Sequential Maximum Mutual Information Linear Discriminant Analysis for Speech Recognition

Yuuki Tachioka[1], *Shinji Watanabe*[2], *Jonathan Le Roux*[2], *and John R. Hershey*[2]

[1]Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura, Japan
[2]Mitsubishi Electric Research Laboratories, Cambridge, US
Tachioka.Yuki@eb.MitsubishiElectric.co.jp, {watanabe,leroux,hershey}@merl.com

Abstract

Linear discriminant analysis (LDA) is a simple and effective feature transformation technique that aims to improve discriminability by maximizing the ratio of the between-class variance to the within-class variance. However, LDA does not explicitly consider the sequential discriminative criterion which consists in directly reducing the errors of a speech recognizer. This paper proposes a simple extension of LDA that is called sequential LDA (sLDA) based on a sequential discriminative criterion computed from the Gaussian statistics, which are obtained from sequential maximum mutual information (MMI) training. Although the objective function of the proposed LDA can be regarded as a special case of various discriminative feature transformation techniques (for example, f-MPE or the bottom layer of a neural network), the transformation matrix can be obtained as the closed-form solution to a generalized eigenvalue problem, in contrast to the gradient-descent-based optimization methods usually used in these techniques. Experiments on large vocabulary continuous speech recognition (Corpus of Spontaneous Japanese) and noisy speech recognition task (2nd CHiME challenge) show consistent improvements from standard LDA due to the sequential discriminative training. In addition, the proposed method, despite its simple and fast computation, improved the performance in combination with discriminative feature transformation (f-bMMI), perhaps by providing a good initialization to f-bMMI.

Index Terms: Maximum mutual information, linear discriminant analysis, region dependent linear transformation

1. Introduction

Feature transformation with dimensionality reduction is usually the first step in the front-end pipeline for automatic speech recognition (ASR). Such methods allow the use of long-context features that can consider the influence across multiple frames directly instead of using traditional delta features. One of the simplest and widely used methods has been linear discriminant analysis (LDA) [1], which maximizes the ratio of the between-class variance to the within-class variance, where the classes are typically derived from the context-dependent phoneme states. An advantage of LDA is that it provides a simple and efficient closed-form solution to estimate the transformation. One of its limitations is the assumption of equal covariance for the classes. To relax the constraint of equal covariance of LDA, heteroscedastic discriminant analysis (HDA) and heteroscedastic LDA (HLDA) have been proposed [2, 3].

Another limitation of LDA is the lack of explicit consideration of speech recognizer (decoder) outputs. The purpose of feature transformation is essentially to provide features appropriate for recognition. LDA aims to improve discriminability of features but standard LDA deals the same way with classes that are easy to distinguish for the recognizer as with classes that are difficult to classify (i.e., easy to confuse).

Owing to the recent progress in discriminative training methods, it is well known that a sequential discriminative training with recognizer error tendencies is effective for various conventional techniques such as acoustic modeling or feature space discriminative training. Maximum mutual information (MMI) criterion [4] or minimum phoneme error (MPE) criterion [5] are effective training criteria because they consider the patterns of error at the recognition level, in order to focus on distinguishing the most important states. During training, these methods typically employ extended-Baum-Welch (EBW) updates, where the sufficient statistics for model parameter estimation are based on functions of the posterior probabilities of the recognition word sequences. Feature transformation based on such methods can improve the ASR performance further.

Linear feature transformation generally consists of projection matrices and offset terms. LDA is a global (single region) linear projection with no offsets. In contrast, region dependent linear transformation (RDLT) [6] first divides the feature space into regions, and for each region separate transforms can be applied. Discriminative approaches such as MPE-HLDA [7], which is an extension of HLDA based on the MPE criterion, feature space MPE (f-MPE) [8], and MMI-SPLICE [9], have been proposed. Such methods typically require iterative gradient-descent optimization. Typically, LDA features are still used as input to such methods since they are simple to compute and provide a reasonable starting point.

The proposed method is an extension of standard LDA based on the MMI objective function in order to consider the recognition posteriors when feature statistics are calculated. The advantages of the proposed method are the existence of a closed-form solution, and the simplicity of implementation which amounts to a simple modification of the sufficient statistics computation.

This paper first describes in Section 2 the conventional LDA [1], mainly from the perspective described in [2, 10]. Next, our proposed MMI approach is described in Section 3. Experiments on two different tasks show that the proposed method improves speech recognition performance on two data sets in Section 4.

2. Maximum Likelihood LDA

When $\boldsymbol{x}_t \in \mathbb{R}^n$ is the tth frame n-dimensional input feature, which is usually obtained by concatenating original MFCC features of contiguous several frames, LDA feature transformation

[1, 10] transforms $\boldsymbol{x}_t$ to lower dimensional feature $\boldsymbol{y}_t \in \mathbb{R}^p$ as

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t, \tag{1}$$

where $\boldsymbol{A}$ is an LDA feature transformation matrix whose dimension is $p \times n$, which $(p < n)$. The objective function of LDA is given as

$$\arg\max_{\boldsymbol{A}} \frac{|\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^\top|}{|\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^\top|}, \tag{2}$$

where $\top$ denotes a transposition and $\boldsymbol{B}$ and $\boldsymbol{W}$ denote $n \times n$ between-class scatter matrix and within-class scatter matrix as defined in Eq. (3), respectively:

$$\begin{aligned} \boldsymbol{B} &= \frac{1}{\sum_j N_j} \sum_j N_j \boldsymbol{\mu}_j^{\mathrm{x}} (\boldsymbol{\mu}_j^{\mathrm{x}})^\top - \bar{\boldsymbol{\mu}}^{\mathrm{x}} (\bar{\boldsymbol{\mu}}^{\mathrm{x}})^\top, \\ \boldsymbol{W} &= \frac{1}{\sum_j N_j} \sum_j N_j \boldsymbol{\Sigma}_j^{\mathrm{x}}, \end{aligned} \tag{3}$$

where $\boldsymbol{\mu}^{\mathrm{x}}$ and $\boldsymbol{\Sigma}^{\mathrm{x}}$ are the mean vector and co-variance matrix in the original vector $\boldsymbol{x}$ space, N_j is the count of elements which belong to the j-th class, and $\bar{\boldsymbol{\mu}}^{\mathrm{x}}$ is the average of all vectors $\boldsymbol{\mu}_j^{\mathrm{x}}$. Generally, $\boldsymbol{\mu}_j^{\mathrm{x}}$ and $\boldsymbol{\Sigma}_j^{\mathrm{x}}$ are computed [10] for class j as

$$\begin{aligned} N_j &= \sum_t \psi_t(j), \\ \boldsymbol{\mu}^{\mathrm{x}} &= \frac{1}{N_j} \sum_t \psi_t(j) \boldsymbol{x}_t, \\ \boldsymbol{\Sigma}_j^{\mathrm{x}} &= \frac{1}{N_j} \sum_t \psi_t(j) \boldsymbol{x}_t \boldsymbol{x}_t^\top - \boldsymbol{\mu}_j^{\mathrm{x}} (\boldsymbol{\mu}_j^{\mathrm{x}})^\top, \end{aligned} \tag{4}$$

where, $\psi_t(j)$ are class membership weights relating $\boldsymbol{x}_t$ to class j. In the classic LDA, the class assignments, given by $j = l(t)$, are hard, so $\psi_t(j)$ can be defined as:

$$\psi_t(j) = \begin{cases} 1\text{: } l(t) = j, \\ 0\text{: otherwise.} \end{cases} \tag{5}$$

Here, we assume that LDA class j is related to the HMM state number as in the most general case. In this case, alignments by the HMM model correspond to the class label.

A solution to LDA is obtained by solving the following generalized eigenvalue problem [11],

$$\boldsymbol{B}v = \lambda \boldsymbol{W}v, \tag{6}$$

and assigning to the rows of $\boldsymbol{A}$ the eigenvectors $v_{1:p}^T$ corresponding to the p largest eigenvalues $\lambda_{1:p}$.

It has been shown by Kumar *et al*. that standard LDA has the same optimum as a maximum likelihood problem [2]. In this problem, the model has tied state-dependent variances in $y_t = \boldsymbol{A}x_t$, and the mean and variance in the orthogonal subspace $y'_t = \boldsymbol{A}'x_t$ are state-independent, where $\boldsymbol{A}'$ is an $(n-p) \times n$ matrix having rows orthogonal to those of $\boldsymbol{A}$.

This result can be generalized to a full HMM model, with tied parameters in the style of Kumar, by considering the maximum likelihood objective function:

$$\mathcal{F}^{\mathrm{MLK}} = \ln P(\boldsymbol{Y}, \omega_r), \tag{7}$$

where $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots\}$ is the sequence of transformed feature vectors, and ω_r is the correct word label. The derivative of this model with respect to a model parameter θ_j is

$$\frac{\partial \mathcal{F}^{\mathrm{MLK}}}{\partial \theta_j} = \sum_t \sum_j \frac{\partial \mathcal{F}^{\mathrm{MLK}}}{\partial \ln p(\boldsymbol{y}_t|j)} \frac{\partial \ln p(\boldsymbol{y}_t|j)}{\partial \theta_j}, \tag{8}$$

$$= \sum_t \sum_j \gamma_t(j) \frac{\partial \ln p(\boldsymbol{y}_t|j)}{\partial \theta_j}, \tag{9}$$

where $p(\boldsymbol{y}_t|j)$ is the acoustic model state conditional probability. Setting these derivatives equal to 0 and solving for model parameters leads to the state-dependent means and variances as calculated in Eq. (4) with

$$\psi_t(j) = \gamma_t^{\mathrm{num}}(j), \tag{10}$$

for state j. This again leads to a solution to the LDA problem using the generalized eigenvalue problem (6), this time with soft class membership determined by the state posteriors. For models estimated using the Baum-Welch algorithm, the above LDA statistics more closely correspond to those used in estimating the model. This means that the matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ are more accurately estimated in this case.

3. Sequential Maximum Mutual Information LDA

3.1. Derivation from MMI objective function

Section 2 describes the linear transformation which maximizes scatter between classes and minimizes scatter within classes based on the correct labels as in Eq. (2). However, because the maximum likelihood statistics are different from the MMI statistics, the resulting $\boldsymbol{B}$ and $\boldsymbol{W}$ are not accurate for MMI-based models. Similar to the MMI discriminative training of acoustic model parameters, posteriors of denominator lattices γ_t^{den} should be taken into account. We call this method sequential MMI LDA (sLDA).

The MMI objective function is given as

$$\mathcal{F}^{\mathrm{MMI}} = \ln \frac{P(\boldsymbol{Y}, \omega_r)}{\sum_\omega P(\boldsymbol{Y}, \omega)}, \tag{11}$$

where ω are the hypotheses of the original system. The derivative of the MMI objective function (11) by state-dependent model parameters θ_j, as in MMI-SPLICE [9], is

$$\begin{aligned} \frac{\partial \mathcal{F}^{\mathrm{MMI}}}{\partial \theta_j} &= \sum_t \sum_j \frac{\partial \mathcal{F}^{\mathrm{MMI}}}{\partial \ln p(\boldsymbol{y}_t|j)} \frac{\partial \ln p(\boldsymbol{y}_t|j)}{\partial \theta_j}, \\ &= \sum_t \sum_j \Delta_t(j) \frac{\partial \ln p(\boldsymbol{y}_t|j)}{\partial \theta_j}, \end{aligned} \tag{12}$$

where $p(\boldsymbol{y}_t|j)$ is the acoustic model state conditional probability. This leads to the same mean and variance estimation as Eq. (4), except that here $\psi_t(j) = \Delta_t(l(t))$. However, since $\Delta_t(j)$ can be negative, usually extended Baum-Welch updates are used, because they maintain positivity. Here we introduce a parameter α $(0 \leq \alpha \leq 1)$ that reduces the strength of the denominator term $\gamma_t^{\mathrm{den}}(j)$:

$$\psi_t(j) = \gamma_t^{\mathrm{num}}(j) - \alpha \gamma_t^{\mathrm{den}}(j). \tag{13}$$

If α equals to zero, this equation reduces to that of LDA.

The proposed method can be interpreted as a form of LDA with a soft feature selection [12] corresponding closely to the MMI model. Little weight is imposed on the data where $\gamma_t^{\mathrm{den}}(j)$

is near one and this corresponds to the correct case for a recognizer. This realizes an adjustment of the weight of the training data according to the errors made by the recognizer. However, as the between-class variance $\boldsymbol{B}$ remains global, it is only slightly affected by the MMI-based weights, and this method still focuses on all classes. Nevertheless, it has a simple closed-form solution, and an easy implementation, so may be useful as a starting point for more advanced discriminative transforms.

3.2. I-smoothing interpretation

Equation (13) can be rewritten as

$$\psi_t(j) = (1-\alpha)\gamma_t^{\text{num}}(j) + \alpha\Delta_t(j). \tag{14}$$

This equation can be interpreted as a smoothing between the difference statistics $\Delta_t(j)$ and the class label posterior $\gamma_t^{\text{num}}(j)$ with interpolation ratio α. Thus, by setting the parameter α less than 1, α helps avoid over-training and is related to I-smoothing [5], which is widely used for discriminative training of acoustic models.

3.3. Boosted MMI extension

In analogy to boosted MMI [8], we can introduce a boosting factor b that boosts the posteriors of hypotheses based on the phoneme accuracy. The boosted MMI objective function is:

$$\mathcal{F}^{\text{bMMI}} = \ln \frac{P(\boldsymbol{Y}, \omega_r)}{\sum_\omega P(\boldsymbol{Y}, \omega) e^{-bH(\omega, \omega_r)}}, \tag{15}$$

where $H(\omega, \omega_r)$ is the phoneme accuracy. The boosted version of the weights can be obtained as in the classical boosted MMI framework, by using the forward-backward algorithm on the denominator lattice, and adding, for each state, $-b$ times the contribution to the sentence level accuracy. Denoting by $\gamma_t^{b,\text{den}}(j)$ the denominator term, we obtain a bMMI version of the weights $\psi_t^b(j)$:

$$\psi_t^b(j) = \gamma_t^{\text{num}}(j) - \alpha\gamma_t^{b,\text{den}}(j). \tag{16}$$

The boosting factor b is typically taken to be negative so as to put more focus on frames with low accuracy than on those with high accuracy.

4. Experiments

4.1. Experimental setup

We evaluated the performance improvement on two corpora: the Corpus of Spontaneous Japanese (CSJ) [13] and the second CHiME challenge Track 2 [14]. The former is one of the most widely used large vocabulary continuous speech recognition (LVCSR) tasks (vocabulary size is about 70k). Three types of test sets are provided and each set consists of 10 speakers' lecture-style speech. Test sets 1, 2, and 3 contain 22,682, 23,226, and 14,896 words, respectively. The first aim of our experiments is to validate the effectiveness of the proposed sLDA compared to the conventional LDA when changing the parameters α and b in Eq. (16). The HMM was trained with maximum likelihood estimation using 0th$\sim$12th order MFCCs + Δ + $\Delta\Delta$, the number of context-dependent HMM states was 3,500 and the total number of Gaussians was 96,000.

The CHiME challenge Track 2 is designed for evaluating the ASR performance of a medium vocabulary task (Wall Street Journal (WSJ0), vocabulary size is 5k) under reverberated and noisy environments with WER. The noise is non-stationary, such as other speakers' utterances, household noise, or music and is added to 'isolated' speech at SNR $= \{-6, -3, 0, 3, 6, 9\}$dB. This task is aimed to validate the performance of the proposed sLDA for noise robust speech recognition task, and the effectiveness of its combinations with discriminative training of acoustic models (Gaussian Mixture Model and Deep Neural Networks (DNN)) and feature space discriminative training (f-bMMI) [8]. We used noise-suppressed single-channel data obtained by prior-based binary masking [15] and used the Kaldi toolkit [16] with the baseline evaluation tool that we provided [15, 17]. The development set (si_dt_05) contained 409 utterances including 6,779 words from 10 speakers, and the evaluation data set (si_et_05) contained 330 utterances including 5,353 words from 12 speakers (Nov'92) for each SNR condition. The number of HMM states was 2,500 and the total number of Gaussians was 15,000. For the DNN, we used Povey's implementation of DNN training in Kaldi with 3 hidden layers and 1,000,000 parameters. The initial learning rate was 0.01 and was decreased to 0.001 at the end of training. The baseline features were 0th$\sim$12th order MFCCs + Δ + $\Delta\Delta$. Moreover, we combine LDA with Maximum Likelihood Linear Transformation (MLLT) [18, 19], which is usually performed with LDA as a set of feature transformation techniques. For the CHiME corpus, speaker adaptation technique, namely Speaker Adaptive Training (SAT) and feature space Maximum Likelihood Linear Regression (fMLLR), were also applied.

Table 1: WER of the conventional LDA ($\alpha = 0$) and the proposed sequential maximum mutual information LDA (sLDA) with different α and b, which are smoothing and boosting factors in Eq. (16), respectively, on CSJ database.

	α	b	test1	test2	test3	Avg.
LDA	0	0	20.42	17.95	19.22	19.20
	0.1	0	**20.39**	17.81	19.49	19.23
	0.3	0	20.47	17.93	19.28	19.23
	0.5	0	20.44	17.81	19.14	19.13
	0.7	0	20.40	17.83	19.03	19.09
	1.0	0	20.51	**17.68**	**18.77**	**18.99**
	0.1	-0.1	20.46	17.86	19.29	19.20
	0.3	-0.1	20.28	17.74	19.21	19.08
	0.5	-0.1	20.38	17.87	19.08	19.11
	0.7	-0.1	20.43	17.63	19.13	19.06
	1.0	-0.1	20.60	17.65	18.91	19.05
LDA	0	0	19.09	16.31	17.21	17.54
+MLLT	0.1	0	19.13	15.96	17.23	17.44
	0.3	0	19.08	**15.91**	17.07	**17.35**
	0.5	0	19.04	16.12	17.25	17.47
	0.7	0	19.09	16.03	17.11	17.41
	1.0	0	18.90	16.24	**16.94**	17.36
	0.1	-0.1	19.20	16.21	17.33	17.58
	0.3	-0.1	19.07	16.21	17.09	17.46
	0.5	-0.1	18.96	16.11	17.07	17.38
	0.7	-0.1	**18.87**	16.09	17.19	17.38
	1.0	-0.1	19.17	16.05	17.11	17.44

4.2. CSJ (LVCSR)

Table 1 shows the experimental results on the CSJ corpus. Although the performance improvement depended on the param-

Table 2: WER[%] for isolated speech (**si_dt_05**) of the CHiME challenge with different αs using ML acoustic model for noisy speech recognition with noise suppression by prior-based binary masking (sLDA+MLLT).

α	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg.
0	64.64	54.24	46.35	37.91	32.75	28.96	44.14
0.1	64.64	53.81	46.45	38.65	32.75	29.15	44.24
0.3	64.88	**53.72**	45.58	**37.13**	**31.89**	**28.43**	**43.61**
0.5	64.71	53.84	46.20	37.81	32.25	28.81	43.94
0.7	64.48	54.43	45.88	37.51	32.44	28.69	43.91
1.0	**64.36**	54.29	**45.01**	37.81	32.59	28.96	43.84

Table 3: WER[%] for isolated speech (**si_dt_05**) using ML and discriminatively trained acoustic model (bMMI) with feature-space discriminative training (f-bMMI). LDA+MLLT (upper), sLDA+MLLT (lower).

	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg.
ML	64.64	54.24	46.35	37.91	32.75	28.96	44.14
bMMI	63.39	52.54	44.56	35.60	30.98	28.10	42.53
f-bMMI	60.92	50.41	41.76	33.59	29.56	25.90	40.36
ML	64.88	53.72	45.58	37.13	31.89	28.43	43.61
bMMI	62.75	51.78	44.24	35.92	30.80	27.32	42.14
f-bMMI	**60.27**	**49.26**	**41.08**	**32.95**	**28.63**	**25.17**	**39.56**

eter α, overall, the proposed sLDA worked better than the conventional LDA ($\alpha = 0$) even when combined with MLLT. For the best case (bold case in the table), absolute 0.21% and 0.19% WER reductions for sLDA and sLDA+MLLT respectively were observed. Unfortunately, the boosted extension had little impact on the results, and for the rest of the experiments, the boosting factor b was set to zero.

4.3. Second CHiME Challenge Track 2 (Noise robust ASR)

Table 2 continues to investigates further the influence of the parameter α on performance through experiments on the CHiME challenge Track 2. MLLT is used in addition to the proposed sLDA. In average, for the cases where α is 0.3 or more, the speech recognition performance was improved and the case $\alpha = 0.3$ achieved the best improvement (0.53% absolute WER reduction), which is the same as in Table 1. From Tables 1 and 2, we validate that the proposed LDA was superior to the conventional LDA on two different ASR tasks.

Table 3 shows the results with discriminative training of acoustic model (bMMI) and feature space discriminative training (f-bMMI). For both cases, the proposed method improved the speech recognition performance, especially for the f-bMMI case (0.8% absolute WER reduction). The combination of the proposed method and f-bMMI achieved an additional improvement. This suggests that preliminary discriminative classification of the proposed method provided a good initialization to f-bMMI, which is also discriminative feature transformation with more precise region-dependent modeling.

Tables 4 and 5 show the results on the development and evaluation sets additionally with speaker adaptive training, fMLLR type speaker adaptation, and DNN system in order to validate the effectiveness of the proposed method in a state-of-the-art ASR system. Although for the DNN system the average ASR performance degraded on the evaluation set, the proposed method improved the performance for all the SNR conditions in the development set, and for half of the SNR conditions (-3, 3, and 9dB) in the evaluation set. Overall, the proposed method improved the average ASR performance by up to 0.9% absolute.

Table 4: WER[%] for isolated speech (**si_dt_05**) with speaker adaptive training, speaker adaptation (fMLLR), and minimum Bayes risk decoding (MBR). LDA+MLLT (upper), sLDA+MLLT (lower).

	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg.
ML	59.94	47.93	39.83	33.01	28.00	23.47	38.70
bMMI	56.90	45.79	37.60	30.31	26.15	21.74	36.42
f-bMMI	52.93	42.62	34.59	27.63	24.27	20.24	33.71
(+MBR)	52.65	42.04	33.75	27.05	23.74	19.91	33.19
DNN	52.78	42.50	34.08	27.05	24.13	20.12	33.44
bMMI	47.34	36.33	28.96	23.40	20.03	17.05	28.85
(+MBR)	46.79	35.68	28.44	22.88	19.91	16.64	28.39
ML	59.21	48.40	39.28	32.41	27.72	22.86	38.31
bMMI	56.14	45.51	36.69	29.55	26.08	21.33	35.88
f-bMMI	53.09	43.34	33.71	27.16	23.93	19.78	33.50
(+MBR)	52.60	42.51	33.03	26.38	23.34	19.18	32.84
DNN	52.91	41.81	32.56	27.73	24.31	19.68	33.17
bMMI	47.31	36.13	28.49	23.50	20.00	16.57	28.67
(+MBR)	**46.59**	**35.31**	**27.84**	**22.82**	**19.69**	**16.49**	**28.12**

Table 5: WER[%] for isolated speech (**si_et_05**) with speaker adaptive training and speaker adaptation (fMLLR). LDA+MLLT (upper), sLDA+MLLT (lower). In this table, DNN is DNN with boosted MMI.

	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg.
ML	50.91	41.64	33.89	26.30	21.61	18.85	32.20
f-bMMI	44.54	35.91	29.24	22.31	17.77	15.88	27.61
(+MBR)	44.51	35.42	28.81	21.46	17.41	14.98	27.10
DNN	37.98	28.26	21.86	17.71	12.61	11.75	21.70
(+MBR)	**37.14**	27.35	**21.41**	16.94	**12.55**	11.54	**21.16**
ML	50.46	42.05	32.80	26.42	21.22	18.61	31.93
f-bMMI	44.85	35.05	27.69	21.43	17.34	14.74	26.85
(+MBR)	44.07	34.09	27.22	20.33	16.85	14.61	26.20
DNN	38.63	27.54	22.55	17.37	13.23	11.69	21.84
(+MBR)	37.98	**27.16**	21.73	**16.93**	12.83	**11.23**	21.31

5. Conclusion and Future Work

This paper proposed to extend LDA based on sequential MMI training methods by using the discriminatively modified sufficient statistics computed from the lattices. The advantages of the proposed method are its low complexity and ease of implementation, in that it boils down to a simple modification of the computation of the sufficient statistics. Experiments on both an LVCSR task and a noise robust ASR task show its effectiveness. Although our approach is based on the closed-form solution of a generalized eigenvalue problem and is in that regard different from other discriminative feature transformation methods based on EBW or gradient-descent optimization techniques, future work will consider in more depth the theoretical relationships between them.

6. References

[1] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proceedings of ICASSP*, 1992, pp. 13–16.

[2] N. Kumar, *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. PhD dissertation, Johns Hopkins University, 1997.

[3] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, pp. 283–297, 1998.

[4] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proceedings of ICASSP*, vol. 11, 1986, pp. 49–52.

[5] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings of ICASSP*, vol. I, 2002, pp. 105–108.

[6] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proceedings of INTERSPEECH*, 2006, pp. 1573–1576.

[7] B. Zhang and S. Matsoukas, "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," in *Proceedings of ICASSP*, 2005, pp. 925–928.

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

[9] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," in *Proceedings of INTERSPEECH*, 2005, pp. 989–992.

[10] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proceedings of ICASSP*, 2000, pp. 1129–1132.

[11] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, 1990.

[12] B. Chen, S.-H. Liu, and F.-H. Chu, "Training data selection for improving discriminative training of acoustic models," *Pattern Recognition Letters*, vol. 30, pp. 1228–1235, 2009.

[13] S. Furui, K. Maekawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," in *Proceedings of ASR*, 2000, pp. 244–248.

[14] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proceedings of ICASSP*, 2013, pp. 126–130.

[15] Y. Tachioka, S. Watanabe, J. Le Roux, and J. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," in *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments*, Jun. 2013, pp. 19–24.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011, pp. 1–4.

[17] Y. Tachioka, S. Watanabe, and J. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," in *Proceedings of ICASSP*, May 2013, pp. 6935–6939.

[18] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proceedings of ICASSP*, 1998, pp. 661–664.

[19] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272–281, 3 1999.