. The deep architecture enables a DNN to learn more invariant and discriminative features before performing classification using the final softmax output layer. However, there has been only limited success in the adaptation of DNN-based acoustic models [5, 6], and in general they have to "learn by seeing" [7] — high recognition accuracy is usually obtained in matched training and test conditions. Thus hybrid DNN/HMM approaches may perform poorly in unseen acoustic conditions, especially if there is limited in-domain training data.

Tandem systems [8] use a neural network to provide features for a conventional Gaussian mixture model (GMM) based system, and a particular example of the neural network derived features is known as the bottleneck features [9, 10]. This method can take advantages of DNN feature extraction while enjoying the efficient adaptation algorithms for GMMs. However, GMMs typically employ diagonal covariance matrices, which limits their ability to learn feature correlations, as well as effectively restricting the bottleneck features to a limited dimensionality for computational reasons. As a result, a pre-processing approach such as principal component analysis (PCA) is often used to decorrelate and reduce the dimensionality of the bottleneck features.

We recently introduced a new acoustic model based on probabilistic linear discriminant analysis (PLDA) [11], which aims to overcome these constraints. It can be viewed as an extension of the GMM which is able to use higher dimensional feature vectors and can learn feature correlations in subspaces. PLDA was originally proposed for face recognition [12], and is now very well studied for speaker recognition using the i-vector framework [13, 14, 15]. PLDA is a probabilistic extension of linear discriminant analysis (LDA) [12]; similar to joint factor analysis (JFA) [16], PLDA factorizes the variability of the observations for a specific class (e.g. one speaker) using two latent variables: a within-class variable which is shared by all the observations of this class, and a between-class variable which is used to explain the variability to each observation. Furthermore when applied to speaker identification JFA operates in the GMM mean supervector domain, while the PLDA-based acoustic model directly operates in the acoustic feature domain.

We have previously demonstrated the feasibility of the PLDA-based acoustic model, its flexibility in using feature vectors of various dimensions, and its ability to learn feature correlations [11]. In this paper, we investigate the use of bottleneck features with PLDA-based acoustic models, in order to take the advantage of DNNs as feature extractors. In experiments on the Switchboard corpus [17], we compare this model to three other acoustic modelling approaches: GMMs, subspace GMMs (SGMMs) [18] and hybrid DNN/HMMs [1], and show that comparable or better recognition accuracy can be obtained.

2. PLDA-based Acoustic Model

The PLDA-based acoustic model is a generative model in which an acoustic feature vector $\mathbf{y}_t \in \mathbb{R}^d$ from the j-th HMM state at time index t is expressed as:

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

where $\mathbf{z}_j \in \mathbb{R}^q$ is the state variable (equivalent to the between-class identity variable in JFA) shared by the whole set of acoustic frames generated by the j-th state. $\mathbf{x}_{jt} \in \mathbb{R}^p$ is the observation variable (equivalent to the within-class channel variable in JFA) which explains the per-frame variability. Usually, the dimensionality of these two latent variables is smaller than that of the feature vector $\mathbf{y}_t$, i.e. $p \le d, q \le d$. $\mathbf{U} \in \mathbb{R}^{d\times p}$ and $\mathbf{G} \in \mathbb{R}^{d\times q}$ 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. $\mathbf{b} \in \mathbb{R}^d$ denotes the bias and $\epsilon_{jt} \in \mathbb{R}^d$ is the residual noise which is assumed to be Gaussian with zero mean and di-

Funded by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

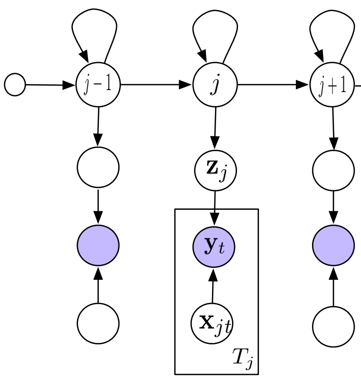

Figure 1: An illustration of a PLDA-HMM acoustic model, where the state variable $\mathbf{z}_j$ depends on j-th HMM state, and each observation $\mathbf{y}_t$ depends both on the state variable $\mathbf{z}_j$ and the observation variable $\mathbf{x}_{jt}$. The residual noise variable ϵ_{jt} is omitted for clarity.

agonal covariance. Figure 1 illustrates the concept of combining PLDA with an HMM for acoustic modelling.

Using a single PLDA has a limited modelling capacity since it only approximates a single Gaussian distribution. In [11], we used a mixture of PLDAs which can be written as

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

$$\epsilon_{jmt} \sim \mathcal{N}(\mathbf{0}, \Lambda_m) \tag{3}$$

where $1 \leq m \leq M$ is the component index. Denoting c to be the component indicator variable, the prior (weight) of each component is written as $P(c = m|j) = \pi_{jm}$. In this case, the model extends a conventional GMM by factorising the variability (2). By using low-rank matrices for $\mathbf{G}_m$ and $\mathbf{U}_m$, PLDA is more flexible in using higher dimensional feature inputs [11]. Moreover, it can learn feature correlations by using approximated full covariance matrices, which can be seen from the computation of the likelihood functions by marginalising out the observation variable $\mathbf{x}_{jmt}$ using its prior distribution:

$$\begin{aligned} & p(\mathbf{y}_t|\mathbf{z}_{jm}, j, m) \\ &= \int p(\mathbf{y}_t|\mathbf{x}_{jmt}, \mathbf{z}_{jm}, j, m) P(\mathbf{x}_{jmt}) d\mathbf{x}_{jmt} && (4) \\ &= \mathcal{N}\Big(\mathbf{y}_t; \mathbf{G}_m\mathbf{z}_{jm} + \mathbf{b}_m, \mathbf{U}_m\mathbf{U}_m^T + \Lambda_m\Big). && (5) \end{aligned}$$

where we have used $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as the prior distribution for $P(\mathbf{x}_{jmt})$, following the practice used for speaker and face recognition using JFA and PLDA [19, 20]. Note that the likelihood can be efficiently computed without inverting matrices $\mathbf{U}_m\mathbf{U}_m^T + \Lambda_m$ directly, but by using the following Woodbury matrix inversion lemma as in [19, 20]:

$$\begin{aligned} & (\mathbf{U}_m\mathbf{U}_m^T + \Lambda_m)^{-1} \\ &= \Lambda_m^{-1} - \Lambda_m^{-1}\mathbf{U}_m(\mathbf{I} + \mathbf{U}_m^T\Lambda_m^{-1}\mathbf{U}_m)^{-1}\mathbf{U}_m^T\Lambda_m^{-1} && (6) \\ &= \Lambda_m^{-1} - \mathbf{L}\mathbf{L}^T && (7) \end{aligned}$$

where $\mathbf{L} = \Lambda_m^{-1}\mathbf{U}_m(\mathbf{I} + \mathbf{U}_m^T\Lambda_m^{-1}\mathbf{U}_m)^{-1/2}$. This makes it computationally feasible when $\mathbf{y}_t$ is high dimensional. As discussed in [11] , this acoustic model is closely related to factor analysed HMMs [21] and SGMMs [18].

3. Model Training

A PLDA-based acoustic model may be trained using an EM algorithm [11], which is based on the assumption that the two latent variables $\mathbf{x}_{jmt}$ and $\mathbf{z}_{jm}$ are conditionally independent. This allows the updates of the projection matrices $\mathbf{U}_m$ and $\mathbf{G}_m$ to be interleaved. Conditional independence of the latent variables was not assumed when using PLDA for face recognition [12, 20], but a joint model training approach was used. For PLDA-based acoustic modelling, such kind of training algorithm can be derived by representing the model from stacking the T frames of j-th HMM state and m-th PLDA component:

$$\underbrace{\left[\begin{array}{c|c} \mathbf{y}_1 & \\ \mathbf{y}_2 & \\ \vdots & j,m \\ \mathbf{y}_T & \end{array}\right]}_{\bar{\mathbf{y}}|j,m} = \underbrace{\begin{bmatrix} \mathbf{G}_m & \mathbf{U}_m & \dots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \vdots \\ \mathbf{G}_m & \mathbf{0} & \dots & \mathbf{U}_m \end{bmatrix}}_{\bar{\mathbf{H}}_m} \underbrace{\begin{bmatrix} \mathbf{z}_{jm} \\ \mathbf{x}_{jm1} \\ \vdots \\ \mathbf{x}_{jmT} \end{bmatrix}}_{\bar{\mathbf{v}}_{jm}}$$

$$+ \underbrace{\begin{bmatrix} \mathbf{b}_m \\ \vdots \\ \mathbf{b}_m \end{bmatrix}}_{\bar{\mathbf{b}}_m} + \underbrace{\begin{bmatrix} \epsilon_{jm1} \\ \vdots \\ \epsilon_{jmT} \end{bmatrix}}_{\bar{\epsilon}_{jm}}$$

or in the form of the new notation

$$\bar{\mathbf{y}}|j,m = \bar{\mathbf{H}}_m\bar{\mathbf{v}}_{jm} + \bar{\mathbf{b}}_m + \bar{\epsilon}_{jm}. \tag{8}$$

This is a factor analysis model, and the model training algorithms for mixtures of factor analysers may be used [22, 16]. Unfortunately, this approach is computationally demanding for acoustic modelling, since the number of frames is normally very large, which results in significant computational and memory demands.

This difficulty can be circumvented if $\mathbf{x}_{jmt}$ and $\mathbf{z}_{jm}$ are assumed to be conditionally independent. In this case, the EM auxiliary function to update $\mathbf{U}_m$ is

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

where k is a constant that is independent of $\mathbf{U}_m$, γ_{jmt} denotes the component posterior probability $P(j,m|\mathbf{y}_t)$, and $\bar{\mathbf{z}}_{jm}$ denotes the mean of the posterior distribution of $\mathbf{z}_{jm}$ which is computed from the previous iteration. $\mathbb{E}[\cdot]$ is the expectation operation over the posterior distribution of $\mathbf{x}_{jmt}$, which can be computed from the Bayes' rule:

$$\begin{aligned} & P(\mathbf{x}_{jmt}|\mathbf{y}_t, \bar{\mathbf{z}}_{jm}, j, m) \\ &\quad = \frac{p(\mathbf{y}_t|\mathbf{x}_{jmt}, \bar{\mathbf{z}}_{jm}, j, m) P(\mathbf{x}_{jmt})}{\int p(\mathbf{y}_t|\mathbf{x}_{jmt}, \bar{\mathbf{z}}_{jm}, j, m) P(\mathbf{x}_{jmt}) d\mathbf{x}_{jmt}}. && (9) \end{aligned}$$

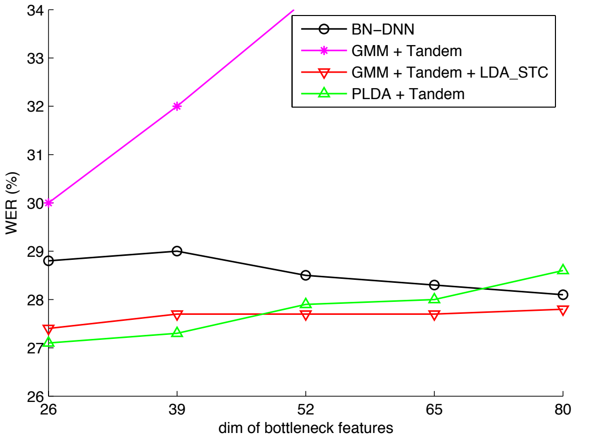

Figure 2: WERs on the Switchboard evaluation set using 33 hours of training data.

Using $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as the prior for $\mathbf{x}_{jmt}$, and with some algebraic rearrangement, we can obtain

$$P(\mathbf{x}_{jmt}|\mathbf{y}_t, \bar{\mathbf{z}}_{jm}, j, m) = \mathcal{N}(\mathbf{x}_{jmt}; \mathbf{V}_m^{-1}\mathbf{w}_{jmt}, \mathbf{V}_m^{-1}) \quad (10)$$

$$\mathbf{V}_m = \mathbf{I} + \mathbf{U}_m^T \Lambda_m^{-1} \mathbf{U}_m \quad (11)$$

$$\mathbf{w}_{jmt} = \mathbf{U}_m^T \Lambda_m^{-1} (\mathbf{y}_t - \mathbf{G}_m \bar{\mathbf{z}}_{jm} - \mathbf{b}_m) \quad (12)$$

We then set $\partial \mathcal{Q}(\mathbf{U}_m)/\partial \mathbf{U}_m = 0$ to obtain the update for $\mathbf{U}_m$

$$\mathbf{U}_m = \left(\sum_{jt} \gamma_{jmt} (\mathbf{y}_t - \mathbf{G}_m \bar{\mathbf{z}}_{jm} - \mathbf{b}_m) \mathbb{E}^T[\mathbf{x}_{jmt}] \right) \times \left(\sum_{jt} \gamma_{jmt} \mathbb{E}\left[\mathbf{x}_{jmt}\mathbf{x}_{jmt}^T\right] \right)^{-1} \quad (13)$$

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

4. Experiments

We have performed experiments using the Switchboard corpus [17], in which the training set comprises about 300 hours of conversational telephone speech. The Hub-5 Eval 2000 data [23] is used as the test set. The experiments were performed using the Kaldi speech recognition toolkit [24], which we have extended with an implementation of the PLDA-based acoustic model. Our current implementation does not yet support speaker adaptation and discriminative training, and hence in the following experiments, we have used maximum likelihood estimation and have not employed speaker adaptation (aside from speaker-based cepstral mean and variance normalisation). We used the pronunciation lexicon that was supplied by the Mississippi State transcriptions [25] which has more than 30,000 words, and a trigram language model was used for decoding.

4.1. Baseline systems

Table 1 shows the word error rates (WERs) of four types of speaker-independent acoustic models without sequence discriminative training: GMM, SGMM, PLDA, and hybrid DNN. Each model was trained using about 33 hours of Switchboard training data, and we show separate results for the Callhome (CHM) and Switchboard (SWB) evaluation sets. The number

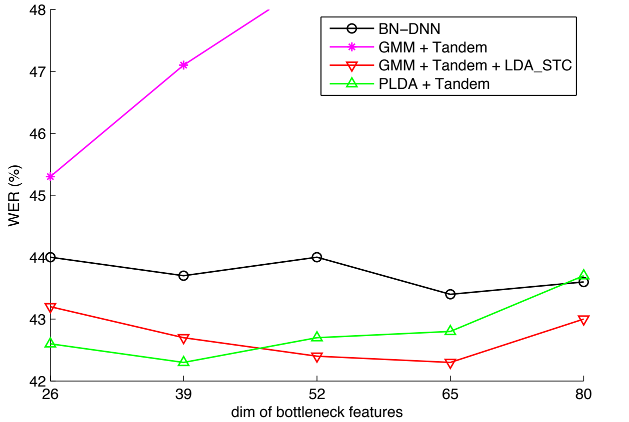

Figure 3: WERs on the CallHome evaluation set using 33 hours of training data.

Table 1: WERs (%) using 33 hours Switchboard training data

System	Feature	CHM	SWB	Avg
GMM	MFCC_0+Δ+ΔΔ	54.0	36.6	45.4
GMM	MFCC_0+LDA_STC (±3)	50.6	33.5	42.2
GMM	MFCC_0+LDA_STC (±4)	50.7	33.3	42.1
GMM	MFCC_0+LDA_STC (±5)	50.9	34.1	42.4
PLDA	MFCC_0 (±3)	49.5	32.4	41.1
PLDA	MFCC_0 (±4)	49.3	31.5	40.6
PLDA	MFCC_0 (±5)	49.7	33.2	41.6
PLDA	MFCC_0+Δ+ΔΔ (±1)	49.9	32.4	41.3
PLDA	MFCC_0+Δ+ΔΔ (±2)	52.2	34.0	43.1
SGMM	MFCC_0+Δ+ΔΔ	48.5	31.4	40.1
DNN hybrid	MFCC_0+Δ+ΔΔ (±4)	43.1	27.6	35.4

of tied HMM states is around 2,400 for each of the acoustic models shown in this table. The GMM system has about 30,000 Gaussian components. The number of components in the background model is 400 for both PLDA and SGMM systems. There were 20,000 sub-states (with 40-dimensional sub-state vectors) in the SGMM system. In the PLDA system, $\mathbf{x}_{jmt}$ and $\mathbf{z}_{jm}$ are also 40-dimensional. In the DNN system, the feature input was obtained by splicing 12-dimensional MFCCs with zeroth, delta and delta-delta coefficients (MFCC_0+Δ+ΔΔ) using a context size of 9 frames (i.e. ±4). Six hidden layers each with 1,024 nodes were used. The size of DNN output was the number of tied HMM states (around 2400), giving a total of about 8 million parameters. We have also studied different forms of feature input for GMM and PLDA acoustic models. From Table 1, we see that the best PLDA system has a consistently lower WER than the GMM systems with and without linear discriminant analysis (LDA) and semi-tied covariance matrix (STC) [26] when using spliced MFCC_0 as feature input. It is also comparable with SGMMs, but is more flexible with respect to the feature vector dimensionality. The DNN system has a significantly lower WER compared to the other acoustic models.

4.2. Bottleneck features

We trained bottleneck DNNs — using the same training data and the same kind of feature input — by reducing the size of the fifth hidden layer. We evaluated different sizes of the bottleneck layer, ranging from 26 to 80. The WERs of BN-DNNs were only slightly higher than the standard DNN system (0.5%

Table 2: WERs (%) using 33 hours Switchboard training data

System	Feature	CHM	SWB	Avg
DNN hybrid	MFCC_0+Δ+$\Delta\Delta$ ($\pm$4)	43.1	27.6	35.4
BN hybrid	MFCC_0+Δ+$\Delta\Delta$ ($\pm$4)	44.0	28.8	36.4
GMM	MFCC_0+Δ+$\Delta\Delta$	54.0	36.6	45.4
GMM	MFCC_0+LDA_STC ($\pm$3)	50.6	33.5	42.2
GMM	Tandem	44.8	30.9	37.9
GMM	Tandem + LDA_STC	43.2	27.4	35.3
SGMM	Tandem + LDA_STC	41.7	26.7	34.3
PLDA	Tandem	42.6	27.1	34.9

Table 3: WERs (%) using 109 hours Switchboard training data

System	Feature	CHM	SWB	Avg
DNN hybrid	MFCC_0+Δ+$\Delta\Delta$ ($\pm$4)	36.3	22.0	29.2
BN hybrid	MFCC_0+Δ+$\Delta\Delta$ ($\pm$4)	37.7	22.7	30.2
GMM	MFCC_0+Δ+$\Delta\Delta$	48.9	31.0	40.1
GMM	MFCC_0+LDA_STC ($\pm$3)	44.9	28.0	36.5
GMM	Tandem	39.7	25.5	32.6
GMM	Tandem + LDA_STC	36.7	22.1	29.5
SGMM	Tandem + LDA_STC	36.2	21.7	29.0
PLDA	Tandem	35.9	21.6	28.8

– 1.0% absolute). We then trained the Tandem GMM and PLDA systems by concatenating the bottleneck and 39-dimensional MFCC_0+Δ+$\Delta\Delta$ coefficients using the same system configuration as proposed in [8]. We also show results of using LDA to reduce the dimensionality of Tandem features to be 40 followed by STC to de-correlate the features. Figures 2 and 3 present the results using different sizes of bottleneck layers. Overall, increasing the size of bottleneck layer from 26 to 80[1] did not bring notable improvement to the three types of acoustic models. In fact, without LDA and STC transform, the "GMM+Tandem" system is prone to overfitting and may achieve much lower accuracy given higher dimensional bottleneck features. Again, the results demonstrates the flexibility of PLDA acoustic models in terms of using input feature vectors of varying dimension.

Table 2 summarises the results of using 33 hours of training data. Overall, the GMM system achieved a significant reduction in WER by using Tandem features. In addition, using the LDA and STC transforms for feature dimension reduction and de-correlation results in another $\sim$ 1.5% absolute WER reduction for the GMM-based system, which is comparable to the DNN hybrid system. The PLDA acoustic model can capture the feature correlations in the subspace, and outperforms the GMM system with LDA_STC transform using the same Tandem features. We also show results of the SGMM system with Tandem features. Since using full covariance matrices is computationally prohibitive and prone to model overfitting, we applied the LDA_STC transform obtained from the GMM system to the Tandem features before training the SGMM acoustic model. This was successful, with around 15% relative WER reduction compared to the results shown in Table 1, and it also slightly outperforms the PLDA and DNN systems.

4.3. Increased training data

To investigate whether the conclusion from the previous experiments holds in case of increased training data, we performed experiments using around 109 hours of Switchboard training data. In this case, we still used 6 hidden layers for the hybrid DNN system, but increase the size of each hidden layer to be 1200. The number of output nodes is around 4000, giving a total of approximately 12.5 million parameters. Again, the bottleneck DNN system has the same configuration as the hybrid DNN except that the size of the bottleneck layer is set to be 26. The GMM systems have around 90,000 Gaussian components, and the SGMM system has about 60,000 sub-state vectors. Again, $M = 400$ for the PLDA system, which is the same size as the UBM in the SGMM acoustic model. A summary of the results is given in Table 3 where we can see a similar trend as in Table 2. The "GMM+Tandem+LDA_STC" system is able to achieve almost the same WER as the hybrid DNN system, while the PLDA and SGMM systems can perform slightly better.

Table 4: WERs (%) using 109 hours Switchboard training data. Here, the bottleneck features were extracted from the bottleneck DNN system trained using 33 hours of data.

System	Feature	CHM	SWB	Avg
GMM	Tandem	41.7	26.3	34.0
GMM	Tandem + LDA_STC	39.0	24.1	31.6
SGMM	Tandem + LDA_STC	38.3	23.3	30.8
PLDA	Tandem	38.3	23.6	31.0

We then investigated the generalisation ability of the bottleneck features. We extracted the bottleneck features for the 109 hours of training data using the bottleneck DNN trained from 33 hours of data discussed in section 4.2. We then reproduced the Tandem systems using the GMM, SGMM and PLDA acoustic modelling approaches. The WERs are given in Table 4. By looking at the GMM systems, we observe that using such kind of bottleneck features can still lead to much higher recognition accuracy compared to using the MFCCs alone, yet the overall WERs are considerably higher than those reported in Table 3. This may indicate that using more and matched training data to train the bottleneck DNN to extract the features is beneficial for the Tandem system. Note that the comparison was performed using the Switchboard database which has limited differences in the acoustic conditions. We may expect a further drop in terms of recognition accuracy when there is a mismatch between the training and test conditions. However, as we discussed in section 1, there are efficient adaptation algorithms for acoustic models within the Gaussian family which can mitigate this problem. It is necessary to point out that the hybrid DNN system be can be significantly improved by using feature space MLLR (fMLLR) transformation [27] and sequence training criterion [28, 29, 30]. It is worthwhile to look at if the bottleneck features extracted from such kind DNN systems can further improve the Tandem system. In addition, the PLDA-based acoustic model may be further improved by tying the state variables across the components [11], which is one of our future works.

5. Conclusions

In this paper, we reviewed the recently proposed PLDA-based acoustic modelling approach, and investigated the use of bottleneck features from a DNN for this model. We demonstrated the flexibility of this model in making use of such kind of feature representation, and have obtained comparable or higher recognition accuracy compared to other acoustic model including GMMs, SGMMs, and hybrid DNNs. The current implementation of PLDA for acoustic modelling may be improved by sharing model parameters, e.g. through tying the state variables across the model components which would be analogous to state vectors used in SGMMs. Future work also include speaker adaptation and discriminative training for this model.

[1]The corresponding Tandem features are 65- to 119-dimensional.

6. References

[1] 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.

[2] 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.

[3] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer, 1994, vol. 247.

[4] 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.

[5] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. SLT*. IEEE, 2012, pp. 366–369.

[6] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 7893–7897.

[7] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," in *Proc. ICLR*, 2013.

[8] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*. IEEE, 2000, pp. 1635–1638.

[9] 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.

[10] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks." in *INTERSPEECH*, 2011, pp. 237–240.

[11] L. Lu and S. Renals, "Probabilistic linear discriminant analysis for acoustic modelling," *IEEE Signal Processing Letters*, 2014.

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

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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, 2009, pp. 4057–4060.

[20] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince, "Probabilistic models for inference about identity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, 2012.

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[28] A.-r. Mohamed, D. Yu, and L. Deng, "Investigation of full-sequence training of deep belief networks for speech recognition." in *INTERSPEECH*, 2010, pp. 2846–2849.

[29] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization." in *INTERSPEECH*, 2012.

[30] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. INTERSPEECH*, 2013.