COMBINATION OF SPEECH FEATURES USING SMOOTHED HETEROSCEDASTIC LINEAR DISCRIMINANT ANALYSIS

Lukáš Burget

Brno University of Technology, Faculty of Information Technology
Božetěchova 2, Brno, 612 66, Czech Republic
burget@fit.vutbr.cz

ABSTRACT

Feature combination techniques based on PCA, LDA and HLDA are compared in experiments where limited amount of training data is available. Success with feature combination can be quite dependent on proper estimation of statistics required by the used technique. Insufficiency of training data is, therefore, an important problem, which has to be taken in to account in our experiments. Besides of some standard approaches increasing robustness of statistic estimation, methods based on combination of LDA and HLDA are proposed. An improved recognition performance obtained using these methods is demonstrated in experiments.

1. INTRODUCTION

For speech recognition, it is still not exactly known which information should be extracted from speech signal in the phase of feature extraction. Attempt to preserve one part of information often leads to loss of another (e.g. resolution in time vs. resolution in frequency). Currently, Mel Frequency Cepstral Coefficients (MFCC) are probably the most popular features. It was observed that recognition system using an alternative feature extraction method can perform better in certain cases (e.g recognition of certain words) even thought the overall performance of the system is worse. This is mainly caused by presence of complementary information in the different features.

It has been proved that combination of different systems can be powerful technique to improve recognition performance. The level of success is however limited by the complementarity of systems combined. In our previous work [1], we have developed technique allowing to measure the system complementarity. The technique was based on comparing symbol sequences recognized by different recognizers. We have shown, that outputs of systems that were recognized as the most complementary according to our measure can be successfully combined using technique known as ROVER [2]. All recognition systems used in those experiments differed only in a feature extraction method, which means that most of their complementarity is encoded directly in the features. Therefore, we have focused directly on combination of features in this work.

2. COMBINATION OF FEATURE STREAMS

The simplest way to perform the combination of feature streams that are time synchronized is the concatenation of corresponding feature vectors. The resulting feature stream may not be, however, suitable for following classification process, which usually requires feature vectors of reasonable dimensionality, and having individual coefficients decorrelated. The feature concatenation is, therefore, only the first step of combination techniques presented in this work. These techniques then differ in postprocessing performed in order to decorrelate concatenated feature vectors and to reduce their dimensionality by removing coefficients with redundant and unimportant information. The following sections deal with these postprocessing methods and related problems.

2.1. Postprocessing using PCA

Principal Component Analysis [3] (PCA) is a standard technique allowing for feature decorrelation and dimensionality reduction using linear projection. Success with PCA can be, however, quite limited for the following reasons: **a)** PCA assumes that features obey (Multivariate) Gaussian distribution. **b)** PCA projection ensures the only the global decorrelation of features. However, for classification process, it is usually desirable that features representing each particular class (e.g. one HMM state) are decorrelated. **c)** Dimensionality reduction is performed by discarding low variances dimensions. Correctness of this approach is therefore contingent on the assumption that variance in the data is directly related to the amount of information important for recognition of speech. This assumption will be often violated in our case, as coefficients coming from different feature streams can be scaled to different dynamic ranges. This problem was partially solved in our experiments by scaling all coefficients of concatenated feature vectors to the unity variance before deriving PCA projection.

Note that PCA projection matrix is given by eigen vectors of global covariance matrix, $\mathbf{\Sigma}$, which can be estimated on concatenated feature vectors available for training. To perform dimensionality reduction, concatenated feature vectors are projected only to several eigen vectors corresponding to largest eigen values.

2.2. Postprocessing using LDA and HLDA

As an alternative to PCA, Heteroscedastic Linear Discriminant Analysis (HLDA) [4] can be used to derive linear projection decorrelating concatenated 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 per-

This research has been partially supported by EC projects Multi-modal meeting manager (M4), No. IST-2001-34485, Augmented Multi-party Interaction (AMI), No. 506811 and by Grant Agency of Czech Republic under project No. 102/02/0124.

forming the dimensionality reduction, HLDA allows to preserve such dimensions, in which feature vectors representing individual classes are best separated. Because the importance of a dimension is given by separability of classes and not by variance of data in the dimension, there is no need to scale feature vector coefficients before deriving HLDA projection. Unlike PCA, HLDA allows to derive such projection that best decorrelates features associated with each particular class (maximum likelihood linear transformation for diagonal covariance modeling [4, 5]).

To perform decorrelation and dimensionality reduction, n-dimensional concatenated 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 [5] is used in our experiments to estimate matrix $\mathbf{A}$, where individual rows are periodically reestimated using following formula:

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

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

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

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

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 $\mathbf{\Sigma}_{AC}$ is across-class covariance matrix and $\mathbf{\Sigma}_{WC}$ is within-class covariance matrix. Again, projection to only several eigen vectors corresponding to largest eigen values can be performed in order to reduce dimensionality of features.

In our experiments, HMM state labels are generated for the training data by state-level forced alignment algorithm using a well-trained HMM system. The labels are then used to define 180 classes for HLDA and LDA.

3. ROBUST STATISTICS ESTIMATION

All the concatenated feature vector postprocessing techniques (PCA, LDA and HLDA) rely on statistics (e.g. global or class covariance matrices) estimated on training data. Success with the feature combination is, therefore, quite dependent on the correct estimation of these statistics. In our experiments, however, only limited amount of training data is available, which may not be sufficient to obtain their good estimates. The estimation will be problematic specially for HLDA, where an estimate of a covariance matrix is required for each individual class. To overcome this problem, following methods increasing robustness of statistics estimation are used in our experiments:

3.1. 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 propose 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. In our experiments, this technique will be refereed as 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:

$$\check{\mathbf{\Sigma}}^{(j)} = \alpha \hat{\mathbf{\Sigma}}^{(j)} + (1 - \alpha) \mathbf{\Sigma}_{WC} \qquad (3)$$

where $\check{\mathbf{\Sigma}}^{(j)}$ is "smoothed" estimate of covariance matrix of j^{th} class used by SHLDA, $\hat{\mathbf{\Sigma}}^{(j)}$ is estimate of ordinary covariance matrix of j^{th} class, $\mathbf{\Sigma}_{WC}$ is estimate of within-class covariance matrix 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.

3.2. Clustered HLDA

We propose also an alternative modification of HLDA increasing its robustness, to which we will refer as to Clustered HLDA (CHLDA). The modification is based on assumption that such clusters (sets of classes) can be found, that all classes belonging to one particular cluster have the same covariance matrix and differ only in mean vectors. Instead of *class covariance matrices* $\hat{\mathbf{\Sigma}}^{(j)}$ and *class occupation counts* N_j, which are the statistics used by ordinary HLDA, statistics used by CHLDA are *cluster within-class covariance matrices* $\hat{\mathbf{\Sigma}}_{CWC}^{C}$ and *cluster occupation counts* N^C. Estimate of *cluster within-class covariance matrix* for cluster C is given by equation:

$$\hat{\mathbf{\Sigma}}_{CWC}^{C} = \frac{\sum_{j \in C} N_j \hat{\mathbf{\Sigma}}^{(j)}}{N^C} \qquad (4)$$

Cluster occupation count is given as the sum of *class occupation counts* of all classes belonging to the cluster: $N^C = \sum_{j \in C} N_j$
In the case CHLDA, the issue is how to divide classes to clusters. For this purpose, a sophisticated clustering method can be used based, for example, on measuring similarities between feature distributions representing individual classes. In our experiment, a simple clustering is tested, where only two clusters are considered: classes (HMM states) representing non-speech parts of utterances and classes representing speech parts.

Note that, leaving each particular class to form a separate cluster, CHLDA becomes HLDA. And the other way around, making only one cluster consisting of all classes, CHLDA becomes LDA.

4. EXPERIMENTAL SETUP

Speech data from TI Connected Digits database [6] were used for both training and testing of all recognition systems. Limited

System	POW	DA4	30B	ENG	BLS	15B	LPCC	DA1	Avg.
WER [%]	2.90	2.91	2.99	3.00	3.04	3.14	3.36	3.51	3.11

Table 1. Word Error Rates for base features.

number of clean speech utterances were selected for training (616 utterances from 8 speakers). Four types of noise (subway, car, exhibition, babble) from AURORA2 TI Digits database [7] were artificially added to speech data at SNR level 20dB and 10dB. The same 616 utterances were used to create data for all noisy conditions. Together $616 \times (1 + 4 \times 2) = 5544$ utterances were used for training. Test data were prepared in a similar manner. Here, 912 utterances from 24 speakers were used. Together $912 \times (1 + 4 \times 2) = 8208$ utterances were used for testing.

To allow for feature combination, the following eight *base feature streams* were extracted from training and test data:

BSL - 15 Mel Frequency Cepstral Coefficients (MFCC) [8] augmented with their first and second order derivatives (delta and double-delta), filter bank applied on magnitude spectrum, 23 bands in Mel filter bank, 25 ms window length, 10 ms frame rate, 5 frames delta and delta-delta window, frame energy is represented by C0 coefficient

LPCC - 15 LPCC augmented with their derivatives (LPC order 15, other parameters similar to BSL features)

The name BSL stays for "baseline", since all seven remaining feature extraction methods are only modifications of BSL methods and always only one of their parameters is changed. In the following list, only the changed parameter of BSL features is described:

DA1	-	delta and delta-delta window is 3 frames
DA4	-	delta and delta-delta window is 9 frames
B15	-	15 bands are used in filter bank
B30	-	30 bands are used in filter bank
POW	-	filter bank is applied on power spectrum
ENG	-	frame energy replaces C0 coefficient

The feature combination is performed in two steps: Pairs of corresponding base feature vectors is concatenated to form $45 + 45 = 90$-dimensional features. These features are then decorrelated using a particular "combination" method (PCA, SHLDA, CHLDA) and its dimensionality is reduced again to 45.

The same recognition system is trained and evaluated for each (combined) feature stream. Continuous HMMs are used with output probability density function modeled by mixture of gaussians (3 components). Whole word model with left-to-right topology (16 states for digits, 3 states for silence) is used.

5. EXPERIMENTAL RESULTS

Word Error Rates (WER) of recognition systems using individual base features are presented in table 1. We can find 28 possible pairs of different base features. For each combination method (PCA, SHLDA, CHLDA), 28 systems using different pairs of base features are evaluated and their WERs are averaged. Average WERs for individual combination method can be found in the second column of table 2. In addition, combination methods are also tested for their ability only to decorrelate individual 45-dimensional base features (without performing any dimensional reduction). The eight possible systems using such features are evaluated for each combination method and their WERs are again averaged (see the first column of table 2). If such value is smaller then the average WER obtained with base features (3.11%), additional decorrelation using corresponding combination method is in average helpful.

System combination method	Average decorrelating system WER	Average combining system WER
PCA	3.70	3.35
LDA ($\alpha = 0$)	2.91	2.87
SHLDA $\alpha = 0.25$	3.08	2.82
SHLDA $\alpha = 0.5$	3.03	2.82
SHLDA $\alpha = 0.75$	3.04	**2.80**
HLDA ($\alpha = 1$)	3.14	2.91
CHLDA	2.96	**2.78**

Table 2. Average WERs for individual combination methods.

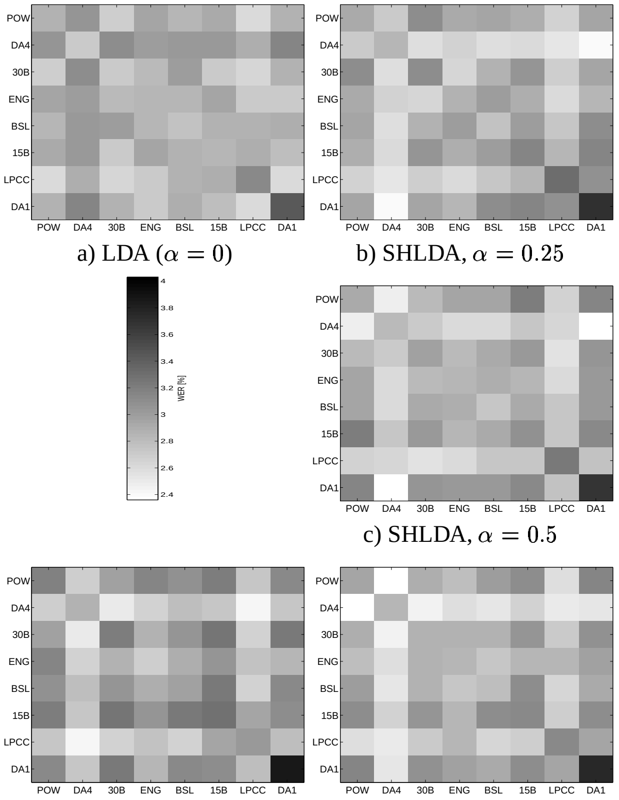
It can be observed from table 2 that **PCA** has clearly failed for both tasks: decorrelation of base features and their combination. In fact, with one exception, all systems using features combined by PCA performed worse then the better system using one of base features participating in the combination. PCA, which is driven only by variances seen in different feature space dimensions, was unable to distinguish important dimensions where variance is given mainly by speech from dimensions where variance is given by noise that present in our training data.

In our experiments with **SHLDA**, the following values were tested for the factor α: 0 (which equals to LDA), 0.25, 0.5, 0.75 and 1 (which equals to HLDA). The average WER for combined features is lowest with $\alpha = 0.75$ (see table 2). For this case, WERs of all 28+8 evaluated systems are also presented in table 3. Each value out-of-diagonal corresponds to one combination of pair of base features. Each diagonal value corresponds to the case where base features are only decorrelated by SHLDA. A graphical representation of table 3 is shown in figure 1d, where brighter color represents lower WER. Combination of DA4 or LPCC features with any other features is especially beneficial (bright rows/columns), which is perfectly in agreement with our previous work [1], where these features were found to be the most complementary to any other features.

LDA performed very well for the task of decorrelation of base features (see table 2). However, out-of-diagonal values in figure 1a do not indicate any consistent advantage of combining pairs of different features. Moreover, combination of DA4 features with any other features leads even to degradation in the performance (dark fields in row/column DA4 except the field on the diagonal), which is in contrast to our expectation that DA4 features should combine the best.

Results with **HLDA** are shown in figure 1e. Here, systems combining features other than DA4 or LPCC provide often worse results than corresponding systems based on LDA. In the average, LDA outperforms HLDA in both the ability to decorrelate base features and the ability to combine different features (see table 2). A probable explanation for this HLDA behavior is the following:

As HLDA requires to estimate statistics of much larger size in comparison to LDA, and as we have only limited amount of the data available for their estimation, the results obtained using HLDA are biased with an additional error caused by more noisy statistic estimates. However, the ability of HLDA to combine complementary information from different feature streams seems to be much better in comparison to LDA. If two highly complementary feature streams are combined using HLDA, the error bias is negli-



a) LDA ($\alpha = 0$) b) SHLDA, $\alpha = 0.25$

c) SHLDA, $\alpha = 0.5$

e) HLDA ($\alpha = 1$) d) SHLDA, $\alpha = 0.75$

Fig. 1. WER of systems using feature combination based on LDA, HLDA and SHLDA.

gible in comparison to the gain obtained by combination of complementary information. In this case HLDA is superior to LDA. In opposite, if two not much complementary feature streams are combined, the strength of HLDA is not employed and HLDA provides worse results than LDA because of the error bias. As we could see in figure 1d, SHLDA, which is making compromise between LDA and HLDA, decrease the error bias significantly in comparison to pure HLDA and at the same time it does not lose the ability to combine complementary information.

WERs obtained with **CHLDA** are presented in figure 2. Ability of CHLDA to decorrelate base features is virtually the same as in the case of LDA (see table 2), which was the method giving so far the best results for this purpose. In average, CHLDA outperforms all previously described systems in its ability to combine different features streams.

6. CONCLUSIONS

In this work, we have presented methods of feature combination. These methods differ only in postprocessing concatenated feature vectors, which is performed in order to decorrelate combined features and to reduce their dimensionality. A mixed results were obtained in experiments based on LDA, which is making wrong assumption of equality of class covariance matrices, and HLDA, where class covariance matrices were estimated poorly. The best results were obtained using SHLDA and CHLDA, which are making compromise between both the mentioned problems.

	POW	DA4	30B	ENG	BSL	15B	LPCC	DA1
POW	2.95	**2.36**	2.89	2.80	3.00	3.11	2.57	3.15
DA4		2.85	2.46	2.59	2.53	2.65	2.50	2.53
30B			2.86	2.87	2.87	3.06	2.72	3.08
ENG				2.85	2.73	2.85	2.85	2.97
BSL					2.80	3.10	2.63	2.92
15B						3.14	2.68	3.11
LPCC							3.13	2.95
DA1								3.76

Table 3. WER of systems using feature combination based on SHLDA for $\alpha = 0.75$.

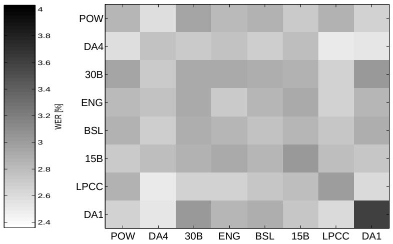


Fig. 2. WER of systems using CHLDA based feature combination. Two clusters are considered: HMM states representing non-speech parts of utterances and states representing speech parts.

7. REFERENCES

[1] L. Burget, "Measurement of complementarity of recognition systems," Tech. Rep., Brno University of Technology, Faculty of Information Technology, 2003.

[2] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. IEEE Workshop on automatic speech recognition and understanding*, 1997.

[3] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic press, Inc., Boston, USA, 1990.

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

[5] M.J.F. Gales., "Semi-tied covariance matrices for hidden markov models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[6] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP'84*, 1984.

[7] D. Pearce H. G. Hirsch, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Automatic Speech Recognition: Challenges for the Next Millennium*. ISCA ITRW ASR2000, Paris, France.

[8] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech & Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.