Feature-space Speaker Adaptation for Probabilistic Linear Discriminant Analysis Acoustic Models

Liang Lu, Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
{liang.lu, s.renals}@ed.ac.uk

10.21437/Interspeech.2015-602

Abstract

Probabilistic linear discriminant analysis (PLDA) acoustic models extend Gaussian mixture models by factorizing the acoustic variability using state-dependent and observation-dependent variables. This enables the use of higher dimensional acoustic features, and the capture of intra-frame feature correlations. In this paper, we investigate the estimation of speaker adaptive feature-space (constrained) maximum likelihood linear regression transforms from PLDA-based acoustic models. This feature-space speaker transformation estimation approach is potentially very useful due to the ability of PLDA acoustic models to use different types of acoustic features, for example applying these transforms to deep neural network (DNN) acoustic models for cross adaptation. We evaluated the approach on the Switchboard corpus, and observe significant word error reduction by using both the mel-frequency cepstral coefficients and DNN bottleneck features.

Index Terms: speech recognition, probabilistic linear discriminant analysis, speaker adaptation, fMLLR, PLDA

1. Introduction

The most successful approaches to speaker adaptation are based on maximum likelihood linear regression (MLLR) transformations [1, 2], usually in the context of a conventional Gaussian mixture model (GMM) based acoustic model. Although hybrid deep neural network (DNN) / hidden Markov model (HMM) approaches now define the state-of-the-art for speech recognition [3, 4, 5, 6], it has been shown that feature-space MLLR (fMLLR) transforms computed from a GMM acoustic model can reduce the word error rate (WER) significantly when used with DNN acoustic models [4, 7]. A well known limitation of GMM-based acoustic models is that they only work well with low-dimensional and decorrelated acoustic features. Consequently, the application of fMLLR based speaker adaptation is limited to only a few types of acoustic features. For instance, DNN hybrid acoustic models can obtain higher recognition accuracy using the log spectral features compared to the widely used mel-frequency cepstral coefficients (MFCCs) [8], however, estimating fMLLR transforms from (correlated) log spectral features requires covariance modelling, which is challenging with large-scale conventional GMM systems.

To overcome the limitations of GMMs, we proposed an acoustic model based on probabilistic linear discriminant analysis (PLDA) [9, 10]. This model 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. For acoustic modelling, PLDA is used to model the HMM state density function directly, in contrast to its application to face or speaker recognition [11, 12, 13]. A PLDA acoustic model factorizes the acoustic variability using HMM state dependent variables which are expected to be consistent across different acoustic conditions, and observation dependent variables which characterise acoustic changes at the frame level [9]. Similar to a subspace GMM (SGMM) [14], the factorisation is based on the inference of subspaces. However, while the SGMM uses a set of full covariance matrices to directly model the per frame acoustic variability, the PLDA model introduces an additional set of projections to model this variability in lower-dimension subspaces, thus making it suitable for higher dimensional features.

Previously we have shown that PLDA is feasible for large vocabulary speech recognition, and that it can accommodate various types of acoustic feature [9, 15]. In this paper, we develop a speaker adaptation approach for PLDA acoustic models using fMLLR transforms. Since PLDA is a factorisation model for a Gaussian distribution, the conventional algorithm to estimate the fMLLR transforms for GMMs may still be applicable [2]. However, we face the difficulty of optimising speaker transforms for full covariance models, since PLDA approximates full covariances using a decomposition into low-rank matrices. There are numerous previous papers on this problem – for instance [16, 17, 18]. In this paper we study two approaches to circumvent this difficulty using the mean approximation and a sampling approach. Experiments were performed on the Switchboard corpus, using both MFCC and DNN bottleneck features. These approaches were demonstrated to be efficient and effective, providing substantial WER reduction with little additional computational cost.

2. PLDA-based Acoustic Model

The PLDA-based acoustic model [9] 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:

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

where $\mathbf{z}_j \in \mathbb{R}^q$ is the state variable 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 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. $p, q \leq d$. $\mathbf{U} \in \mathbb{R}^{d\times p}$ 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 linear discriminant analysis (LDA) formulation, but are estimated probabilistically.

Funded by EPSRC Programme Grant EP/I031022/1, Natural Speech Technology (NST). The NST research data collection may be accessed at http://datashare.is.ed.ac.uk/handle/10283/786.

Copyright © 2015 ISCA

$\mathbf{b} \in \mathbb{R}^d$ denotes the bias and $\epsilon_{jt} \in \mathbb{R}^d$ is the residual noise which is assumed to be Gaussian with zero mean and diagonal covariance Λ. By marginalising out the residual noise variable ϵ_{jt}, we obtain the following likelihood function:

$$p(\mathbf{y}_t|j) = \mathcal{N}(\mathbf{y}_t; \mathbf{U}\mathbf{x}_{jt} + \mathbf{G}\mathbf{z}_j + \mathbf{b}, \Lambda)\,. \quad (2)$$

And if we also marginalise out the observation variable x_{jt}, we obtain:

$$p(\mathbf{y}_t|j) = \mathcal{N}(\mathbf{y}_t; \mathbf{G}\mathbf{z}_j + \mathbf{b}, \mathbf{U}\mathbf{U}^T + \Lambda), \quad (3)$$

which approximates the full covariance model by the low-rank matrix $\mathbf{U}$. Our previous results demonstrate that this approach can achieve a significant decrease in WER [9].

2.1. Mixture and tied models

A single PLDA has a limited modelling capacity since it only approximates a single Gaussian distribution. The modelling capacity can be increased by using mixture models, for example, an M-component PLDA mixture model results in the following component distribution:

$$\mathbf{y}_t|j, m = \mathbf{U}_m\mathbf{x}_{jmt} + \mathbf{G}_m\mathbf{z}_{jm} + \mathbf{b}_m + \epsilon_{jmt}, \quad (4)$$
$$\epsilon_{jmt} \sim \mathcal{N}(\mathbf{0}, \Lambda_m)\,. \quad (5)$$

If c is the component indicator variable, then the prior (weight) of each component is $P(c = m|j) = \pi_{jm}$. Given the latent variables $\mathbf{x}_{jmt}$ and $\mathbf{z}_{jm}$, the state-level distribution over features is:

$$p(\mathbf{y}_t|j) = \sum_m \pi_{jm}\mathcal{N}(\mathbf{y}_t; \mathbf{U}_m\mathbf{x}_{jmt} + \mathbf{G}_m\mathbf{z}_{jm} + \mathbf{b}_m, \Lambda_m)\,.$$

Since the projection matrices $\mathbf{U}_m$ and $\mathbf{G}_m$ are globally shared, a large number of components can be used to improve the model capacity, e.g. $M = 400$ [9].

Previously, we observed that PLDA mixture models are prone to overfitting when using a large number of components. Motivated by the parameter sharing approach in SGMMs [14], we studied a tied version of the PLDA mixture model [19], in which the state variable $\mathbf{z}_{jm}$ is tied across all the components for each HMM state:

$$\mathbf{y}_t|j, m = \mathbf{U}_m\mathbf{x}_{jmt} + \mathbf{G}_m\mathbf{z}_j + \mathbf{b}_m + \epsilon_{jmt}, \quad (6)$$
$$\epsilon_{jmt} \sim \mathcal{N}(\mathbf{0}, \Lambda_m)\,. \quad (7)$$

This approach can significantly reduce the number of state-dependent parameters as well as the computational cost. Li et al [10] presented a similar approach, which was applied to face recognition. Using a global state variable may over-simplify the model, so a further mixing-up strategy can be applied similar to SGMMs [14]. We refer the readers to [19] for further details. In this paper, to simplify the notation, we use the PLDA mixture model for the discussion of speaker adaptation, but the extension to tied PLDA models is straightforward.

3. Estimation of fMLLR transforms

For a PLDA mixture model, learning fMLLR transforms is equivalent to speaker adaptation with full covariance GMMs if we use a likelihood evaluation function which marginalises out the observation latent variable x_{jmt} (3). To simplify the problem, in this work we derive the auxiliary function of fMLLR transforms without marginalization of $\mathbf{x}_{jmt}$; this latent variable is marginalised out only when estimating acoustic model parameters and decoding given the fMLLR transforms. There is an obvious mismatch between fMLLR estimation and application; however, we find the approach works reasonably well. Our approach relies on a point estimate of $\mathbf{x}_{jmt}$, detailed below.

3.1. Posterior mean approximation

The first approach is to use the posterior mean of $\mathbf{x}_{jmt}$ as the point estimate to evaluate the likelihood function during fMLLR estimation, which gives:

$$p(\mathbf{y}_t|j, m, \mathcal{T}) = |\mathbf{A}|\mathcal{N}(\hat{\mathbf{y}}_t; \mathbf{U}_m\mathbf{x}_{jmt} + \mathbf{G}_m\mathbf{z}_{jm} + \mathbf{b}_m, \Lambda_m), \quad (8)$$

where $\mathcal{T}$ denotes the transformation parameters $(\mathbf{A}, \mathbf{c})$, $\hat{\mathbf{y}}_t$ is the adapted feature vector $\hat{\mathbf{y}}_t = \mathbf{A}\mathbf{y}_t + \mathbf{c}$, and $\mathbf{x}_{jmt}$ is the mean of its posterior distribution which can be obtained as:

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

The rationale behind this approach is to move the model along the basis spanned by $\mathbf{U}_m$ to explain the adapted observation better. Using the prior distribution $P(\mathbf{x}_{jmt}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ [1] we obtain:

$$P(\mathbf{x}_{jmt}|\hat{\mathbf{y}}_t, \mathbf{z}_{jm}, j, m) = \mathcal{N}(\mathbf{x}_{jmt}; \mathbf{V}_m^{-1}\mathbf{w}_{jmt}, \mathbf{V}_m^{-1}) \quad (10)$$
$$\mathbf{V}_m = \mathbf{I} + \mathbf{U}_m^T\Lambda_m^{-1}\mathbf{U}_m \quad (11)$$
$$\mathbf{w}_{jmt} = \mathbf{U}_m^T\Lambda_m^{-1}(\hat{\mathbf{y}}_t - \mathbf{G}_m\mathbf{z}_{jm} - \mathbf{b}_m)\,. \quad (12)$$

By replacing $\mathbf{x}_{jmt} = \mathbf{V}_m^{-1}\mathbf{w}_{jmt}$, we can write the auxiliary function to optimise $\mathcal{T}$ as follows:

$$\mathcal{Q}(\mathcal{T}) = \sum_{jmt}\gamma_{jmt}\Big(-\frac{1}{2}(\hat{\mathbf{y}}_t - \boldsymbol{\mu}_{jmt})^T\Lambda_m^{-1}(\hat{\mathbf{y}}_t - \boldsymbol{\mu}_{jmt}) + \log|\mathbf{A}|\Big) + \text{const}, \quad (13)$$

where $\boldsymbol{\mu}_{jmt} = \mathbf{U}_m\mathbf{x}_{jmt} + \mathbf{G}_m\mathbf{z}_{jm} + \mathbf{b}_m$, and γ_{jmt} is the posterior probability. Since Λ_m is diagonal, the auxiliary function can be optimised using the conventional approach [2].

3.2. Prior mean approximation

The posterior mean approximation approach relies on the estimation of the posterior distribution of $\mathbf{x}_{jmt}$ given the observation $\hat{\mathbf{y}}_t$. However, the posterior distribution is usually very flat since it is computed only using one observation, and the estimation is likely to be inaccurate. Our previous work [9] shows that the posterior mean approximation approach performs significantly worse than marginalising out $\mathbf{x}_{jmt}$ using its prior distribution in speaker-independent systems. Therefore, we consider the other approach that uses the prior rather than the posterior of $\mathbf{x}_{jmt}$ for mean approximation. This means that no knowledge from the observations will be used and it may be potentially more robust. Since we use $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as the prior, the likelihood function in this case is:

$$p(\mathbf{y}_t|j, m, \mathcal{T}) = |\mathbf{A}|\mathcal{N}(\hat{\mathbf{y}}_t; \mathbf{G}_m\mathbf{z}_{jm} + \mathbf{b}_m, \Lambda_m)\,. \quad (14)$$

Estimating the transform parameters $\mathcal{T}$ is then straightforward.

[1] Note that using $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as a prior is reasonable since, after convergence, a non-zero mean can be accounted for by $\mathbf{b}_m$, and the variance can be modified by rotating and scaling the matrix $\mathbf{U}_m$.

3.3. Unscented transform based sampling approximation

Finally, we may use a sampling approach to approximate the integrated likelihood function. For instance, using a naive sampling approach:

$$p(\mathbf{y}_t|j,m,\mathcal{T}) = \sum_{k=0}^{K} w_k |\mathbf{A}| \mathcal{N}(\hat{\mathbf{y}}; \mathbf{U}_m \mathbf{x}_{jmt}^k + \mathbf{G}_m \mathbf{z}_{jm} + \mathbf{b}_m, \Lambda_m), \tag{15}$$

where $\mathbf{x}_{jmt}^1, \cdots, \mathbf{x}_{jmt}^K$ are samples independently drawn from the prior distribution $P(\mathbf{x}_{jmt}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, and w_k is the sampling weight. Given an infinite number of samples, the approximation error is close to zero which is equivalent to marginalising out $\mathbf{x}_{jmt}$ using its prior distribution. However, this is infeasible for computational reasons. In this work we employ the unscented transform (UT) [20], a deterministic sampling approach which draws many fewer samples compared to alternative sampling approaches. UT draws samples deterministically from the *sigma points* – a set of points chosen to have the same mean and covariance as the original distribution. This approach has been used in noise adaptation for robust speech recognition [21, 22, 23, 24]. For a Gaussian distribution $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, UT draws samples as

$$\mathbf{x}_0 = \boldsymbol{\mu}, \tag{16}$$

$$\mathbf{x}_i = \boldsymbol{\mu} + \left[\sqrt{(2d+\kappa)\boldsymbol{\Sigma}}\right]_i, \tag{17}$$

$$\mathbf{x}_{i+p} = \boldsymbol{\mu} - \left[\sqrt{(2d+\kappa)\boldsymbol{\Sigma}}\right]_i, \tag{18}$$

where $i = 1, \ldots, d$, and $\sqrt{\boldsymbol{\Sigma}}$ and $[\boldsymbol{\Sigma}]_i$ denote the Cholesky decomposition and i_{th} column of the covariance matrix $\boldsymbol{\Sigma}$ respectively. κ is a tuning parameter, p is the dimensionality of $\mathbf{x}$, where the weights are defined in UT as

$$w_0 = \frac{\kappa}{p+\kappa}, \quad w_i = \frac{1}{2(p+\kappa)}. \tag{19}$$

In this work, we set $\kappa = 1/2$ to give the equal weight to all the samples [20]. Again, the estimation of fMLLR transforms is readily available using equation (15).

4. Experiments

The experiments were performed using the Switchboard corpus [25], a conversational telephone speech database released by LDC with the catalog number as LDC97S62. The training set comprises about 300 hours of conversational telephone speech, and the Hub-5 Eval 2000 data [26] is used as the test set. Our systems were built using the Kaldi speech recognition toolkit [27] with the additional implementation of the PLDA acoustic model. In the following experiments, we used tied PLDA, which works better and is also computational cheaper compared to a PLDA mixture model [9]. We only report the results using speaker adaptive training (SAT) [28] which interleaves the estimation of the fMLLR transforms and the acoustic model parameters. We used the 30,000 word pronunciation lexicon that was supplied with the Mississippi State transcriptions [29], and a trigram language model for decoding.

4.1. MFCC features

We first show baseline results for GMM and PLDA systems using MFCC features in Table 1. These initial systems were trained using about 33 hours of Switchboard training data, and we show separate results for the CallHome (CHM) and Switchboard (SWB) evaluation sets. We have used MFCCs with dynamic features which are 39-dimensional, and also spliced static 13-dim MFCC_0 with or without LDA and semi-tied covariance(STC) modelling [30]. As demonstrated previously [14], the PLDA acoustic model is applicable to higher dimensional feature vectors, but LDA and STC transforms are still beneficial to this system using MFCCs.

We used the MFCC_0($\pm$3)+LDA_STC features for SAT systems. The initial fMLLR transform was estimated from the speaker-independent acoustic model, and then the acoustic model was retrained for 4 iterations before updating the fMLLR transforms. The process was repeated until convergence was reached. Figure 1 illustrates the training process and the results of fMLLR re-estimation, where the prior mean approximation training approach was used. We observed consistent improvement in both log likelihood scores and recognition accuracies. The SAT system converges after 2–3 re-estimations of the fMLLR transforms. Table 2 shows the results of using the three proposed fMLLR estimation approaches discussed in Section 3. As a comparison, we also report the result of cross adaptation where the fMLLR transforms were borrowed from the GMM system of Table 1. Similar to our previous results [9], the

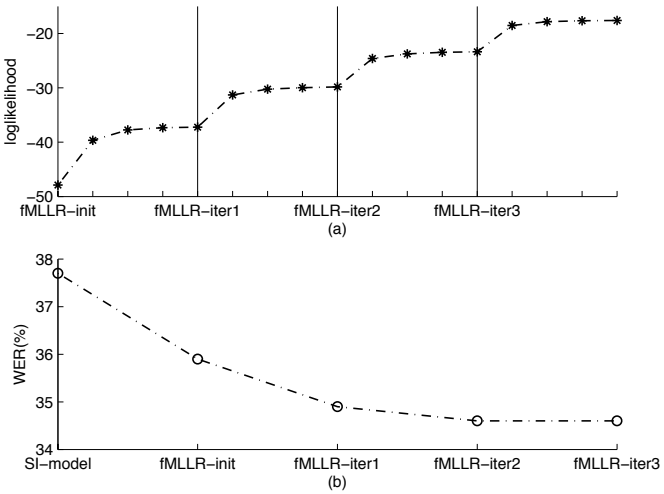

Figure 1: The speaker adaptive training process for a PLDA system. A speaker-independent (SI) acoustic model was used to estimate the initial fMLLR transform, and then the acoustic model was retrained given the speaker transforms for 4 iterations before updating the fMLLR transforms. (a) shows the improvement of the log-likelihood on the training data after updating the transforms and acoustic model. (b) shows the reduction of the word error rate.

Table 1: WERs (%) of baseline systems using 33 hours Switchboard training data.

System	Feature	Dim	CHM	SWB	Avg
GMM	MFCC_0+Δ+$\Delta\Delta$	39	54.0	36.6	45.4
GMM	MFCC_0($\pm$3)+LDA_STC	40	50.6	33.5	42.2
+ SAT	MFCC_0($\pm$3)+LDA_STC	40	43.2	27.9	35.6
PLDA	MFCC_0 ($\pm$3)	91	49.5	32.4	41.1
PLDA	MFCC_0 ($\pm$4)	117	49.3	31.5	40.6
PLDA	MFCC_0 ($\pm$5)	143	49.7	33.2	41.6
PLDA	MFCC_0($\pm$3)+LDA_STC	40	45.7	29.5	37.7

Table 2: Speaker adaptive training results of PLDA systems using 33 hours Switchboard training data. The acoustic features are 40-dim MFCC_0($\pm$3)+LDA_STC.

System	CHM	SWB	Avg
Posterior mean approximation	43.1	28.3	35.7
Prior mean approximation	42.1	27.0	34.6
UT-based sampling	41.8	27.3	34.6
Cross adaptation	39.6	25.3	32.5

Table 3: WERs (%) using 300 hours Switchboard training data

System	Feature	CHM	SWB	Avg
DNN hybrid	MFCC_0+Δ+$\Delta\Delta$ ($\pm$5)	28.5	14.9	21.8
BN hybrid	MFCC_0+Δ+$\Delta\Delta$ ($\pm$5)	28.4	15.2	21.9
GMM	MFCC_0($\pm$3)+LDA_STC	41.9	25.0	33.5
+SAT	MFCC_0($\pm$3)+LDA_STC	36.3	21.4	28.9
+bMMI	MFCC_0($\pm$3)+LDA_STC	33.1	18.5	25.9
GMM	BN	31.7	18.6	25.2
+SAT	BN	30.6	18.4	24.5
+bMMI	BN	29.7	17.9	23.8
GMM	BN + STC	31.7	18.0	24.9
+SAT	BN + STC	30.1	17.7	23.9
PLDA	BN	30.6	17.2	24.0
+SAT	BN	28.5	16.3	22.5
+bMMI	BN	27.4	15.3	21.4

posterior mean approximation approach is considerably worse than the other two approaches, while UT-based sampling did not considerably outperform the prior mean approximation approach, which was not expected. In the future, we shall investigate other sampling approaches. Finally, we observed that the fMLLR transforms from the GMM system resulted in much lower WER. This is probably because the features were decorrelated by STC, which is beneficial to GMMs with diagonal covariances to estimate the fMLLR transforms.

4.2. Bottleneck features

To evaluate the speaker adaptation approaches on stronger systems, we trained the PLDA system using DNN bottleneck features [31, 32] on the full Switchboard training set of 300 hours. Our bottleneck features were extracted from a DNN acoustic model, with 6 hidden layers of 2048 hidden units, except the fifth hidden layer which is the bottleneck layer of 40 hidden units. In order to evaluate the effect of speaker adaptation in PLDA and GMMs using the bottleneck features, we did not apply a feature space speaker transformation when training the bottleneck DNN, in case that the bottleneck features themselves were already speaker normalised. The input features for the DNNs were 11 concatenated 39-dimensional MFCC vectors. We also report results using the standard DNN hybrid model without a bottleneck layer (Table 3). We observed that the two DNN systems achieved similar WERs.

We used the prior mean approximation approach to estimate the fMLLR transforms for the PLDA+SAT acoustic model. To further improve the accuracy, the model was then trained discriminatively using the boosted MMI criterion [33] for 3 iterations. Table 3 compares the results of PLDA and GMM systems using bottleneck features, from which we can see that the PLDA system can consistently obtain around 10% relative improvement, and is comparable or slightly better than the DNN baseline. As a comparison, we have also shown the WERs of GMM systems using the MFCC features. We observe that improvement form GMM+SAT over the speaker independent system is much larger relatively for MFCC features than bottleneck features, probably because of the decorrelating effect of the global STC transform. To validate this hypothesis, we retrained the GMM system using bottleneck features, followed by a global STC transform. The improvement from SAT is relatively larger in this case; however, the relative improvement is still much smaller than using the MFCC features, and the GMM+STA+STC system is still worse than the PLDA+SAT system. We have also trained the PLDA+SAT system by cross adaptation where the fMLLR transforms were borrowed from the GMM system. In this case, we obtained the average WER 22.4%, which is comparable to PLDA+SAT using self-estimated fMLLR transforms.

Table 4: System combination and related results

System	CHM	SWB	Avg
DNN hybrid baseline	28.5	14.9	21.8
GMM$\oplus$PLDA	27.1	15.7	21.4
DNN$\oplus$GMM	27.6	15.8	21.7
DNN$\oplus$PLDA	26.4	14.2	20.5
DNN$\oplus$PLDA$\oplus$GMM	26.2	14.4	20.4
DNN hybrid (36M)[35]	27.1	15.1	21.2
DNN hybrid (36M) + dropout [35]	26.7	14.7	20.8
DNN hybrid (100M)[35]	26.7	14.7	20.7
DNN hybrid (100M) + dropout[35]	26.3	14.6	20.5
DNN hybrid + sMBR [37]	-	13.3	-
DNN hybrid + fMLLR + sMBR [38]	24.1	12.6	18.4

4.3. System combination

Finally, we investigate if the GMM, DNN and PLDA systems are complementary to each other. We used minimum Bayes risk decoding [34] implemented in Kaldi for system combination by combing the word lattices from each sub-system. As shown in Table 4, combining the GMM system can only marginally improve the DNN and PLDA systems on the CallHome subset, while making them worse on the Switchboard subset. The PLDA system, however, is highly complementary to the DNN system, and results in 1.3% absolute improvement on average. We have also shown some recent reported results on Switchboard using DNN hybrid acoustic models. In [35], it is shown that increasing the size of DNN from 36 million to 100 million parameters and using dropout regularisation [36] only marginally improve the accuracy, while sequence training and feature space speaker adaptation can significantly reduce the WER [37, 38]. In the future, we shall investigate applying our fMLLR transforms for feature space adaptation of DNN hybrid models using different types of features.

5. Conclusions

In this paper, we have developed a speaker adaptation approaches for PLDA acoustic models. Our methods circumvent the difficulty and complexity of full covariance speaker adaptation by taking point approximations of the observation latent variable when estimating the fMLLR transformations. Our approaches were evaluated on the Switchboard conversational telephone speech transcription task, and we have studied both MFCC and DNN bottleneck features. The fMLLR estimation approaches are simple to implement, require little additional computation, and our results demonstrate that these approaches are efficient. Given the flexibility of PLDA acoustic models in using different types of acoustic features, in the future we shall investigate feature-space adaptation for other feature types (e.g. log filter-bank features) and study cross adaptation for DNN hybrid models.

6. References

[1] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, no. 2, p. 171, 1995.

[2] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[3] 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.

[4] 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," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

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

[6] 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.

[7] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Proc IEEE ICASSP*, 2013.

[8] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at Microsoft," in *Proc. IEEE ICASSP*, 2013, pp. 8604–8608.

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

[10] 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.

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

[12] 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.

[13] 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. IEEE ICASSP*, 2011, pp. 4828–4831.

[14] 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.

[15] L. Lu and S. Renals, "Probabilistic linear discriminant analysis with bottleneck features for speech recognition," in *Proc. INTERSPEECH*, 2014.

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

[17] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance gaussians." in *Proc. INTERSPEECH*, 2006.

[18] A. Ghoshal, D. Povey, M. Agarwal, P. Akyazi, L. Burget, K. Feng, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "A novel estimation of feature-space MLLR for full-covariance models," in *Proc. IEEE ICASSP*, 2010, pp. 4310–4313.

[19] L. Lu and S. Renals, "Tied probabilistic linear discriminant analysis for speech recognition," *arXiv:1411.0895 [cs.CL]*, 2014.

[20] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[21] Y. Hu and Q. Huo, "An HMM compensation approach using unscented transformation for noisy speech recognition," *Proc. ISCSLP*, pp. 346–357, 2006.

[22] J. Li, D. Yu, Y. Gong, and L. Deng, "Unscented transform with online distortion estimation for HMM adaptation," in *Proc. INTERSPEECH*, 2010.

[23] F. Faubel, J. McDonough, and D. Klakow, "On expectation maximization based channel and noise estimation beyond the vector taylor series expansion," in *Proc IEEE ICASSP*, 2010.

[24] L. Lu, A. Ghoshal, and S. Renals, "Joint uncertainty decoding with unscented transform for noise robust subspace Gaussian mixture models," in *Proc. SAPA-SCALE Workshop*, 2012.

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

[26] 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.

[27] 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.

[28] Y. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996.

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

[30] 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.

[31] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE ICASSP*, 2007.

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

[33] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, 2008, pp. 4057–4060.

[34] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[35] A. Maas, P. Qi, Z. Xie, A. Hannun, C. T. Lengerich, D. Jurafsky, and A. Y. Ng, "Building DNN Acoustic Models for Large Vocabulary Speech Recognition," in *arXiv preprint arXiv:1406.7806v2*, 2015.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] 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.

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