Using Smoothed Heteroscedastic Linear Discriminant Analysis in Large Vocabulary Continuous Speech Recognition System *

Martin Karafiát, Lukáš Burget, and Jan Černocký

Faculty of Information Technology, Brno University of Technology
{karafiat,burget,cernocky}@fit.vutbr.cz

Abstract. In the state-of-the-art speech recognition systems, Heteroscedastic Linear Discriminant Analysis (HLDA) is becoming popular technique allowing for feature decorrelation and dimensionality reduction. However, HLDA relies on statistics, which may not be reliably estimated when only limited amount of training data is available. Recently, Smoothed HLDA (SHLDA) was proposed as a robust modification of HLDA. Previously, SHLDA was successfully used for feature combination in small vocabulary recognition experiments [1]. In this work, we verify that SHLDA can be advantageously used also for Large Vocabulary Continuous Speech Recognition.

1 Introduction

The Heteroscedastic Linear Discriminant Analysis is getting more popular in the state-of-the-art recognition systems. However, its success is quite dependent on the correct estimation of the needed statistics. In case a limited amount of training data is available, it may be difficult to obtain good estimates. The estimation will be problematic especially for HLDA, where an estimate of covariance matrix is required for each individual class. To overcome this problem, we have proposed Smoothed Heteroscedastic Linear Discriminant Analysis (SHLDA) [1] which is advantageously combining the robust estimation of statistics in LDA and relaxing of some statistical assumptions (identical covariance matrix of all classes) in HLDA. So far, SHLDA has not been tested on a big LVCSR task (which was also the most important critical point we received at ICSLP'2004). This paper describes such tests. It is organized as follows: section 2 described the SHLDA with its roots in HLDA and LDA. Section 3 presents the training of SHLDA in the GMM/HMM training framework. The following section 4 gives an overview of LVCSR systems the SHLDA was tested in, experimental setups and results. Section 5 concludes the paper.

* This work was partially supported by EC project Augmented Multi-party Interaction (AMI), No. 506811 and Grant Agency of Czech Republic under project No. 102/05/0278. Thanks to Sheffield University for generating LVCSR lattices. We further thank Cambridge University Engineering Department who have the h5train03 CTS training set available as well as the right to use Gunnar Evermann's HDecode at the University of Sheffield.

2 Smoothed Heteroscedastic Linear Discriminant Analysis

2.1 HLDA

The Heteroscedastic Linear Discriminant Analysis [2] can be used to derive linear projection de-correlating feature vectors and performing the dimensionality reduction. For HLDA, each feature vector that is used to derive the transformation must be assigned to a class. When performing the dimensionality reduction, HLDA allows to preserve useful dimensions, in which feature vectors representing individual classes are best separated (Figure 2). HLDA allows to derive such projection that best de-correlates features associated with each particular class (maximum likelihood linear transformation for diagonal covariance modeling [2, 3]).

To perform de-correlation and dimensionality reduction, n-dimensional feature vectors are projected into first $p < n$ rows, $\mathbf{a}_{k=1...p}$, of $n \times n$ HLDA transformation matrix, $\mathbf{A}$. An efficient iterative algorithm [3] is used in our experiments to estimate matrix $\mathbf{A}$, where individual rows are periodically re-estimated using following formula:

$$\hat{\mathbf{a}}_k = \mathbf{c}_k \mathbf{G}^{(k)-1} \sqrt{\frac{T}{\mathbf{c}_k \mathbf{G}^{(k)-1} \mathbf{c}_k^T}} \tag{1}$$

where $\mathbf{c}_i$ is the i^{th} row vector of co-factor matrix $C = |\mathbf{A}|\mathbf{A}^{-1}$ for current estimate of $\mathbf{A}$ and

$$\mathbf{G}^{(k)} = \begin{cases} \sum_{j=1}^{J} \frac{\gamma_j}{\mathbf{a}_k \hat{\boldsymbol{\Sigma}}^{(j)} \mathbf{a}_k^T} \hat{\boldsymbol{\Sigma}}^{(j)} & k \leq p \\ \frac{T}{\mathbf{a}_k \hat{\boldsymbol{\Sigma}} \mathbf{a}_k^T} \hat{\boldsymbol{\Sigma}} & k > p \end{cases} \tag{2}$$

where $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}^{(j)}$ are estimates of global covariance matrix and covariance matrix of j^{th} class, γ_j is number of training feature vectors belonging to j^{th} class and T is the total number of training feature vectors.

In our experiments, the classes are defined by each Gaussian mixture component m of each state s. The selection, that feature vector $\mathbf{o}(t)$ belong to class j, is given by the value of occupation probability $\gamma_j(t)$. These occupation probabilities, are used to re-estimate transition probabilities, and mixture component weights according to standard Baum-Welsh algorithm. In this algorithm, occupation probabilities, $\gamma_j(t)$, and feature vectors $\mathbf{o}(t)$ are used to estimate n-dimensional mean vector, $\boldsymbol{\mu}_j$, and full covariance $n \times n$ matrix, $\boldsymbol{\Sigma}_j$, of each Gaussian mixture component, j, according to the following equations:

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{t=1}^{T} \gamma_j(t) \mathbf{o}(t)}{\gamma_j}, \tag{3}$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{t=1}^{T} \gamma_j(t) (\mathbf{o}(t) - \hat{\boldsymbol{\mu}}_j)(\mathbf{o}(t) - \boldsymbol{\mu}_j)^T}{\gamma_j}, \tag{4}$$

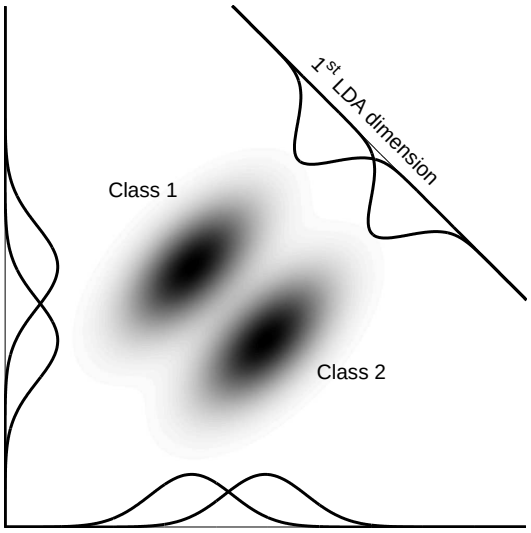

Fig. 1. *Linear Discriminant Analysis for 2-Dimensional Data.*

$$\gamma_j = \sum_{t=1}^{T} \gamma_j(t) \tag{5}$$

where T is the number of feature vectors used for training. New HLDA projection, $\mathbf{A}$, is then derived using the occupation probabilities and the estimated class covariance matrices, $\hat{\boldsymbol{\Sigma}}_j$.

To obtain the correct estimates of HMM parameters in feature space correspond to the newly derived transformation, $\mathbf{A}_p$, p-dimensional mean vector, $\hat{\boldsymbol{\mu}}_j^{HLDA}$, and variance vector, $\hat{\boldsymbol{\sigma}}_j^{HLDA}$, of each Gaussian mixture component is updated according to the following equations.

$$\hat{\boldsymbol{\mu}}_j^{HLDA} = \mathbf{A}_p \hat{\boldsymbol{\mu}}_j, \tag{6}$$

$$\hat{\boldsymbol{\sigma}}_j^{HLDA} = diag(\mathbf{A}_p \hat{\boldsymbol{\Sigma}}_j \mathbf{A}_p^T). \tag{7}$$

Where $\mathbf{A}_p$ is matrix consisting of first p rows of matrix $\mathbf{A}$.

2.2 LDA

Well known Linear Discriminant Analysis (LDA) can be seen as special case of HLDA, where it is assumed that covariance matrices of all classes are the same (see Figure 1). In contrast to HLDA, closed form solution exists in this case.

Base vectors of LDA transformations are given by eigen vectors of a $\boldsymbol{\Sigma}_{ac} \times \boldsymbol{\Sigma}_{wc}^{-1}$. The within-class covariance matrix, $\boldsymbol{\Sigma}_{wc}$, which represents the unwanted

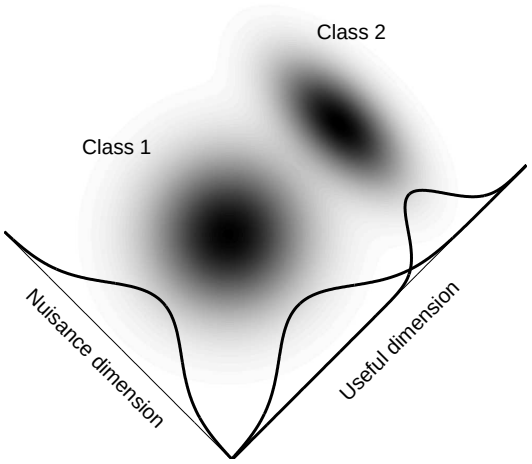

Fig. 2. *Heteroscedastic Linear Discriminant Analysis for 2-Dimensional Data.*

variability in data, is estimated as weighted average of covariance matrices of all classes:

$$\hat{\boldsymbol{\Sigma}}_{wc} = \frac{1}{N} \sum_j \gamma_j \hat{\boldsymbol{\Sigma}}_j, \tag{8}$$

The across-class covariance matrix $\boldsymbol{\Sigma}_{ac}$ represents the wanted variability in data and it is computed as a covariance matrix of weighted mean vectors of all classes:

$$\hat{\boldsymbol{\Sigma}}_{ac} = \frac{1}{T} \sum_j \gamma_j (\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}})^T = \hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{wc}, \tag{9}$$

where $\hat{\boldsymbol{\mu}}$ is the global mean vector and $\hat{\boldsymbol{\Sigma}}$ is the global covariance matrix.

Again, projection to only several eigen vectors corresponding to largest eigen values can be performed in order to reduce dimensionality of features.

2.3 Smoothed HLDA

HLDA requires the covariance matrix to be estimated for each class. The higher number of classes is used, the fewer feature vector examples are available for each class and class covariance matrix estimates become more noisy. LDA overcomes this problem by assuming that there is the same (within-class) covariance matrix for all classes. The within-class covariance matrix is computed as the weighted average of all class covariance matrices, which ensures its more robust estimate. On the other hand, assumption of the same covariance matrix for all classes is usually not fulfilled for real speech features, and therefore, transformation derived using LDA is not the optimal one.

We have recently proposed [1] a technique based on combination of HLDA and LDA, where class covariance matrices are estimated more robustly, and at the same time, (at least the major) differences between covariance matrices of different classes are preserved. We call it Smoothed HLDA (SHLDA). SHLDA differs from HLDA only in the way of class covariance matrices estimation. In the case of SHLDA, estimate of class covariance matrices is given by equation:

$$\check{\boldsymbol{\Sigma}}_j = \alpha\hat{\boldsymbol{\Sigma}}_j + (1-\alpha)\boldsymbol{\Sigma}_{WC} \tag{10}$$

where $\check{\boldsymbol{\Sigma}}_j$ is "smoothed" estimate of covariance matrix of class used by SHLDA and given by Gaussian mixture m at state s, $\hat{\boldsymbol{\Sigma}}_j$ is estimate of covariance matrix given by equation 4, $\boldsymbol{\Sigma}_{WC}$ is estimate of within-class covariance matrix given by equation 8 and α is smoothing factor, which is a value in the range of 0 to 1. Note that for α equal to 0, SHLDA becomes LDA and for α equal to 1, SHLDA becomes HLDA.

3 SHLDA/HLDA training

The algorithm of training SHLDA/HLDA feature transform is the following:

1. Well trained HMM models (no HLDA) are used to compute the occupation probabilities, posterior probability the feature vector belongs to class (Gaussian mixture). In our case, an original feature stream (PLP + Δ + $\Delta\Delta$) was augmented with $\Delta\Delta\Delta$ coefficients. Therefore, for the features read by HMM, a input transform (rows of identity matrix cutting-off $\Delta\Delta\Delta$) was applied to produce the original feature vectors to avoid the mismatch between models and data. The computation of full covariance statistics were performed for whole feature vectors.
2. Computation of SHLDA projection. Covariance matrices of each class are "smoothed" by global within-class covariance one (see equation 10) and the standard HLDA estimation follows. Finally, the dimensionality reduction was applied to limit the feature space into the original size.
3. HMM models are updated in the new feature space.
4. Additional iterations of standard Baum-Welsh re-estimations are performed to update mean and variances in projected space (without updating the transform).

4 Systems, experiments and results

4.1 Baseline system

Our recognition system was trained on ctstrain04 training set, a subset of the h5train03 set, defined at the Cambridge University as a training set for Conversation Telephone Speech (CTS) recognition systems. It contains about 278 hours of well transcribed speech data from Switchboard I,II and Call Home English (see Table 1).

Database	Amount of data [hours]	
	h5train03	ctstrain04
Switchboard I	263.61	248.52
Switchboard II - cellular	16.18	15.27
Call Home English	15.77	13.93
Total	295.56	277.72

Table 1. Train data description.

Corpus	Num. of words	LM weights	
		2gram	3gram
SWB	3.5M	0.733	0.639
Hub4 LM96	220M	0.266	0.360

Table 2. Number of words and weights per corpus for LVCSR language model.

13th order PLP cepstral coefficients, including 0th one, were extracted and first and second derivatives were added. This gives a standard 39 dimension feature vector. Cepstral mean and variance normalization were applied.

The baseline cross-word triphone HMM models were trained by Baum-Welsh re-estimation and mixture splitting. We used a standard 3-state left-to-right phoneme setup, with 16 Gaussian mixtures per state. 7598 tied states were obtained by decision tree clustering. Each Gaussian mixture was taken as different class for HLDA experiment. Therefore, we had $N = 16 \times 7598 = 121568$ classes.

The Language model used in decoding setup was computed by interpolation form Switchboard I,II + Call Home English and Hub4 (Broadcast news) transcriptions (see Table 2). The size of recognition vocabulary was 50k words.

All models selections were tested on the Hub5 Eval01 test set composed of 3 subsets of 20 conversations from the Switchboard-1, Switchboard-2, and Switchboard-cellular corpora, for a total length of about 6 hours of audio data.

The recognition output was generated in two passes:

- At first, lattice generation with baseline HMMs and bigram language model was performed. Further, the lattices were expanded by more accurate trigram language model. The pruning process was applied to reduce them to reasonable size.
- In the second pass, lattice were re-scored with our models.

To obtain the baseline number, the recognition results obtained with models used for lattice generation were simply re-scored by the same models. The baseline WER was **36.7%** WER.

4.2 Basic SHLDA system

As input for SHLDA experiments, the baseline HMM models (feature vectors with 39 dimensions) were used. We added the third derivatives into the feature

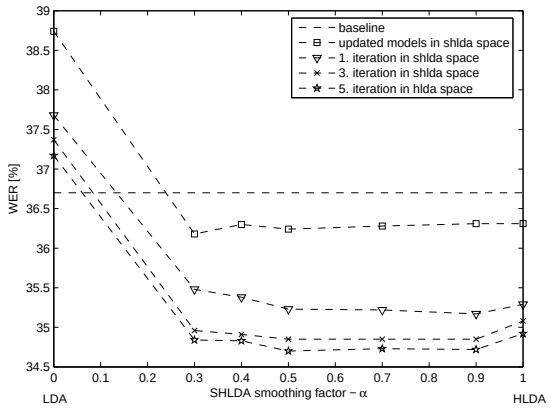

Fig. 3. Dependency of WER on the smoothing factor.

stream, which gave us 52 dimensional feature vector. To avoid the mismatch between the HMM models and feature vector size, we created the input transform, which truncates these last 13 coefficients.

SHLDA transform was then trained to perform the projection from 52 dimensions to 39 dimension. Smoothing factor values 0.0 (LDA), 0.3, 0.4, 0.5, 0.7, 0.9, 1.0 (HLDA) were tested in our experiments. Figure 3 shows dependency of WER on SHLDA smoothing factor α. Pure LDA failed, probably due to bad assumption of the same Gaussian distribution in all classes. The best system performance was obtained for smoothing factor 0.5 as the best balance between LDA and HLDA. The relative improvement of this system is 5% compared to the baseline and 0.5% compared to the clean HLDA setup. For the experiments with full system setup (section 4.3), the smoothing factor fixed to 0.5.

4.3 Full system setup

To improve the system accuracy we implemented VTLN speaker normalization and MLLR speaker adaptation:

1. **VTLN speaker normalization** - Warping factors for each speaker were estimated by Brent method in range 0.8 - 1.2 in similar way as in [4]. The training part of the database was recoded and new HMMs re-estimated. The SHLDA was applied as in the previous experiments (section 3 on VTLN'ed data. Smoothing factor was set to 0.5 (section 4.2).
   The ASR output from the best non-VTLN SHLDA system and SHLDA-VTLN models were used to estimate the new warping factors for eval01 test set. The data was recoded and new ASR output generated.
2. **MLLR speaker adaptation** - Output from previous decoding pass with VTLN-SHLDA models was used to estimate MLLR transformation per

Decoding pass	No-SHLDA	SHLDA(0.5)
Baseline	36.7	34.8
VTLN	33.7	32.3
VTLN+MLLR	31.5	30.6

Table 3. Comparison of systems without and with SHLDA.

speaker [5, 6]. Two transforms per speaker were applied - one for silence and one for speech.

WER **30.6%** was obtained with SHLDA-VTLN-MLLR system, which corresponds to 16% relative improvement with respect to baseline value. For comparison, we run the same adaptation techniques on baseline models without SHLDA, the WER was 31.5%WER. Therefore, 2.8% relative improvement was reached compared to the same system without SHLDA transform.

5 Conclusion

After encouraging results obtained on small task [1], we have shown that robust estimation given by smoothing of HLDA outperforms standard common HLDA method also in large LVCSR system. It is encouraging, that SHLDA helps even if case enough data is available (even if we can argue that is there not anything like enough data...). We find also good that SHLDA integrates well with standard speaker adaptation techniques.

References

1. L. Burget, "Combination of speech features using smoothed heteroscedastic linear discriminant analysis," in *8th International Conference on Spoken Language Processing*, (Jeju island, KR), oct 2004.
2. N. Kumar, *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. PhD thesis, John Hopkins University, Baltimore, 1997.
3. M. Gales., "Semi-tied covariance matrices for hidden markov models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.
4. T. Hain, P.Woodland, T. Niesler, and E. Whittaker, "The 1998 htk system for transcription of conversational telephone speech," in *Proc. IEEE ICASSP*, 1999.
5. C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hmms," in *Computer, Speech and Language, , Vol. 9, pp. 171 186*, 1995.
6. M. Gales and P. Woodland, "Mean and variance adaptation within the mllr framework," in *Computer Speech and Language, Vol. 10, pp. 249 264*, 1996.