Robust heteroscedastic linear discriminant analysis and LCRC posterior features in large vocabulary continuous speech recognition

Martin Karafiát, František Grézl, Petr Schwarz, Lukáš Burget, and Jan Černocký

Speech@FIT group, Faculty of Information Technology, Brno University of Technology
{karafiat,grezl,schwarzp,burget,cernocky}@fit.vutbr.cz

Abstract

This paper deals with feature extraction in speech recognition. Three robust variants of popular HLDA transform are investigated. Influence of adding posterior features to PLP feature stream is studied. The experimental results are obtained on CTS (continuous telephone speech) data. Silence-reduced HLDA and LCRC phoneme-state posterior features together provide more than 4% absolute improvement in word error rate.

Robustna heteroroskedastična linearna diskriminantna analiza (HLDA) in LCRC posteriorne značilke pri razpoznavanju tekočega govora z velikim besednjakom

Prispevek se ukvarja z izločanjem značilk v razpoznavanju govora. Raziskane so tri robustne različice priljubljene transformacije HLDA. Obravnavan je vpliv dodajanja posteriornih znailk zaporedju značilk PLP. Experimentalni rezultati so dobljeni na podlagi podatkov zveznega telefonskega govora. HLDA in LCRC posteriorne znailke stanja fonema skupaj prinašata več kot 4% absolutno izboljšanje pri stopnji zanesljivosti razpoznavanja besed.

1. Introduction

Speech feature extraction is important part of every large vocabulary continuous speech recognition system (LVCSR). Performance gains obtained thanks to this block are quite welcome as (on contrary to adding data or changing training or decoding algorithms), feature extraction is considered as "cheap" part of speech recognition system.

One of key problems in feature extraction is to reduce the dimensionality of feature vectors while preserving the discriminative power of features. Linear transforms such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are mostly used for this task. In recent years, Heteroscedastic Linear Discriminant Analysis (HLDA) has gained popularity in the research community (Kumar, 1997; Burget, 2004) for its relaxed constraints on statistical properties of classes (unlike LDA, HLDA does not assume the same covariance matrix for all classes). To compute HLDA transformation matrix, however, more statistics need to be estimated and the reliability of such estimations becomes an issue. Section 2. discusses robust variants of HLDA.

Second part of the paper is devoted to the use of posterior-features. Posteriors generated by neural networks (NN) and converted into features are also increasingly popular in small (Adami et al., 2002) and large (Zhu et al., 2005) recognition systems for their complementarity with classical PLP or MFCC coefficients. Section 3. introduces phoneme-state posterior estimator based on split temporal context (Schwarz et al., 2004; Schwarz et al., 2006) that has already proved its quality in different tasks ranging from language identification to keyword spotting.

2. HLDA

HLDA allows to derive such projection that best decorrelates features associated with each particular class (maximum likelihood linear transformation for diagonal covariance modeling (Kumar, 1997)). To perform decorrelation and dimensionality reduction, n-dimensional feature vectors are projected into first $p < n$ rows, $\mathbf{a}_{k=1\ldots p}$, of $n \times n$ HLDA transformation matrix, $\mathbf{A}$. An efficient iterative algorithm (Gales., 1999; Burget, 2004) is used in our experiments to estimate matrix $\mathbf{A}$, where individual rows are periodically re-estimated using the 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

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

where $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Sigma}}^j$ are estimates of the 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 and γ_j are their occupation counts.

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. In contrast to HLDA, closed form solution exists in this case. Basis of LDA transformation are given by eigen-vectors of matrix $\mathbf{\Sigma}_{AC} \times \mathbf{\Sigma}_{WC}^{-1}$, where $\mathbf{\Sigma}_{WC}$ is within-class covariance matrix and $\mathbf{\Sigma}_{AC}$ is across-class covariance matrix.

2.1. SHLDA

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 — class covariance matrix estimates become more noisy. We have recently proposed (Burget, 2004) 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. Smoothed HLDA (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:

$$\check{\mathbf{\Sigma}}_j = \alpha\hat{\mathbf{\Sigma}}_j + (1-\alpha)\mathbf{\Sigma}_{WC} \tag{3}$$

where $\check{\mathbf{\Sigma}}_j$ is "smoothed" estimate of covariance matrix for class j. $\hat{\mathbf{\Sigma}}_j$ is estimate of covariance matrix, $\mathbf{\Sigma}_{WC}$ is estimate of within-class covariance matrix and α is smoothing factor — 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.

2.2. MAP-SHLDA

SHLDA gives more robust estimation than standard HLDA but optimal smoothing factor α depends on the amount of data for each class. In extreme case, α should be set to 0 (HLDA) if infinite amount of training data is available. With decreasing amount of data, optimal α value will slide up to LDA direction.

To add more robustness into the smoothing procedure, we implemented maximum a posteriori (MAP) smoothing (Gauvain and Lee, 1994), where within-class covariance matrix $\mathbf{\Sigma}_{WC}$ is considered as the prior. Estimate of the class covariance matrix is then given by:

$$\check{\mathbf{\Sigma}}_j = \mathbf{\Sigma}_{WC}\frac{\tau}{\gamma_j+\tau} + \hat{\mathbf{\Sigma}}_j\frac{\gamma_j}{\gamma_j+\tau} \tag{4}$$

where τ is a control constant. Obviously, if insufficient data is available for current class, the prior $\mathbf{\Sigma}_{WC}$ is considered more reliable than the class estimation $\hat{\mathbf{\Sigma}}_j$. In case of infinite data, only the class estimation of covariance matrix $\hat{\mathbf{\Sigma}}_j$ is used for further processing.

2.3. Silence Reduction in HLDA

From the point of view of transformation estimation, silence is a "bad" class as its distributions differ significantly from all speech classes. Moreover, training data (even if end-pointed) contains significant proportion of silence. Therefore, we have experimented with limiting the influence of silence.

Rather than discarding the silence frames, the occupation counts γ_j of silence classes, which take part in computation of global covariance matrix $\hat{\mathbf{\Sigma}}$, and in Equation 2 are scaled by silence reduction factor $1/SR$. Setting $SR = \infty$ corresponds to complete elimination of silence statistics.

3. Posterior features

Several works have shown that using posterior-features generated by NNs is advantageous for speech recognition (Adami et al., 2002; Zhu et al., 2005). We have experimented with two setups to generate posteriors. The first one is based on a simple estimation of phoneme posterior probabilities from a block of 9 consecutive PLP-feature vectors (FeatureNet).

The second one uses our state-of-the-art phoneme-state posterior estimator based on modeling long temporal context (Schwarz et al., 2006). Details of the posterior estimator are shown in Fig. 1. Mel filter bank log energies are obtained in conventional way. Based on our previous work in phoneme recognition (Schwarz et al., 2004), the context of 31 frames (310 ms) around the current frame is taken. This context is split into 2 halves: Left and Right Contexts (hence the name "LCRC"). This allows for more precise modeling of the whole trajectory while limiting the size of the model (number of weights in the NN) and reducing the amount of necessary training data. For both parts, temporal evolutions of critical band log energies are processed by discrete cosine transform to de-correlate and reduce dimensionality. Two NNs are trained to produce phoneme-state posterior probabilities for both context parts. We use 3 states per phoneme which follows similar idea as states in phoneme HMM. Third NN functions as a merger and produces final set of phoneme-state posterior probabilities[1]

For both approaches, the resulting posteriors are processed by log and by a linear transform to de-correlate and reduce dimensionality (details are given in the experimental section below).

4. Experiments

Our recognition system was trained on ctstrain04 training set, a subset of the h5train03 set, defined at the Cambridge University as training set for Conversation Telephone Speech (CTS) recognition systems (Hain et al., 2005). It contains about 278 hours of well transcribed speech data from Switchboard I, II and Call Home English. All systems were tested on the Hub5 Eval01 test set composed of 3 subsets of 20 conversations from Switchboard I, II, and Switchboard-cellular, for a total length of about 6 hours of audio data.

The baseline features are 13th order PLP cepstral coefficients, including 0th one, with first and second derivatives added. This gives a standard 39 dimension feature vector. Cepstral mean and variance normalization was applied. Baseline cross-word triphone HMM models were trained by Baum-Welch re-estimation and mixture splitting. We used a standard 3-state left-to-right phoneme setup, with 16 Gaussian mixture components per state. 7598 tied states were obtained by decision tree clustering. Each Gaussian mixture was taken as a different class for HLDA experiment. Therefore, we had N = 16×7598 = 121568 classes.

The trigram language model used in decoding was estimated at University of Sheffield by interpolation from Switchboard I, II, Call Home English and Hub4 (Broadcast news) transcriptions. The size of recognition vocabulary was 50k words.

The recognition output was generated in two passes: At first, lattice generation with baseline HMMs and bigram language model was performed. The lattices were

[1] Neural nets are trained using QuickNet from ICSI and SNet – a parallel NN training software being developed in Speech@FIT

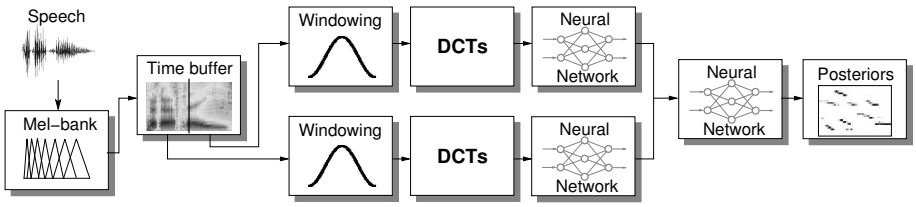

Figure 1: Phoneme-state posterior estimator based on split left and right contexts.

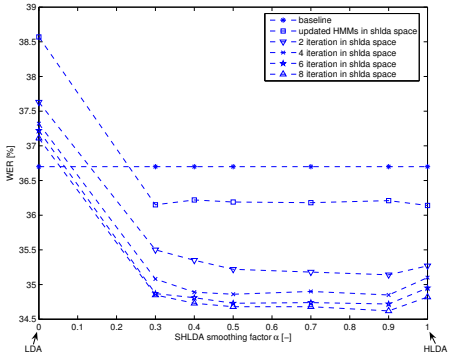

Figure 2: Dependency of WER on the SHLDA smoothing factors.

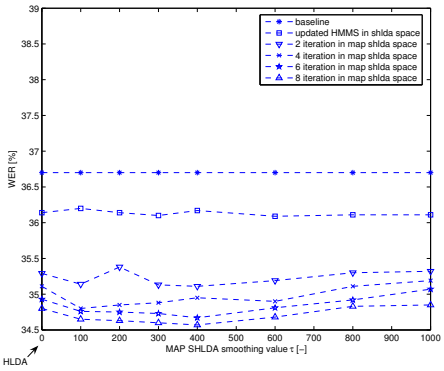

Figure 3: Dependency of WER on the MAP-SHLDA (right) smoothing factors.

expanded by more accurate trigram language model. The pruning process was applied to reduce them to reasonable size. In the second pass, lattices were re-scored with tested features and models.

4.1. Flavors of HLDA

We added the third derivatives into the feature stream, which gave us 52 dimensional feature vectors. **SHLDA** transform was then trained to perform the projection from 52 to 39 dimension. Smoothing factors α in Eq. 3 of 0.0 (LDA), 0.3, 0.4, 0.5, 0.7, 0.9, 1.0 (HLDA) were tested. Figure 2 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 (Table 1) was obtained for smoothing factor 0.9. The relative improvement of this system is 7.9% compared to the baseline and 0.6% compared to the clean HLDA setup.

MAP-SHLDA test setup was built in same way as SHLDA system, only the smoothing procedure (Equation 3) was replaced by MAP approach (Equation 4). The average value of all class occupation counts was 820. Therefore $\tau = 820$ in MAP-SHLDA should have the same behavior as $\alpha = 0.5$ in SHLDA if all classes had the same number of observations. The optimal smoothing values for SHLDA were in range 0.5—0.9 (Figure 2). Therefore, we decided to test smoothing control constant τ on values 0 (HLDA), 100, 200, 300, 400, 600, 800 and 1000. The results are shown in Figure 3. The best system performance (Table 1) was obtained for $\tau = 400$. The relative improvement of this system is 8% compared to the baseline and 0.7% compared to the clean HLDA setup.

Silence reduction in HLDA (SR-HLDA) was tested with factors SR equal to 1 (no reduction), 2, 10, 100 and ∞ (removing all silence classes). For $SR = 1$, the WER is obviously 34.8%, for $SR = 2$ it drops to 34.6% and from $SR = 10 \ldots \infty$ it is constant: 34.5%.

System	WER [%]
Baseline (no HLDA)	36.7
HLDA	34.8
SHLDA	34.6
MAP-SHLDA	34.6
SR-HLDA	34.5

Table 1: Comparison of HLDA systems.

4.2. Posterior features

Posterior features were always used together with base PLP features. Table 2 summarizes the results.

Upper part of Figure 4 shows the way the two feature streams were combined in FeatureNet experiments. The upper branch corresponds to the previous section. To compute posterior features, 9 frames of PLP+Δ+$\Delta\Delta$ were stacked and processed by a neural net with 1262 neurons in the hidden layer (this number was chosen to have approximately 500k weights in the NN). There are 45 phoneme classes, which determines the size of the output layer. Log-posteriors are processed by KLT or HLDA and then concatenated with PLP+HLDA features to form the final 64-dimensional feature vectors.

Lower panel of Figure 4 presents the setup with LCRC-posterior features. The PLPs were derived directly with Δ, $\Delta\Delta$ and $\Delta\Delta\Delta$, and down-scaled by HLDA to 39 dimensions. The detail of LCRC-posterior feature derivation is in Fig. 1, all nets had 1500 neurons in the hidden layer. For

System	WER [%]
PLP SR-HLDA	34.5
PLP SR-HLDA + PLP-posteriors KLT	33.8
PLP SR-HLDA + PLP-posteriors HLDA	33.3
PLP SR-HLDA + LCRC-posteriors HLDA	32.6

Table 2: Performance of posterior features in the CTS system.

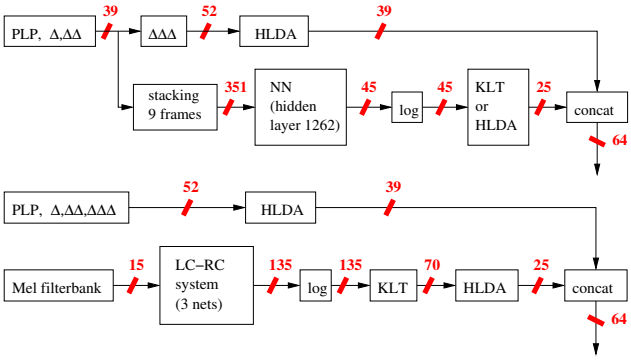

Figure 4: Configuration of the system with PLP- (upper panel) LCRC-posteriors (lower panel).

each frame, the output of LCRC system are estimates of 135 phoneme-state[2] posterior probabilities. As the number of phoneme-state posteriors is too high to fit the statistics necessary for HLDA estimation into the memory, the output dimensionality of LCRC system is first reduced by KLT from 135 to 70. The following HLDA reduces this size to 25, and the results are concatenated with PLP+HLDA features to form again 64-dimensional feature vectors.

We see, that the posterior features improve the results by almost 1% absolutely, and that there is clear preference of HLDA to KLT. With the new LCRC features, we have confirmed good results they provide in phoneme recognition (Schwarz et al., 2006) — with these features, the results are almost 2% better than the PLP SR-HLDA baseline.

5. Conclusion

In this paper, we have investigated robust variants of HLDA and use of classical and novel posterior features in speech recognition.

In the HLDA part, 2 approaches of HLDA smoothing were tested: Smoothed HLDA (SHLDA) and MAP variant of SHLDA, taking into account the amounts of data available for estimation of statistics for different classes. Both perform better than the basic HLDA. We have however found, that removing the silence class from the HLDA estimations (Silence-reduced HLDA) is equally effective and cheaper in computation. Testing SHLDA and MAP-SHLDA on the top of SR-HLDA did not bring any further improvement, therefore we stick with SR-HLDA as the most suitable transformation in our LVCSR experiments.

Two kinds of posterior features were tested – "classical" FeatureNet approach with stacked 9 frames of PLPs and novel approach using more elaborate structure to phoneme-state posterior modeling. The later scheme provided significant reduction of word error rate.

Our current work focuses on using the described feature extraction schemes in meeting data recognition along with speaker adaptative training scheme based on constrained maximum likelihood linear regression (CMLLR) and discriminative training using Minimum Phoneme Error (MPE) criterion. First results indicate that the improvement obtained by SHLDA and posterior features carries on through both adaptation and discriminative training steps.

[2]see (Schwarz et al., 2006) for details on splitting each of phonemes to 3 phoneme-states

6. Acknowledgments

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. Lukáš Burget was supported by post-doctoral grant of Grant Agency of Czech Republic No. 102/06/P383. Thanks University of Sheffield for generating LVCSR lattices. We further thank Cambridge University Engineering Department making the h5train03 CTS training set available for granting the right to use Gunnar Evermann's HDecode to the University of Sheffield.

7. References

A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas. 2002. Qualcomm-ICSI-OGI features for ASR. In *Proc. ICSLP 2002*, Denver, Colorado, USA.

L. Burget. 2004. Combination of speech features using smoothed heteroscedastic linear discriminant analysis. In *8th International Conference on Spoken Language Processing*, Jeju island, KR, oct.

M.J.F. Gales. 1999. Semi-tied covariance matrices for hidden markov models. *IEEE Trans. Speech and Audio Processing*, 7:272–281.

J. Gauvain and C. Lee. 1994. Maximum a posteriori estimation for multivariate gaussian mixture. *IEEE Trans. Speech and Audio Processing*, 2:291–298.

T. Hain, J. Dines, G. Gaurau, M. Karafiat, D. Moore, V. Wan, R.J.F. Ordelman, and S. Renals. 2005. Transcription of conference room meetings: an investigation. In *In Proceedings of Interspeech 2005*, Lisabon, Portugal.

N. Kumar. 1997. *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. Ph.D. thesis, John Hopkins University, Baltimore.

P. Schwarz, P. Matějka, and J. Černocký. 2004. Towards lower error rates in phoneme recognition. In *Proc. International Conference on Text, Speech and Dialogue*, pages 465–472, Brno, Czech Republic, September.

Petr Schwarz, Pavel Matějka, and Jan Černocký. 2006. Hierarchical structures of neural networks for phoneme recognition. In *Proc. ICASSP 2006*, Toulouse, France.

Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan. 2005. Using mlp features in sri's conversational speech recognition system. In *In Proceedings of Interspeech 2005*, pages 2141–2144, Lisabon, Portugal.