

Improved Robustness of Automatic Speech Recognition using a New Class Definition in Linear Discriminant Analysis

M. Schafföner, M. Katz, S. E. Krüger, and A. Wendemuth

Cognitive Systems Group, Dept. of Electrical Engineering,
Otto-von-Guericke-University Magdeburg,
PF 4120, 39016 Magdeburg, Germany
schaffoe@iesk.et.uni-magdeburg.de

Abstract

This work discusses the improvements which can be expected when applying linear feature-space transformations based on Linear Discriminant Analysis (LDA) within automatic speech-recognition (ASR). It is shown that different factors influence the effectiveness of LDA-transformations. Most importantly, increasing the number of LDA-classes by using time-aligned states of Hidden-Markov-Models instead of phonemes is necessary to obtain improvements predictably. An extension of LDA is presented, which utilises the elementary Gaussian components of the mixture probability-density functions of the Hidden-Markov-Models' states to define actual Gaussian LDA-classes. Experimental results on the TIMIT and WSJCAM0 recognition task are given, where relative improvements of the error-rate of 3.2% and 3.9%, respectively, were obtained.

1. Introduction

In Automatic Speech Recognition (ASR), the parameter vectors usually have a great number of dimensions, typically in the order of 20–50. Parameter-vectors belonging to classes to be recognised may not be well distributed in this high-dimensional space, i.e. the distribution does not facilitate easy separation of classes. To overcome this problem, methods such as Linear Discriminant Analysis (LDA) and derivatives thereof can be used to transform the original high-dimensional space into a different, possibly lower-dimensional one while retaining, or even improving, class-separability.

The use of LDA for feature-space transformations in ASR-applications was first introduced in [1]. It was later the subject of numerous other publications, most notably [2,3]. Also, modifications of LDA and other discriminative techniques were then investigated in the framework of ASR [4–6].

1.1. Feature Selection, Theory of LDA, and Nomenclature

Dynamic features are used which characterise the temporal change in the vicinity of the analytic window. If

$$\dot{\mathbf{x}}^{(m)} = \left(x_1^{(m)}, \dots, x_d^{(m)} \right) \quad (1)$$

is the d -dimensional vector of static short-term features for the m th analytic window, the general procedure is to compute some new features by the use of a matrix of temporal neighbourhood, and to combine them, together with the original feature-vector $\dot{\mathbf{x}}^{(m)}$, into a new feature-vector $\mathbf{x}^{(m)}$.

The simplest method to obtain such an augmented vector is to concatenate a number of neighbouring vectors into one new

vector. More sophisticated methods include the use of first and higher-order derivatives, filtering and Fourier-transformation. The feature-vectors which are augmented by dynamic properties possess a high dimensionality with strongly correlated components. It is, therefore, advisable to apply techniques to reduce this dimensionality while retaining as much discriminative information as possible. In the context of ASR, we discuss LDA here only.

LDA assumes that the classificatory information contained in the original feature-vectors $\mathbf{x} \in \mathbb{R}^n$ can be fully described by vectors $\mathbf{y} \in \mathbb{R}^p$ with $p \leq n$. The necessary linear transformation is achieved by a matrix Θ [7] which must be estimated by a sufficiently large number of training-samples:

$$\mathbf{y} = \Theta^T \mathbf{x} \quad (2)$$

Consider J classes with N_j elements (feature-vectors) $\mathbf{x}_{j,i}$ in class j , $j \in \{1, \dots, J\}$ and $1 \leq i \leq N_j$. Let $N = \sum_{j=1}^J N_j$ be the total number of feature-vectors. Then the classes can be characterised by their arithmetic means $\bar{\mathbf{x}}_j$ and covariances \mathbf{S}_j , assuming they are Gaussian:

$$\bar{\mathbf{x}}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{x}_{j,i} \quad (3)$$

$$\mathbf{S}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} (\mathbf{x}_{j,i} - \bar{\mathbf{x}}_j)(\mathbf{x}_{j,i} - \bar{\mathbf{x}}_j)^T \quad (4)$$

Furthermore, let $\bar{\mathbf{x}}$ be the mean of all feature-vectors, disregarding their class:

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{j=1}^J \sum_{i=1}^{N_j} \mathbf{x}_{j,i} \quad (5)$$

The class information can then be condensed into two scatter-matrices called *within-class scatter-matrix* \mathbf{W} and *between-class scatter-matrix* \mathbf{B} , as follows:

$$\mathbf{W} = \frac{1}{N} \sum_{j=1}^J N_j \mathbf{S}_j \quad (6)$$

$$\mathbf{B} = \frac{1}{N} \sum_{j=1}^J (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T \quad (7)$$

Standard LDA now tries to find the transformation Θ that maximises the objective function

$$J(\Theta) = \frac{|\Theta^T \mathbf{B} \Theta|}{|\Theta^T \mathbf{W} \Theta|}, \quad (8)$$

i.e. it strives to maximise the between-class scatter over the within-class scatter. Ultimately, the classes will be compact in the projected space, with the class-means well separated.

Although the objective function in (8) is non-linear, there is a closed-form solution composed from by the eigenvectors \mathbf{v} corresponding to the p largest eigenvalues λ of the generalised eigenvalue-problem

$$\mathbf{B}\mathbf{v} = \lambda\mathbf{W}\mathbf{v}, \quad (9)$$

which can be solved by the standard eigenvalue-problem

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{V} = \mathbf{V}\mathbf{\Lambda} \quad (10)$$

where

$$\mathbf{V} = (\mathbf{v}^1 \dots \mathbf{v}^n) \quad (11)$$

$$\mathbf{\Lambda} = \text{diag}(\lambda^1 \dots \lambda^n) \quad (12)$$

$$\mathbf{\Theta} = (\mathbf{v}^1 \dots \mathbf{v}^p) \quad (13)$$

2. Class-Definition for LDA

The question of appropriate class-definitions and their interactions with the estimation of the LDA-matrix have only been answered on an argumentative level. In general, better results are obtained if more classes than only the monophones are used. This is quite clear as LDA tries to model each class with a single Gaussian PDF. However, the ASR-systems themselves are examples that single Gaussian PDFs are not sufficient to accurately model sub-word elements. ASR-systems have several distinct states for each monophone, with each state possibly using mixture PDFs consisting of elementary Gaussians. The choice of sub-monophone elements, therefore, seems to be justified. The authors of [1, 2, 8] suggest different combinations of context-dependant triphones, monophones, and their respective segments as used in the recogniser.

Sub-phoneme class-definitions can be found as follows: During the re-alignment process, the time-alignment of both the monophones and their respective emitting states are recorded. The number of the resulting classes depend on the HMM-system used in the re-alignment process. It is clear that, unless the HMM-states' emission-probabilities are modelled with single Gaussians, the HMM-states are not perfect choices for LDA class-identifiers.

However, if the HMM-states emission-PDFs are modelled by mixtures of elementary Gaussian distributions, it is quite easy to find a class definition which perfectly fits LDA's assumption of Gaussian class-distributions by using the mixture-components as LDA class-identifiers. This approach directly targets the Gaussicity of the LDA-classes, in contrast to previous solutions which only targeted this indirectly by the sheer number of classes [1, 2, 8].

The problem is that an alignment, which is necessary to assign feature-vectors to LDA-classes, is only available at the level of models and their states. A time-alignment of mixture-components cannot be obtained this way, because recognisers have no knowledge about the structure of emission-PDFs of HMM-states. Therefore, a sort of post-classification scheme must be employed which assigns feature-vectors to mixture-components, provided that the model and state is known from the alignment.

2.1. Modified LDA for HMM-based Speech Recognition

If a mixture-PDF of HMM m in state s has G_{ms} elementary Gaussian PDFs, the mixture-PDF can be written as

$$p(\mathbf{x}; m, s) = \sum_{g=1}^{G_{ms}} c_{msg} \mathcal{N}(\mathbf{x}; \hat{\boldsymbol{\mu}}_{msg}, \hat{\boldsymbol{\Sigma}}_{msg}) \quad (14)$$

with $\mathcal{N}(\mathbf{x}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}_{msg})$ a single Gaussian with mean $\hat{\boldsymbol{\mu}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{msg}$.

Assuming that we know the model m 's state s from the time-alignment, we can dissect this state's mixture-PDF into its elementary normal distributions (which we want to use as classes) and compute the probabilities of the feature-vector \mathbf{x} being produced by the individual elementary PDFs. These probabilities are then

$$p(\mathbf{x}, g|m, s) = c_{msg} \mathcal{N}(\mathbf{x}|\hat{\boldsymbol{\mu}}_{msg}, \hat{\boldsymbol{\Sigma}}_{msg}) \quad (15)$$

Previously, each vector \mathbf{x} could be uniquely assigned to a class using the time-alignment. With mixture-components as LDA-classes, the state-time-alignment leaves us with a set of G_{ms} possible classes to which the vector \mathbf{x} could be assigned. This assignment must be determined from the probabilities computed using (15).

Two ways to make this decision were investigated, a "soft" decision were the probabilities are only normalised w.r.t. the sum of the G_{ms} probabilities, according to Bayes' rule:

$$w_g(\mathbf{x}_{ms}) = p(g|\mathbf{x}_{ms}, m, s) = \frac{p(\mathbf{x}_{ms}, g|m, s)}{\sum_{y=1}^{G_{ms}} p(\mathbf{x}_{ms}, y|m, s)}, \quad (16)$$

and a "hard" decision were only the (mixture-component-) class with the greatest probability is assigned the vector, as in

$$w_g(\mathbf{x}_{ms}) = \begin{cases} 1 & \text{if } g = \text{argmax}_y(p(\mathbf{x}_{ms}, y|m, s)) \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

Since the contributions of feature-vectors to classes can now also be from the real-valued interval $[0, 1]$, the LDA-equations were adapted to reflect this situation. The notation is changed to make class-dependant variables explicitly name the model, state and mixture-component they were derived from. The whole-number counters N and N_{ms} count the number of vectors as determined by the state-time-alignment, while the contributions of the vectors to the (mixture-component-) classes are tracked by the real-number variables W_{msg} and W

$$W_{msg} = \sum_{i=1}^{N_{ms}} w_g(\mathbf{x}_{ms,i}) \quad (18)$$

$$W = \sum_{m=1}^M \sum_{s=1}^{S_m} \sum_{g=1}^{G_{ms}} W_{msg} \quad (19)$$

If we choose to re-estimate the LDA-classes' parameters $\bar{\mathbf{x}}$ and \mathbf{S} , the formulae, using the weighting of the vectors, are now:

$$\bar{\mathbf{x}}_{msg} = \frac{\sum_{i=1}^{N_{ms}} w_g(\mathbf{x}_{ms,i}) \mathbf{x}_{ms,i}}{W_{msg}} \quad (20)$$

$$\mathbf{S}_{msg} = \frac{\sum_{i=1}^{N_{ms}} w_g(\mathbf{x}_{ms,i}) (\mathbf{x}_{ms,i} - \bar{\mathbf{x}}_{msg})(\mathbf{x}_{ms,i} - \bar{\mathbf{x}}_{msg})^T}{W_{msg}} \quad (21)$$

The global mean is now computed from the class-means, to make the easy recycling of HMM-parameters possible.

$$\bar{\mathbf{x}} = \frac{1}{W} \sum_{m=1}^M \sum_{s=1}^{S_m} \sum_{g=1}^{G_{ms}} W_{msg} \bar{\mathbf{x}}_{msg} \quad (22)$$

The within-class and the between-class scatter-matrices from (6) and (7), respectively, can now be rewritten to reflect the new hierarchical structure of the class-definition and the weighting:

$$\mathbf{W} = \frac{1}{W} \sum_{m=1}^M \sum_{s=1}^{S_m} \sum_{g=1}^{G_{ms}} W_{msg} \mathbf{S}_{msg} \quad (23)$$

$$\mathbf{B} = \frac{1}{W} \sum_{m=1}^M \sum_{s=1}^{S_m} \sum_{g=1}^{G_{ms}} W_{msg} (\bar{\mathbf{x}}_{msg} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{msg} - \bar{\mathbf{x}})^T \quad (24)$$

With these two matrices rewritten to the new application, we are still able to maximise (8) through the eigenvalue-problem (10).

3. Experiments

3.1. Training and Test Conditions

The HMM-systems used for experiments on the TIMIT-corpus [9] have 50 HMMs which model 1 phoneme each. The HMMs have a left-right structure with 6 emitting states. The emission-PDF of each state is modelled with mixture-PDFs consisting of 5 pooled elementary Gaussian PDFs with diagonal variances. The training bootstraps the HMMs with one Gaussian for each emission-PDF, re-estimates the parameters of the HMMs four times, and splits the elementary Gaussian PDF into a mixture of two Gaussians. This new system is re-estimated four times and “mixed-up” to 3 Gaussians per mixture-PDF. This process is repeated until a system with five Gaussians per mixture has been re-estimated four times, which results in the final systems to test. During testing, some of the phonemes are treated as being equivalent, and silence and short-pauses are ignored, so that 39 phonemic entities remain for testing.

On WSJCAM0 [10], the systems were similar, with the sole difference that only 45 phonemes were modelled with 3 emitting states per mixture. Testing on WSJCAM0 was carried out on the word level.

The LDA base-vectors had $n = 39$ components, consisting of 13 static components (12 cepstral coefficients and 1 log-energy) plus first- and second-order time-derivatives. Even though the literature suggests higher-dimensional input-vectors [1, 2, 5], this approach could not be followed due to limited feature-construction capabilities of the ASR-system used here. The properties of the ASR-systems are summarised in table 1.

3.2. Results

On both TIMIT and WSJCAM0, five experiments were conducted. The first was to determine the base recognition-performance without LDA, the other four were LDA-experiments with phonemes, HMM-states, and mixture-components (using both hard and soft weighting from (16) and (17)) as LDA-classes. No dimensionality reduction was applied, only simple transformations, because experiments show that a target dimensionality of 30–40 is optimal [11].

As already mentioned in the introduction, the eigenvalues are measures for the strength of the class-separation in

Configuration variable	Value
Parametrisation	10ms
Sampling-Period	
Window kind	Hamming
Window size	25ms
# Phonemes (HMMs)	TIMIT:50, WSJCAM0:45
HMM type	TIMIT:6, WSJCAM0:3 emitting states, left-to-right, no tying
Mixture components	5 pooled Gaussians, diagonal variances
# Training iterations	4
# Tested entities	TIMIT: 39 phonemes, WSJCAM0: 5000 words

Table 1: Properties of TIMIT- and WSJCAM0-systems

the direction of the corresponding eigenvector. Therefore, an inspection of the eigenvalue-spectra should give a hint about the recognition-improvement to be expected. Figure 1 shows the first 10 eigenvalues of LDA-estimations for three different class-identifiers (phonemes, HMM-states, Gaussian mixture-components) on the TIMIT corpus. Results on WSJCAM0 are of comparable nature.

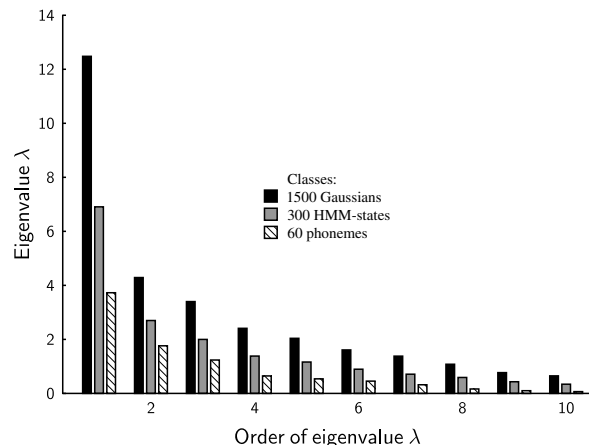


Figure 1: Eigenvalue-spectra of LDA-estimations using different LDA class-identifiers on TIMIT

The new class definition using Gaussian mixture-components yields greatly increased eigenvalues when compared to the traditional class-identifiers HMM-states and phonemes. The impact of the new LDA-transformations, which are expected to decrease the recognition-error, is shown in figure 2 for TIMIT and in figure 3 for WSJCAM0.

The main conclusion which can be drawn from the recognition-experiments is that it is indeed possible to improve the effectiveness of LDA-transformations using Gaussian mixture-components as LDA-classes. In the case of TIMIT, only the hard weighting-scheme could improve the recognition performance, albeit moderately. On WSJCAM0, the error was decreased for either weighting scheme, although here, in contrast to TIMIT, the soft weighting-scheme performed better.

The decision which weighting-scheme to prefer in general is not easy. Theoretically, every vector should be used for the parameter-estimation of each mixture-component class ac-

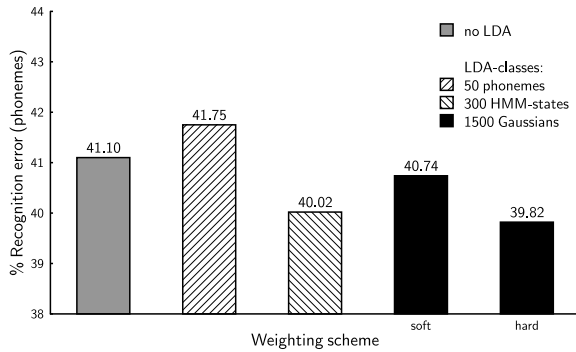


Figure 2: Recognition-error on TIMIT depending on the LDA class-identifier and the weighting-scheme

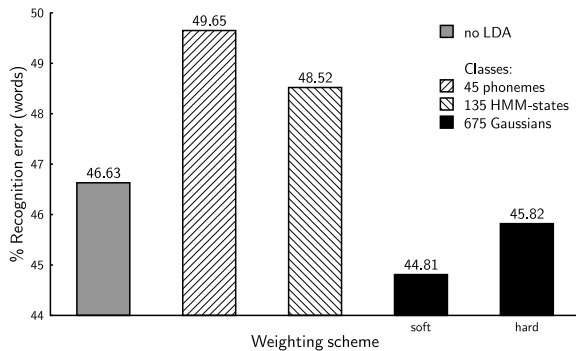


Figure 3: Recognition-error on WSJCAM0 depending on the LDA class-identifier and the weighting-scheme

ording to its posterior probability, which would equal the soft weighting-scheme. However, from experiments presented here and in [11] it is found that only the hard weighting-scheme for mixture-component classes can reliably and predictably decrease the recognition-error when compared to phonemes and HMM-states as LDA-classes. Also, the hard weighting is more economic in terms of computation-time because each vector is only considered in the parameter-estimation of *one* single Gaussian mixture-component class.

4. Conclusion

This work discussed the influence of different class-identifiers on the effectiveness of feature-space transformations obtained from Linear Discriminant Analysis. It was found that time-aligned HMM-states provide better LDA-transformations than phonemes because they better fit LDA's assumption of Gaussian class-distributions.

A novel method of defining LDA-classes by assessing the class-membership as posterior probabilities from Gaussian components of the HMM-states' mixture probability-density functions was introduced, which directly targets LDA's requirement of Gaussian classes. This allowed easy and flexible generation of LDA-classes to improve the LDA-transformation with respect to the recognition-rates. By turning mixture-components into LDA-classes using the presented assessment-method, the recognition-result was improved by 3.2% (relative) on TIMIT, where a similar relative improvement of 2.7% had already been reached using HMM-states as LDA class-identifiers. On WSJCAM0, however, a relative improvement of 3.9% was

obtained, where other LDA class-definitions had only resulted in deteriorated recognition-results.

5. Acknowledgements

I would like to thank the following people for helping me throughout the various stages of this work: Marcel Katz for his patience and valuable remarks about the setup of the experiments, Andreas Wendemuth for his comments on the direction of this work and for revising the manuscript, and Sven Krüger for important hints on the methodology of experiments in general.

6. References

- [1] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for large vocabulary speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, San Francisco, 1992, pp. 13–16.
- [2] K. Beulen, L. Welling, and H. Ney, "Experiments with linear feature extraction in speech recognition," in *Proc. Europ. Conf. on Speech Communication and Technology*, Madrid, Spain, 1995, pp. 1415–1418.
- [3] N. Kumar and A. G. Andreou, "A generalization of linear discriminant analysis in maximum likelihood framework," in *Proceedings of the Joint Statistical Meeting*. Chicago: Statistical computation group, 1996.
- [4] N. Kumar, "Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, Johns Hopkins University, Baltimore, MD, USA, 1997.
- [5] M. Loog and R. Haeb-Umbach, "Multi-class linear dimension reduction by generalized Fisher criteria," in *Proceedings of the International Conference on Spoken Language Processing*, 2000.
- [6] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. ICASSP2000*, 2000.
- [7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., ser. Computer Science and Scientific Computing, W. Rheinboldt, Ed. San Diego: Academic Press, 1990.
- [8] J. Duchateau, K. Demuyne, D. van Campennolle, and P. Wambacq, "Class definition in discriminant feature analysis," in *Proc. European Conference on Speech Communication and Technology*, vol. 3, Aalborg, Denmark, September 2001, pp. 1621–1624.
- [9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus," U.S. Dept. of Commerce, NIST, Gaithersburg, MD, February 1993.
- [10] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 corpus and recording description," Cambridge University Engineering Department (CUED) Speech Group, Trumpington Street, Cambridge CB2 1PZ, UK, Tech. Rep., September 1994.
- [11] M. Schafföner, "Improved robustness of automatic speech recognition using linear discriminant analysis," Master's thesis, Otto-von-Guericke-University Magdeburg, March 2003.