Comparison of Linear Discriminant Analysis Approaches in Automatic Speech Recognition

N. Jakovljevic[1], D. Miskovic[1], M. Janev[2], M. Secujski[1], V. Delic[1]
[1]*Department of Power, Electronic and Telecommunication Engineering, Faculty of Technical Sciences,*
Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia
[2]*Mathematical Institute, Serbian Academy of Sciences and Arts,*
Kneza Mihaila 36, 11000 Belgrade, Serbia
jakovnik@uns.ac.rs

***Abstract*—Speech recognition systems are commonly modelled by hidden Markov models with Gaussian mixture models as observation density functions. These models have a significant number of parameters, which usually leads to the problem of data sparsity, especially for under-resourced languages such as Serbian. One of the ways to overcome the problem of data sparsity is the reduction of the number of features. Linear discriminant analysis (LDA) and heteroscedastic LDA (HLDA) are two common ways to reduce the dimensionality in an automatic speech recognition task. The paper compares the properties of speech recognition systems for Serbian in which both techniques are applied with variable types of input features as well as the number of output features of (H)LDA. The best results are obtained in the case of HLDA with input vectors consisting of concatenations of feature vectors across 7 successive frames, where each feature vector contains 12 mel frequency cepstral coefficients (MFCCs) and normalized energy, and the number of output features is 32 or 35.**

***Index Terms*—Speech recognition, linear discriminant analysis.**

I. INTRODUCTION

One of the issues in automatic speech recognition as well as in pattern recognition is the dimensionality of the feature space. High dimensionality results in high computational and memory complexity as well as a loss of model generality. Although no subspace can contain more discriminative information than any larger space which includes it, since the training set does not contain the whole population the model captures only discriminative information on training data. The larger the deviation of the statistics of training samples from those of the whole population, the more severe this problem becomes. Dimensionality reduction techniques remove the dimensions which are unreliable for classification tasks [1].

Many methods have been proposed for dimensionality reduction, including principal component analysis (PCA) [2], linear discriminant analysis (LDA) [2], heteroscedastic LDA (HLDA) [3], multiple LDA (MLDA) [4], locality preserving projections [5] and marginal Fisher analysis [6]. In speech recognition the common approaches are LDA and HLDA [3], [7], [8]. A study which covers theoretical aspects of LDA and HLDA can be found in [9]. It evaluates the performance of these methods on synthetic Gaussian data under a variety of conditions such as the amount of training data and the degree of overlap between classes. It was observed that HLDA outperforms LDA if class-conditional distributions have unequal covariance matrices.

This paper presents a comparison of a range of LDA and HLDA methods proposed in literature, which differ in the type of input features and projected space dimensionality. All evaluations are made using software tools described in [10], on a Serbian telephone speech database, which includes various types of utterances such as digits, dates, proper names, commands and other common phrases [11].

The rest of the paper is organized as follows. Section II presents the mathematical foundations of LDA and HLDA methods. Experimental setup is described in Section III. Finally, Section IV details experimental results on a variety of input features and algorithm modifications, and is followed by conclusions given in Section V.

II. MATHEMATICAL FOUNDATIONS

The goal of LDA is to find a transformation (projection) matrix $\mathbf{\Theta}$ such that it maximizes the ratio of between-class scatter to the within-class scatter, defined by the following expression

$$J(\Theta) = |\mathbf{\Theta}^{\mathrm{T}}\mathbf{T}\mathbf{\Theta}| \big/ |\mathbf{\Theta}^{\mathrm{T}}\mathbf{W}\mathbf{\Theta}|, \quad (1)$$

where $\mathbf{T}$ is the total scatter matrix, $\mathbf{W}$ is the average within-class scatter matrix, and $|\cdot|$ denotes the determinant. Scatter matrices are defined by:

$$\mathbf{T} = \frac{1}{N}\sum_{i}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}, \quad (2)$$

$$\mathbf{W} = \frac{1}{N}\sum_{c}\sum_{i\in c}(\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}}, \quad (3)$$

where $\mathbf{x}_i$ is the i-th (d-dimensional) observation, $\boldsymbol{\mu}$ is the

Manuscript received January 29, 2013; accepted April 9, 2013.
This research was supported by Serbian Ministry of Education, Science and Technological Development grant No. TR32035.

mean value of all observations, $\boldsymbol{\mu}_c$ is the mean value of observations belonging to the class c, and N is the total number of observations.

The solution of the maximization of $J(\boldsymbol{\Theta})$ with the assumption that projected features ($\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{x}_i$) are uncorrelated is the matrix whose columns are the eigenvectors of the matrix $\mathbf{W}^{-1}\mathbf{T}$ [2]. In order to reduce the dimensionality, only p eigenvectors corresponding to the highest eigenvalues are used. Generally, since the matrix $\mathbf{W}^{-1}\mathbf{T}$ is not symmetric, its eigenvalues can be complex. This problem can be overcome by finding the eigenvectors $\mathbf{v}_k$ of the matrix $\mathbf{L}^{-1}\mathbf{T}(\mathbf{L}^{-1})^{\mathrm{T}}$, where $\mathbf{L}$ is the lower triangular matrix of $\mathbf{W}$ obtained by Cholesky decomposition ($\mathbf{W} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$), so the columns of matrix $\boldsymbol{\Theta}$ are $\boldsymbol{\theta}_k = (\mathbf{L}^{-1})^{\mathrm{T}}\mathbf{v}_k$.

In case of a Gaussian mixture model (GMM), the assumption about the lack of classification information for some features is equivalent to the assumption that the means and variances of the class distributions for these features are the same for all classes [3]. For notational convenience, the feature space of the means and variances is partitioned as follows:

$$\boldsymbol{\mu}_c = [\boldsymbol{\mu}_{cp}^{\mathrm{T}} \boldsymbol{\mu}_0^{\mathrm{T}}]^{\mathrm{T}}, \tag{4}$$

$$\boldsymbol{\Sigma}_c = \begin{bmatrix} \boldsymbol{\Sigma}_{cp} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_0 \end{bmatrix}, \tag{5}$$

where $\boldsymbol{\mu}_{cp}$ is the mean value of the p discriminative features of class c, $\boldsymbol{\mu}_0$ is the mean value of $d - p$ non-discriminative features, $\boldsymbol{\Sigma}_c$ is the $d \times d$ covariance matrix of class c, $\boldsymbol{\Sigma}_{cp}$ is the $p \times p$ covariance matrix of the discriminative features of class c, and $\boldsymbol{\Sigma}_0$ is the covariance matrix of non-discriminative features. It should be noted that $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are common for all classes. The goal of HLDA is to find the matrix $\boldsymbol{\Theta}$ which maximizes the likelihood function

$$L(\boldsymbol{\Theta}) = -\frac{1}{2}\sum_i \left(\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\mu}_{c(i)}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{c(i)}^{-1} \left(\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\mu}_{c(i)}\right) + -\frac{1}{2}\sum_i \ln\left(\left|\boldsymbol{\Sigma}_{c(i)}\right|\right) - \frac{Nd}{2}\ln(2\pi) + N\ln\left(\left|\boldsymbol{\Sigma}_{c(i)}\right|\right), \tag{6}$$

which can be simplified into

$$\boldsymbol{\Theta}_F = \arg\max_{\boldsymbol{\Theta}} \left\{ -\frac{N}{2}\ln\left|\boldsymbol{\Theta}_{n-p}^{\mathrm{T}}\mathbf{T}\boldsymbol{\Theta}_{n-p}\right| + -\sum_c \frac{N_c}{2}\ln\left|\boldsymbol{\Theta}_p^{\mathrm{T}}\mathbf{W}_c\boldsymbol{\Theta}_p\right| + N\ln\left|\boldsymbol{\Theta}\right| \right\}, \tag{7}$$

where N_c is the number of instances in class c, $\mathbf{W}_c$ is the scatter matrix for class c and $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_p\ \boldsymbol{\Theta}_{n-p}]$. However, since there is no closed form of a solution of (7), the steepest descent algorithm is needed for its computation [3]. If it is assumed that the projected features are non-correlated, transformed matrices are diagonal and the problem can be further simplified into

$$\boldsymbol{\Theta}_F = \arg\max_{\boldsymbol{\Theta}} \left\{ -\frac{N}{2}\ln\left|\mathrm{Diag}(\boldsymbol{\Theta}_{n-p}^{\mathrm{T}}\mathbf{T}\boldsymbol{\Theta}_{n-p})\right| + -\sum_c \frac{N_c}{2}\ln\left|\mathrm{Diag}(\boldsymbol{\Theta}_p^{\mathrm{T}}\mathbf{W}_c\boldsymbol{\Theta}_p)\right| + N\ln\left|\boldsymbol{\Theta}\right| \right\}. \tag{8}$$

For this variant of HLDA, a very efficient algorithm for the maximization of (8) is presented in [13]. It can be shown that when Gaussian distributions are constrained so as to have equal covariance matrices, the maximization of (8) is equivalent to the maximization of (1).

Since non-discriminative features are not used for GMM modelling, the assumption of correlation between non-discriminative features will not increase model complexity compared to the diagonal covariance model. This is the reason why in this paper an additional variation of HLDA is analysed. In this case the problem is

$$\boldsymbol{\Theta}_F = \arg\max_{\boldsymbol{\Theta}} \left\{ -\frac{N}{2}\ln\left|\boldsymbol{\Theta}_{n-p}^{\mathrm{T}}\mathbf{T}\boldsymbol{\Theta}_{n-p}\right| + -\sum_c \frac{N_c}{2}\ln\left|\mathrm{Diag}(\boldsymbol{\Theta}_p^{\mathrm{T}}\mathbf{W}_c\boldsymbol{\Theta}_p)\right| + N\ln\left|\boldsymbol{\Theta}\right| \right\}. \tag{9}$$

III. EXPERIMENTAL SETUP

All analysed systems are speaker-independent continuous speech recognizers based on hidden Markov models (HMMs) and GMMs. The number of states per model varies depending on the average model duration in the training set. The acoustic model consists of 4156 states obtained by tree-based clustering. The number of states as well as model structure are the same in all examined variants. The number of mixture components per state is set heuristically, i.e. the number of mixtures gradually rises until the average likelihood on a validation set starts to fall. As a result of this procedure the number of mixtures per state usually differs in the examined systems. For the estimation of system parameters the maximum likelihood criterion is used. More details about the applied procedure can be found in [10].

LDA and HLDA are methods which take into account the information about observation classes. The most common approach is to use triphone HMM states as (H)LDA classes, which was confirmed to outperform other approaches in [12]. The (H)LDA input vectors include standard 12 mel frequency cepstral coefficients (MFCCs) with normalized energy as well as their first and second derivatives [3], and vectors formed by concatenation of successive frames containing 12 MFCCs and normalized energy (the number of successive frames ranges from 3 to 11). The frames are 30 ms long and they are extracted every 10 ms.

The original LDA algorithm [2] calculates the columns of the transformation matrix as eigenvectors of matrix $\mathbf{W}^{-1}\mathbf{T}$. In the rest of the paper, such an LDA transformation matrix will be referred to as unnormalized. In the iterative algorithm for HLDA proposed in [3], the initial transformation matrix is normalized so as to set its determinant to 1. In order to compare the results of LDA and HLDA, experiments with the normalized variant of the transformation matrix are conducted as well. It should be noted that such a normalization does not change the LDA objective function.

The results presented in [3] suggest that there is no significant improvement of the system performance if LDA with GMMs with full covariance matrices are used, thus all the experiments in this paper are restricted to GMMs with diagonal approximations of covariance matrices. This is

equivalent to the assumption that features in the projected space are uncorrelated. To avoid any differences in the estimation of the transformation matrix based on (8) and (9) which would be due to the differences in the algorithm, the iterative algorithm proposed in [3] is used in both cases. In Tables III and V the performances of the systems which used original HLDA, defined by (8), are given in the column "All", and the performances of the relaxed variants, defined by (9), are given in the column "Discriminative".

IV. RESULTS

The word error rates (WER) of the referent systems are shown in Table I. The first system with 39 features (12 MFCCs, normalized energy and their first and second derivatives) is standard in a speech recognition task. In the further text it will be referred to as REF1. On the other side, the system with 26 features (12 MFCCs, normalized energy and their first derivatives) shows better performance on the Serbian telephone speech corpus (it will be referred to as REF2). This is probably a consequence of data sparsity in a 39-dimensional feature space, because the training corpus is relatively small, containing only 12 hours of speech [10].

In Table II, the performances of the systems based on LDA with concatenated input features for different numbers of input frames and output dimensionalities are shown. By comparing them with the WERs of the referent systems, it can be seen that the input vector should consist of at least 7 frames. This is surprising because for the calculation of the first derivative only 5 successive frames are used, which is sufficient for REF2. The high values of WER for the systems with 39 output features confirm the assumption of data sparsity in a 39-dimensional feature space. In many cases, the normalization of the LDA transformation matrix results in a lower WER compared to the unnormalized one, but the difference is not significant. The performances of the systems with the input vector consisting of 9 and 11 frames are similar. This means that the additional 2 frames do not contain new discriminative information.

Experiments with HLDA are conducted only for the 2 variants of input vectors which show the best performance in LDA tests, and their results are presented in Table III. The worst performance for both variants of input features is obtained for 39 output features, although in the version where input vectors contain 9 frames and HLDA assumes correlation of non-discriminative projected features WER is surprisingly low. These high WERs as well as their high deviation can be explained by data sparsity in the training set. The best performances are obtained when the output vector has 32 or 35 dimensions. In these cases the relaxation of the constriction related to the correlation in the projected space leads to a slight improvement of the performances, which unfortunately is not observed in all cases.

In Fig. 1 the values of WERs for the systems based on the LDA and corresponding HLDA are compared. In some cases the HLDA procedure results in a higher WER than the one obtained by the corresponding LDA procedure, which is unexpected since the initial value of transformation matrix for HLDA is the one obtained by LDA. This is probably due to poor estimation of within-class scatter matrices for high-

TABLE I. WORD ERROR RATES (WER) FOR THE REFERENT SYSTEMS

Features	WER[%]
12 MFCCs, normalized energy and 1st and 2nd derivatives	4.73
12 MFCCs, normalized energy and 1st derivatives	4.04

TABLE II. WORD ERROR RATES (WER) FOR THE SYSTEMS BASED ON LDA TRANSFORMED FEATURES WITH NORMALIZED AND UNNORMALIZED TRANSFORMATION MATRIX.

Number of successive frames	Dimension of reduced feature space	WER[%]	
		Normalized	Unnormalized
3	39	7.25	7.17
	35	6.14	6.24
	32	5.57	5.55
	26	5.72	5.11
5	39	5.32	5.12
	35	4.75	4.39
	32	4.32	4.22
	26	4.30	4.35
7	39	4.38	4.41
	35	3.91	3.76
	32	3.90	3.78
	26	3.93	4.14
9	39	4.45	4.62
	35	3.83	3.88
	32	3.83	4.10
	26	4.10	4.15
11	39	4.87	5.23
	35	3.80	3.97
	32	3.83	3.98
	26	4.20	4.25

Note: input LDA vectors are concatenated across successive frames, where each frame contains 12 MFCCs and normalized energy.

TABLE III. WORD ERROR RATES (WER) FOR THE SYSTEMS BASED ON HLDA TRANSFORMED FEATURES.

Number of successive frames	Number of discriminative features	WER[%]	
		Discriminative	All
7	39	4.88	4.93
	35	3.35	3.52
	32	3.25	3.72
	26	4.33	4.16
9	39	3.85	4.79
	35	3.63	4.09
	32	3.89	4.16
	26	4.28	4.41

Note: transformation matrices are estimated with assumptions that only discriminative features are non-correlated (column "Discriminative") and all transformed features are non-correlated (column "All"). Input LDA vectors are concatenated across successive frames, where each frame contains 12 MFCCs and normalized energy.

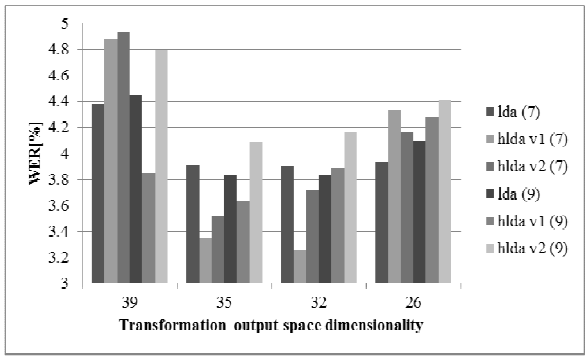

Fig. 1. Word error rates for the systems based on LDA and HLDA features obtained for different dimensionalities of the input and output vectors. The number of successive frames which compose the input (H)LDA vector are given in brackets, and "disc." and "all" indicate the HLDA variant ("disc": only discriminative features are uncorrelated, "all": all features in projected space are uncorrelated).

TABLE IV. WORD ERROR RATES (WER) FOR THE SYSTEMS BASED ON LDA TRANSFORMED FEATURES WITH NORMALIZED AND UNNORMALIZED TRANSFORMATION MATRIX.

Dimension of reduced feature space	WER[%]	
	Normalized	Unnormalized
39	4.03	4.26
35	3.81	3.95
32	4.04	3.63
26	4.11	4.22

Note: Input LDA vectors contain 12 MFCCs, normalized energy, and their first and second derivatives.

TABLE V. WORD ERROR RATES (WER) FOR THE SYSTEMS BASED ON HLDA TRANSFORMED FEATURES

Number of discriminative features	WER[%]	
	Discriminative	All
39	3.95	
35	3.67	3.39
32	3.41	3.63
26	3.85	3.95

Note: transformation matrices are estimated with assumptions that only discriminative features are non-correlated (column "Discriminative") and all transformed features are non-correlated (column "All"). Input LDA vectors contain 12 MFCCs, normalized energy, and their first and second derivatives. In the case of 39-dimensional output vector all features are discriminative.

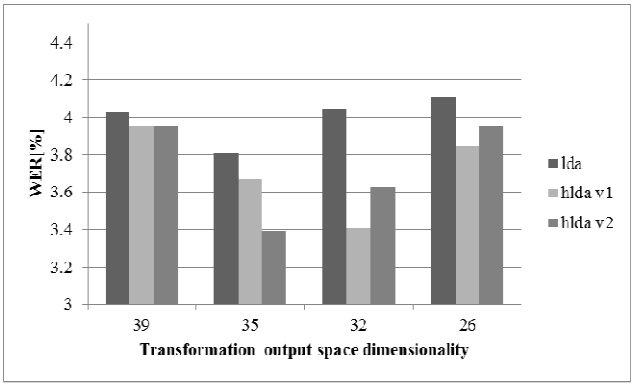

Fig. 2. Word error rates for the systems based on LDA and HLDA features obtained for different dimensionality of output vectors in case the input vector contains 12 MFCCs, normalized energy and their first and second derivatives. The labels "disc." and "all" indicate the HLDA variant ("disc": only discriminative features are uncorrelated, "all": all features in projected space are uncorrelated.

The performances of the systems based on LDA with the input vector containing 12 MFCCs and normalized energy and their derivatives for different output dimensions are shown in Table IV, and it can be seen that the WER for all systems is lower than for REF1. It is also interesting to note that the systems with 26 projected features show inferior performance in comparison with REF2, which also has 26 features. Since most other systems based on LDA have lower WERs than REF2, it can be concluded that projected features have more discriminative properties but that the first 26 with highest eigenvalues are not sufficient for discrimination. As in the case of input features concatenated over successive frames, the use of the normalized transformation matrix usually decreases the WER.

Unlike the systems based on HLDA with input features concatenated over successive frames, all systems based on HLDA with features which include derivatives show an improvement compared to the corresponding systems based on LDA (see Fig. 2 and Table V). On the other hand, there is no significant difference between systems based on HLDA with or without the assumption of correlation between non-discriminative projected features, and the best results are obtained when the dimensions of the output features are 32 and 35.

V. CONCLUSIONS

The paper presents the results of a comparison of a range of LDA and HLDA methods proposed in literature applied to speech recognition in Serbian. It has been found that in case of a relatively small training set, which is common for under-resourced languages, the best results are obtained if the input vectors are concatenated across 7 frames, with each of them containing 12 MFCCs and normalized energy. However, more consistent results are obtained in case where input features are the standard ones, because of poor estimation of within-class scatter matrices for high-dimensional input features. The lowest WER is obtained if 32 output features are used. The introduction of the assumption of correlation between non-discriminative projected features for HLDA can slightly improve system performance if the input features are high-dimensional.

REFERENCES

[1] X. Jiang, "Linear Subspace Learning-Based Dimensionality Reduction", *IEEE Signal Process. Mag.,* vol. 28, no. 2, pp. 16–26, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1109/MSP.2010.939041

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York NY: Springer, ch. 4, 2006.

[3] N. Kumar, A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rand HMMs for improved speech recognition", *Speech Communication*, vol. 26, pp. 283–297, 1998. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(98)00061-2

[4] M. Gales, "Maximum likelihood multiple subspace projections for hidden Markov models", *IEEE Trans. on Speech and Audio Process,* vol. 10, no. 2, pp. 37–47, Feb. 2002. [Online]. Available: http://dx.doi.org/10.1109/89.985541

[5] M. Belkin, P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods", *J. Comput. Syst. Sci.*, vol. 74, no. 3, pp. 328–340, Mar. 2005.

[6] S. Yan. D. Xu, B. Zhang, Q. Yang, H. Zhang, S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction", *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 29, no. 1, pp. 40–51, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2007.250598

[7] R. Haeb-Umbach, H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition", in *Proc. of Inter. Conf. Acoustic Speech and Signal Process.,* San Francisco, 1992, pp. 13–16.

[8] M. Westphal, "TC-Star Recognition baseline results", *TC-STAR Proj. Report*, 2004, [Online]. Available: http://www.tcstar.org/documents/deliverable/deliverable_updated14april05/D6.pdf

[9] H. Zhou, D. Karakos, S. Khundapur, A. Andreou, C. Priebe, "On projections of Gaussian distributions using maximum likelihood criteria", in *Proc. of Information Theory and Applications Workshop* 2009, pp. 431–438.

[10] V. Delić, M. Sečujski, N. Jakovljević, M. Janev, R. Obradović, D. Pekar, "Speech technologies for Serbian and kindred South Slavic languages", *Advances in Speech Recognition.*. Rijeka, Croatia: InTech, ch. 9, 2010. [Online]. Available: http://www.intechopen.com/books/advances-in-speech-recognition/speech-technologies-for-serbian-and-kindred-south-slavic-languages

[11] N. Đurić, D. Pekar, Lj. Jovanov, "Struktura srpske SpeechDat(E) govorne baze snimljene preko fiksne telefonske mreže", in *Proc. of Digit. obrada govora i slike*, Bečej, 2002, pp. 57–60.

[12] N. Jakovljević, D. Mišković, M. Janev, D. Knežević, T. Grbić, "Primena linearne diskriminativne analize u prepoznavanju govora", in *Proc. of Digit. obrada govora i slike,* Kovačica, 2012, pp. 40–43.

[13] M. Gales, "Semi-Tied Covariance Matrices for Hidden Markov Models", *IEEE Trans. Speech Audio Process.,* vol. 7, no. 3, pp. 272–279, May 1999. [Online]. Available: http://dx.doi.org/10.1109/89.759034