Development and Integration of the LDA-Toolkit into the COST249 SpeechDat (II) SIG Reference Recognizer

Bojan Kotnik, Zdravko Kačič, and Bogomir Horvat

University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova ulica 17, SI-2000 Maribor, Slovenia
bojan.kotnik@uni-mb.si

Abstract

This paper presents the development of Linear Discriminant Analysis toolkit (LDA-Toolkit) and its integration into widely used COST249 SpeechDat(II) Task Force Reference Recognizer (RefRec). The crucial parts of the LDA, the determination of LDA classes, as well as the influence of the level of dimensionality reduction on automatic speech recognition performance, are discussed. Evaluation of proposed LDA-RefRec procedure is performed using the Slovenian, German, and Spanish SpeechDat (II) databases. HTK (Hidden Markov Model Toolkit) is used in training and recognition processes. Features are computed using Advanced Front End (AFE) feature extraction procedure, proposed by Motorola, France Telecom, and Alcatel (AFE has been also standardized by ETSI organization). Automatic speech recognition results achieved with LDA-RefRec procedure show performance improvement and simultaneously dimensionality reduction when compared to baseline RefRec procedure. Proposed multilingual LDA classes, equal for all the three databases, perform only slightly worse than monolingual LDA classes, constructed and used separately for particular database. The results show benefits of the usage of the proposed LDA-RefRec procedure for evaluation or development of the automatic speech recognition systems based on SpeechDat (II) compliant databases.

1. Introduction

The structure of a typical automatic continuous-speech recognition system consists of a front-end speech parameterization block, followed by a statistical pattern classifier (usually based on Hidden Markov Models - HMMs). The interface between these two, the feature vector, should ideally contain all the linguistic information of the speech signal relevant to subsequent classification, be insensitive to irrelevant variations (e.g. due to changes in the acoustic environment, inter- and intra- speaker variations), and at the same time have low dimensionality in order to minimize the computational demands of the classification procedure (Welling, 1999). Principal Component Analysis (PCA) and Linear Discriminant Analysis are two common techniques applied in feature post-processing stage for classification and dimensionality reduction. PCA is a feature classification technique in which the data in the input space is transformed to a new feature space where the features are decorrelated. PCA extracts the dimensions along which the data vary most (the dimensions with highest co-variance). Therefore, it is possible to reduce the dimension of the data, if the dimensions with lowest variances are discarded. On the other hand, the optimization criterion for LDA attempts to maximize class separability, which not only reduces the dimensionality of the data, but also reduces the confusion error. The optimizing criterion to obtain the LDA transform, which is represented as a ratio of average between class variations over average within-class scatter, should therefore be maximized. Therefore, it is expected that with the application of the LDA better automatic speech recognition accuracy as well as lower computational requirements should be achieved (Schafföner et al., 2003). The paper describes the development of the LDA-Toolkit and its integration into widely used COST249 Reference Recognizer *RefRec* (Lindberg et al., 2000). First experiments with the RefRec scripts and SpeechDat (II) databases were performed using classical MFCC feature extraction procedure. Recently, in the Aurora standardization group, a novel high - performance feature extraction procedure – Advanced Front-End (AFE) – has been standardized (ETSI standard document, 2002). Therefore, in the proposed paper the AFE will be used to extract baseline speech feature vectors and to perform baseline experiments.

The rest of the paper is organized as follows. Section 2 presents a short overview of the RefRec automatic speech recognition system. Afterwards in the Section 3 the LDA-Toolkit will be described. The integration procedure of the LDA-Toolkit into the RefRec will also be given. Sections 4 describes the experimental setup and presents the results using the proposed LDA-RefRec training/testing procedure and Slovenian, German, and Spanish SpeechDat (II) databases. Finally, Sections 5 and 6 provide discussion of the results and conclude the paper.

2. The Reference Recognizer – RefRec

The RefRec is implemented as a set of Perl scripts. The main training script is, in the latest version (0.96) of the RefRec, called *NoiseTrain* (Johansen et al., 2000; Lindberg et al., 2000). The training procedure starts with the database preparation and feature extraction. For each frame m of the input speech file the feature vector $\mathbf{a}^{\mathbf{m}}$ of length n elements is produced using the Advanced Front-End (AFE) feature extraction procedure. The prototype acoustical monophone model consists of a three state left-to-right diagonal-covariance Gaussian HMM, without skip transitions (Young et al., 2000). HMMs are trained from orthographic (word-level) transcriptions using a pronunciation lexicon. Training starts from context-independent, single Gaussian monophones. The Gaussians are all boot-strapped to the global mean and variance of the training set, followed by supervised embedded Baum-Welch re-estimation. To reduce the problem with unlabelled silence between words, only the phonetically balanced sentences (subcorpus S1-9) are used in the boot-strapping stage. Afterwards, a full state Viterbi realignment (Young et al., 2000) is performed on the whole training set. The output label file `align_32_2.mlf` is generated. This file presents the connection between RefRec and proposed LDA-Toolkit. As will be described in Section 3 the LDA discrimination classes are produced on the basis of `align_32_2.mlf`. Furthermore, the realignment procedure allows lexicon pronunciations other than the canonical ones to be chosen and also identifies potentially erroneous annotations. From the single-mixture monophone models, training proceeds by building word-internal context-dependent models for all triphones occurring in the training set. Word boundaries are modeled with left- or right-context dependent models (biphones). The monophone models are first cloned, then re-estimated with context-dependent supervision. In order to reduce the total number of HMM states and improve generalization ability, state tying is performed. A top-down decision tree clustering approach ensures that unseen words can be modeled without retraining the models, as required for flexible vocabulary recognition (Lindberg et al., 2000). In a final training stage, the tied state triphone models are improved by Gaussian mixture density modeling. Mixture models are generated by successive mixture splitting and re-estimation. The result is a sequence of models with 2, 4, 8, 16 and 32 mixture components, respectively.

3. The LDA-Toolkit and its Integration into the RefRec

Linear discriminant analysis (LDA) has been applied for the transformation of the input feature vector $\mathbf{a}^{\mathbf{m}}$ to the final output feature vector $\mathbf{b}^{\mathbf{m}}$, and to enhance the discriminant power between *K* discrimination classes. This is in order to reduce the computational load of the automatic speech recognition system and to enhance the classification process. The basic idea of the LDA is to reduce the variances within the classes whereas the variances between the classes should be as large as possible (Welling, 1999). Figure 1 represents the block diagram of the usage of LDA-Toolkit for the LDA transformation procedure and its integration into the RefRec. The proposed LDA-Toolkit consists of 13 tools written in C language. The source code is therefore compilable and executable on Linux as well as on Windows platforms. There are 8 main processing tools (*`Class_Covariance`*, *`Class_Mean_Sub`*, *`Full_Class_Generate`*, *`LDA_Matrix_Generate`*, *`MLF_Class_Generate`*, *`WCS_Matrix_Generate`*, *`BCS_Matrix_Generate`*, *`LDA_Transform`*) and 5 general-purpose inspection and emulation tools (*`Class_Covariance_Read`*, *`BWL_Matrix_Read`*, *`Class_Mean_Read`*, *`Class_Matrix_Read`*, *`Emulate_Gen_Feat`*). The following steps describe the LDA processing procedure using particular tool from the LDA-Toolkit.

3.1 Determination of LDA classes

In the proposed LDA procedure, *K* classes correspond to emitting states of hidden Markov model (HMM) of all phonemes in the dictionary, the only exceptions are /sp/ and /sil/, short pause and silence models respectively. Static feature vectors are concatenated with their first and second order time derivatives (Δ-deltas, ΔΔ-delta-deltas) to constitute training feature vectors $\mathbf{a}^{\mathrm{m}}$ of length of 3*13 = 39 elements. Afterwards, the two consecutive feature vectors [$\mathbf{a}^{\mathrm{m}}$, $\mathbf{a}^{\mathrm{m-1}}$] are concatenated to form the one feature vector $\hat{\mathbf{a}}^{m}$ of 78 elements (Welling, 1999). Further, the *NoiseTrain* training script of the RefRec toolkit is executed until the Viterbi full state realignment is performed and `align_32_2.mlf` is created. The *`MLF_Class_Generate`* tool is then used to divide the super-feature vectors of the training material into *K* LDA classes.

3.2 LDA transformation matrix

Once *K* classes are determined, the *K* class-mean vectors (*`Class_Mean_Sub`*) and global-mean vector are determined (*`Full_Class_Generate`* together with *`Class_MeanDetermine`*). The covariance of each class is estimated as follows (*`Class_Covariance`*):

$$\mathbf{C}_i = E\left\{\left(\hat{\mathbf{b}}^m - \mathbf{m}_i\right)\left(\hat{\mathbf{b}}^m - \mathbf{m}_i\right)^T\right\} \text{, where } i \in K \ . \quad (1)$$

In (1) the operator *E*{} represents the expectation operator, $\mathbf{m}_{\mathrm{i}}$ is the mean vector of class $i \in$ K. The mean *within class* scatter matrix $\mathbf{S}_{\mathrm{W}}$, and the mean *between class* scatter matrix $\mathbf{S}_{\mathrm{B}}$ are defined as:

$$\begin{aligned} \mathbf{S_W} &= \sum_{i=1}^{K} P(i)\mathbf{C}_i \\ \mathbf{S_B} &= \sum_{i=1}^{K} P(i)\left((\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T\right) \end{aligned} \ . \quad (2)$$

The two matrices are generated using the *`WCS_Matrix_Generate`* and *`BCS_Matrix_Generate`* executables respectively. The procedure involved in computation of (2) is as follows: In (2) *P*(i) represents the *a priori* probability of class *i*. In the proposed algorithm all classes have equal *a priori* probability *P*(i)=1/K. Since the between-class scatter matrix is calculated from the class mean vectors, those axes should be found, which keep apart these mean vectors as far as possible. Additionally, axes are enforced to be orthogonal, due to the usage of a diagonal covariance matrix in HMM training/recognition procedures. Firstly, decorrelation and variance normalization is performed using a determination of transformation matrix **B**:

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}} \quad (3)$$

where **U** consists of the eigen vectors of the matrix $\mathbf{S_W}$, determined by the solution of the following eigenvalue problem:

$$\mathbf{S_W U} = \mathbf{U\Lambda} \ . \quad (4)$$

Then, the new between-class scatter matrix $\hat{\mathbf{S}}_{\mathbf{B}}$ with transformed class means is determined using the transformation matrix **B**:

$$\hat{\mathbf{S}}_{\mathbf{B}} = \mathbf{B}^T\mathbf{S_B}\mathbf{B} \ . \quad (5)$$

The set of optimal axes, with respect to maximum variance between means, corresponds to the eigenvectors **V** of the following eigenvalue problem:

$$\hat{\mathbf{S}}_{\mathbf{B}}\mathbf{V} = \mathbf{V}\hat{\mathbf{\Lambda}} \ . \quad (6)$$

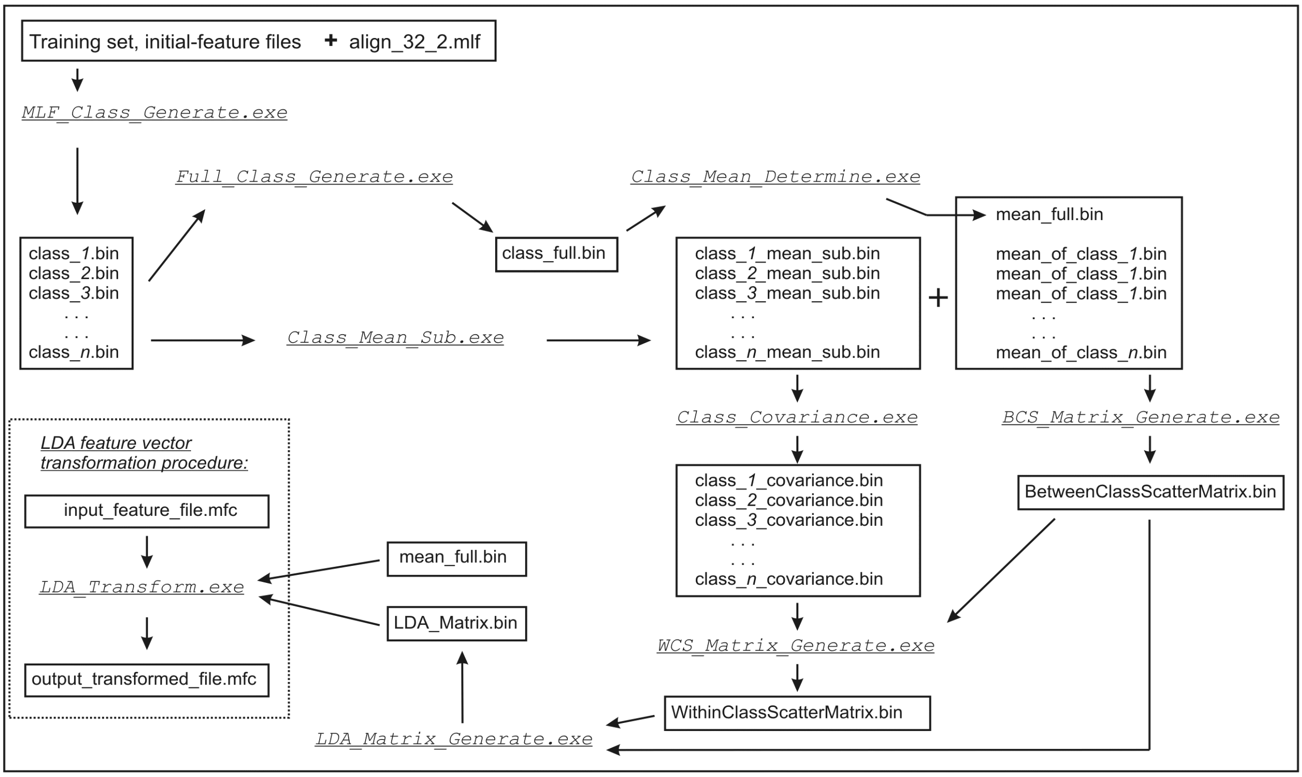

Figure 1: The block diagram of the LDA matrix generation procedure using the proposed LDA-RefRec Toolkit

Since the rank of the matrix $\hat{\mathbf{S}}_\mathbf{B}$ is at maximum K-1, only K-1 axes exist. Therefore, the relevant information is compressed into K-1 eigenvectors $\mathbf{V}$=($\mathbf{v}_1$,$\mathbf{v}_2$,…,$\mathbf{v}_{K-1}$), which correspond to K-1 largest eigenvalues $\widehat{\mathbf{\Lambda}}$. Finally, the linear discriminant analysis transformation matrix $\mathbf{\Omega}$ is defined as (`LDA_Matrix_Generate`):

$$\mathbf{\Omega} = \mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V} = \mathbf{B}\mathbf{V} \quad . \tag{7}$$

After regarding the subtraction of the mean, the final output feature vector is computed using linear discriminant analysis (`LDA_Transform`):

$$\mathbf{b}^m = \mathbf{\Omega}^T\left(\hat{\mathbf{a}}^m - \mathbf{m}\right) \quad . \tag{8}$$

In the proposed algorithm the dimensionality of the final output feature vector $\mathbf{b}^\mathbf{m}$, is reduced from 78 elements (original feature vector $\mathbf{a}^\mathbf{m}$) to L elements. Therefore, the matrix $\mathbf{V}$=($\mathbf{v}_1$,$\mathbf{v}_2$,…,$\mathbf{v}_{78}$) in (6) corresponds to the L largest eigenvalues $\widehat{\mathbf{\Lambda}}$. Transformed final output feature vectors $\mathbf{b}^\mathbf{m}$ with reduced dimension of L elements are then used in repeated training process (*NoiseTrain*). The same feature transformation procedure is applied also to transform the test feature vectors used in the recognition procedure.

4. Experimental Framework and Results

Evaluation of the proposed LDA-RefRec procedure was performed using Slovenian, German, and Spanish SpeechDat (II) databases. The set of 1000 speakers was used in the case of particular database. Automatic speech recognition tests were performed using connected digit test sets B1 and C1. Table 1 presents baseline automatic speech recognition results using AFE (Advanced Front-End) feature extraction procedure. Feature vectors consist of 39 elements (13 static coefficients + Δ + ΔΔ). No feature vector postprocessing algorithms were applied in this experiment. Acoustical models were trained with *NoiseTrain* and test were performed using *NoiseSVWL* test script. The results presented in Table 1 are word error rates (WER) achieved with tied-triphone acoustical models with 1, 2, 4, 8, 16 and 32 mixture components, respectively.

SpeechDat (II) B1, C1 tests % WER	Slovenian SpeechDat FDB 1000	German SpeechDat FDB 1000	Spanish SpeechDat FDB 1000
Tied_1_2	4.61	3.78	3.77
Tied_2_2	4.47	3.04	2.79
Tied_4_2	3.91	2.39	2.38
Tied_8_2	3.63	2.12	1.45
Tied_16_2	2.96	1.93	1.60
Tied_32_2	2.61	2.12	2.17

Table 1: Baseline results (WER) using Advanced Front-End (39 elements in the feature vector) on B1, C1 test sets

Table 2 represents the Slovenian, German, and Spanish SpeechDat (II) automatic speech recognition results (WERs) of connected digit strings (B1, C1) achieved with the usage of the LDA-RefRec Toolkit. Initial-feature files were extracted using AFE. Afterwards, the super-feature files of the dimension of 78 elements were composed.

% WER	Slovenian SpeechDat (II)				German SpeechDat (II)				Spanish SpeechDat (II)			
Dimension	26	30	34	39	26	30	34	39	26	30	34	39
Tied_1_2	6.25	5.19	4.84	6.02	5.21	3.33	3.52	4.11	4.22	3.75	3.32	3.68
Tied_2_2	5.72	4.68	4.55	5.85	4.87	3.00	3.21	2.71	3.96	3.38	2.54	3.47
Tied_4_2	4.40	3.34	4.02	4.67	4.12	2.84	2.84	3.16	3.35	3.03	2.32	2.77
Tied_8_2	3.23	2.51	3.79	4.01	3.08	2.29	2.53	2.72	2.87	2.44	1.88	2.23
Tied_16_2	2.98	2.34	2.88	3.09	2.66	2.05	2.12	2.14	2.26	2.13	1.48	2.02
Tied_32_2	3.12	2.27	2.45	2.84	2.81	1.95	2.27	2.67	2.13	2.49	2.10	2.28

Table 2: The Slovenian, German, and Spanish SpeechDat (II) automatic speech recognition results (WERs) of connected digit strings (B1, C1) achieved with the usage of the LDA-RefRec Toolkit at different dimensions (26, 30, 34, and 39 elements) of the output feature vector.

With the usage of the Viterbi full state realignment procedure performed in the baseline experiment, the super-features were divided into *K* classes, where *K* is the product of the number of phoneme models and the number of emitting states for each phoneme model in particular database. Then, the LDA matrix is computed and LDA transformation is performed. In particular experiment the dimensions of the final output feature vectors of 26, 30, 34, and 39 were considered. The best results are achieved using feature vectors with 30 elements (Slovenian, German FDB), or 34 elements (Spanish FDB). Table 3 represents automatic speech recognition results achieved with the usage of multilingual LDA classes. In this case the single multilingual LDA matrix is used to transform the data of all SpeechDat (II) databases considered. Additionally, the dimension of the final output feature vectors is reduced to 30 elements as this dimension is found to produce the best results in the monolingual case (Table 2).

B1, C1 tests % WER	Slovenian SpeechDat FDB 1000	German SpeechDat FDB 1000	Spanish SpeechDat FDB 1000
Tied_1_2	6.11	3.76	4.09
Tied_2_2	4.69	3.17	3.23
Tied_4_2	4.27	2.59	2.37
Tied_8_2	3.15	2.31	1.99
Tied_16_2	2.55	1.85	1.58
Tied_32_2	2.74	2.08	1.54

Table 3: The B1, C1 results with the usage of multilingual LDA classes (30 elements in the output feature vectors)

5. Discussion

It is evident from the baseline automatic speech recognition results (Table 1) that with the usage of AFE (Advanced Front-End) feature extraction procedure better performance than with the basic MFCC feature extraction procedure (RefRec home, 1999) is achieved. The performance of the original RefRec reference recognizer is further improved with the application of the proposed LDA-RefRec toolkit. It is evident from the comparison of automatic speech recognition results presented in Tables 2 and 3 that with the usage of multilingual LDA matrix slightly worse performance than with monolingual LDA is achieved. Nevertheless, in the case of multilingual LDA only one LDA matrix needs to be constructed for all the three databases (reduced computational load).

6. Conclusion

The automatic speech recognition experiments using LDA-RefRec Toolkit and Slovenian, German, and Spanish SpeechDat (II) databases show performance improvement when compared to basic RefRec reference recognizer results. Additionally, the computational requirements of the automatic speech recognition system as well as the real-time factor are improved due to lower order of the final output feature vector. Proposed LDA-RefRec Toolkit is therefore found to be a powerful tool for construction and evaluation of computationally efficient automatic speech recognition systems based on SpeechDat (II) databases. Nevertheless, the proposed LDA-Toolkit could be used also with other speech databases.

7. References

ETSI standard document – ETSI ES 202 050 v1.1.1 (2002). Speech Processing, Transmission and Quality aspects (STQ), Distributed speech recognition, Advanced front-end feature extraction algorithm, Compression algorithm, 2002.

Johansen, F.T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., Salvi, G. (2000). The COST 249 SpeechDat Multilingual Reference Recogniser. In: *Proc. LREC'2000*, Athens.

Lindberg, B., Johansen, F.T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., Salvi, G. (2000). A Noise Robust Multilingual Reference Recogniser Based on SpeechDat(II). In: *Proc. ICSLP 2000*, Beijing, China.

Refrec home, http://www.telenor.no/fou/prosjekter/taletek/refrec/

Schafföner, M., Katz, M., Krüger, S.E., and Wendemuth, A. (2003). Improved Robustness of Automatic Speech Recognition using a New Class Definition in Linear Discriminant Analysis. In: *Proc. Eurospeech 2003*, Geneva, Switzerland.

Welling, L. (1999). Merkmalsextraction in Spracherkennungssystemen fűr grossen Wortschatz. PhD thesis, RWTH, Aachen, Germany.

Young S., Kershaw D., Odell J., Ollason D., Valtchev V, Woodland P. (2000). The HTK Book - Version 3.0. Microsoft Corporation, USA.