Feature Combination using Linear Discriminant Analysis and its Pitfalls

Ralf Schlüter, András Zolnay, and Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Department
RWTH Aachen University, Aachen, Germany
schlueter@cs.rwth-aachen.de

Abstract

In this paper, Linear Discriminant Analysis (LDA) is investigated with respect to the combination of different acoustic features for automatic speech recognition. It is shown that the combination of acoustic features using LDA does not consistently lead to improvements in word error rate. A detailed analysis of the recognition results on the Verbmobil (*VM II*) and on the English portion of the European Parliament Plenary Sessions (*EPPS*) corpus is given. This includes an independent analysis of the effect of the dimension of the input to LDA, the effect of strongly correlated input features, as well as a detailed numerical analysis of the generalized eigenvalue problem underlying LDA. Relative improvements in word error rate of up to 5% were observed for LDA-based combination of multiple acoustic features.

1. Introduction

In [1], Linear Discriminant Analysis (LDA) was first applied successfully to find an optimal linear combination of successive vectors of a feature stream for automatic speech recognition. LDA could equally well be used to combine different features. In [2], direct combination of different cepstral features was done using LDA, however without significant improvements in word error rate (WER) compared to using the MFCCs alone. LDA also is used successfully in the RWTH ASR system to take advantage of an additional voicing feature [3]. Nevertheless, further experiments presented here show improvements in word error rate are not guaranteed using LDA to combine different acoustic features. Specifically, we will present cases, in which LDA based feature combination leads to degradations. A major difference to using LDA on single feature systems is the large increase in the dimension of the input. This might induce numerical problems with respect to the estimation problem to be solved within LDA, especially if the features to be combined are strongly correlated. The robustness of LDA with respect to increasing the input dimension has been addressed in earlier investigations. In [4] a decrease in WER was observed when overly increasing the LDA input window length. In [5], addition of random coefficients to the feature vectors for an artificial recognition problem also showed degradations. The latter experiment was repeated for speech recognition on real data in this work and could not be confirmed. We review the above-mentioned inconsistencies observed w.r.t. the performance of LDA-based feature combination from the point of view of numerical stability of the underlying eigenvalue problem to be solved within LDA.

The paper is structured as follows. In Section 2, the usage of LDA for feature combination is introduced. An overview of the experimental setup is given in Section 3. Recognition results using LDA for feature combination are presented in Section 4, followed by a detailed analysis of the results in Section 5, including further experiments motivated by the analysis. The paper is concluded by a summary in Section 6.

2. LDA-based Feature Combination

In the following, we describe a straightforward way to use LDA for feature combination, and we discuss problems that might arise when combining acoustic features using LDA.

To combine different acoustic features, for each time frame t the feature vectors $x_t^{f_i}$ of each feature f_i with $i = 1, \ldots, I$ are concatenated to build multi-feature vectors $x_t = (x_t^{f_1}, x_t^{f_2}, \ldots, x_t^{f_I})$. To also take into account the acoustic context, then the multi-feature vectors for a number of successive time frames are concatenated to build a multi-feature vector $X_t = (x_{t-n}, x_{t-n+1}, \ldots, x_{t+n})$ centered around time frame t, covering the acoustic input of all combined features within a window of $2n + 1$ time frames. Finally, a combined feature vector y_t is created by projecting X_t into a subspace of reduced dimension: $y_t = V^T X_t$. The transformation matrix V is determined by LDA such that it conveys the most relevant classification information to the transformed feature vectors y_t. The resulting acoustic vectors are used both in training and in recognition.

3. Experimental Setup

Characteristics of the RWTH recognition system are summarized in Table 1 for two large vocabulary speech corpora: the *VerbMobil II* (*VM II*) and the English partition of the European Parliament Plenary Sessions (*EPPS*) corpus. The *VM II* corpus consists of German conversational speech whereas the *EPPS* corpus contains plenary session speeches of the European Parliament in British English. Acoustic modeling for both corpora is summarized in the following. The optimized LDA output dimension is 45. Gender independent cross-word triphone sub-word units are used. Triphones are clustered using a Classification and Regression Tree (CART). The resulting generalized triphone states are modeled by Gaussian mixture distributions with a pooled diagonal covariance matrix. Further properties differing for both systems are summarized in Table 1. On the *VM II* corpus, tests have been performed only on the *evaluation* set. On the *EPPS* corpus, we present recognition results on both the *development* and the *evaluation* sets. The baseline experiments using a single feature apply LDA in the same way as the feature combination experiments. The only difference is that we use one feature resulting in a smaller LDA input dimension. Nevertheless, the size of the projected feature vectors is kept constant throughout different experiments to ensure comparability of the corresponding numbers of parameters and recognition results.

Table 1: Settings of the RWTH recognition systems for the *VM II* and the *EPPS* corpus.

corpora	name	*VM II*		*EPPS*		
	partition	train	eval	train	dev	eval
size	speech [h]	61.5	1.6	40.8	3.7	3.5
	# speakers	857	16	154	16	36
lexicon	vocabulary	10,157		54,265		
lang. model	type	class-trigram		trigram		
	perplexity	62.0		87		99
LDA	window	11 frames		9 frames		
	output dim.	45		45		
HMM	topology	3 states w/ skip		6 states		
	silence	1 state		1 state		
	# states	3,501		4,501		
	# densities	≈396k		≈446k		

4. Recognition Results

The experiments presented in the following are meant to test the ability of LDA to combine an increasing number of different acoustic features. Table 2 shows results for LDA-based feature combination of MFCC, vocal tract length normalized MFCC (VTLN), voicing (V), and spectrum derivative (SD) features [6] compared to the best single-feature result. On both corpora, the subsequent combination of the VTLN, voicing, and spectrum derivative features results in consistent successive improvements of WER. Finally, we have added the MFCC feature to test the robustness of LDA against increasing input size. We have not expected any significant change in WER since the MFCC feature is strongly related to the VTLN one. As shown in Table 2, the additional MFCC feature has yielded neither in a significant improvement nor degradation in WER. In contrast to the above consistent improvements in WER, we have obtained unexpected degradation when combining the MFCC, MF-PLP, and PLP [7] features. Table 3 summarizes the baseline recognition results of the individual features and the results obtained by the LDA-based combination denoted by Σ_{LDA}. On both corpora, we have obtained strong degradation in WER. This observation does not comply with the results shown in Table 2. There, the combination of the VTLN, V, and SD features with the much weaker performing MFCC feature did not cause any significant degradation in WER. A possible explanation can be found if we consider the correlation between the features as discussed in the following section.

Table 2: Consistent improvements in WER obtained by LDA-based combination of increasing number of acoustic features.

corpus	acoustic feature	error rates [%]					
		dev			eval		
		del	ins	WER	del	ins	WER
VM II	VTLN				3.8	2.9	19.1
	+V				4.1	2.7	18.7
	+SD				3.9	2.9	18.4
	+MFCC				3.6	2.8	18.3
EPPS	VTLN	4.3	1.3	14.2	3.7	1.5	14.1
	+V	4.0	1.5	13.8	3.3	1.6	14.0
	+SD	3.6	1.6	13.7	3.1	1.8	14.0
	+MFCC	3.7	1.6	13.8	3.3	1.9	14.1

Table 3: Degradation in WER obtained by LDA-based combination of baseline features MFCC, MF-PLP, and PLP.

corpus	acoustic feature	error rates [%]					
		dev			eval		
		del	ins	WER	del	ins	WER
VM II	MFCC				4.5	2.9	21.0
	MF-PLP				5.2	2.3	21.0
	PLP				5.9	2.3	21.4
	Σ_{LDA}				4.7	3.3	21.6
EPPS	MFCC	4.3	1.4	14.7	3.8	1.7	15.3
	MF-PLP	4.2	1.5	14.8	3.7	1.7	15.3
	PLP	4.3	1.6	15.4	3.5	1.8	15.8
	Σ_{LDA}	4.8	1.4	15.8	4.1	1.6	16.2

5. Analysis of Results

As mentioned in the introduction, aspects of the application of LDA in speech recognition have already been addressed before. In [4], experiments have been presented with increasing LDA window length i.e. with increasing number of successive concatenated feature vectors. Instead of converging improvements in WER, a clear optimum was found at 11 concatenated feature vectors. In [5], feature vectors of an artificial recognition task have been augmented with an increasing number of white noise components. The classification error rate was doubled when augmenting a two-dimensional feature vectors with 200 white noise components. Also, a real speech recognition system was used to test the effects of an increasing LDA window length. The authors found that increasing the LDA window length requires an increasing amount of training data to retain the best recognition result. In the following, the problems with LDA-based feature combination presented here and in literature are analyzed in more detail.

5.1. Combination with White Noise Components

A possible explanation for the degradation in WER for using LDA with increasing input dimension could be instabilities of the underlying *generalized eigenvalue problem* resulting from the increased dimension. If we assume that the degradation in WER is caused by numerical instabilities then additional artificial white noise components should be able to induce these problems respectively cause increasing WER. Corresponding experiments were performed using the MFCC feature and constant LDA window length. In order to increase the input dimension of LDA, the 16 baseline MFCC components have been augmented with white noise components simulating additional features for each time frame. To rule out singularities resulting from dependent features, we required the added random features to be independent. Standard random number generators do not comply with this requirement. E.g. for a corpus of $\approx$ 60 hours of training data, using a 100-dimensional white noise extension for every MFCC vector requires a random number generator with a periodicity greater than 3.6×10^9. In our experiments, we have used a random number generator from [8], which ensured a sufficient minimum period of 2×10^{18}. Recognition results are summarized in Table 4. We have extended the 16 MFCC components with up to 90 white noise coefficients per time frame. The resulting concatenated LDA input vectors have grown up to 954 components per time frame. For both corpora, the white noise components have not caused any degradation in WER, i.e., increasing the feature vector size by

Table 4: WER obtained by LDA-based combination of MFCC features and increasing number of randomly generated coefficients.

corpus	# rnd cmp	# LDA input	error rates [%]					
			dev			eval		
			del	ins	WER	del	ins	WER
VM II	0	176				4.5	2.9	21.0
	$15 * 11$	341				4.5	3.0	20.9
	$30 * 11$	506				4.5	3.0	20.9
	$38 * 11$	594				4.8	3.0	21.0
EPPS	0	176	4.3	1.4	14.7	3.8	1.7	15.3
	$30 * 9$	414	4.3	1.3	14.6	3.8	1.7	15.2
	$60 * 9$	684	4.2	1.4	14.6	3.9	1.7	15.2
	$90 * 9$	954	4.3	1.4	14.8	3.9	1.7	15.4

up to a factor of nearly 7 apparently does not introduce numerical problems to LDA when numerically solving the generalized eigenvalue problem.
In contrast to [5], the experiments on real data presented here have not led to significant changes in WER. Nevertheless, it should be mentioned that the average number of observations per LDA class observed here (4500) differs strongly from [5], where only 100 observations were presented per LDA class.

5.2. Sensitivity of Eigenvalues and Vectors

The application of LDA to feature combination has led to both improvements and degradations in WER when combining an increasing number of features. In Section 5.1 we have shown that augmenting the features with further uncorrelated random features does not lead to degradations. So what about additional correlated features? A strong correlation between features can lead to singular scatter matrices. Generally, an indefinite symmetric matrix pair (A, B) may lead to complex eigenvalues and may not have a complete set of generalized eigenvectors. Note that if both A and B are (close to) singular, then any complex number λ is a valid eigenvalue. In all our experiments, we have used the linear algebra software library *LAPACK* [9] to solve the generalized eigenvalue problem. Although the matrices involved are symmetric, we have applied the more general *dggevx* algorithm of *LAPACK* developed for generalized non-symmetric eigenvalue problems, since it is designed to cope with indefinite matrix pairs. The algorithm can be summarized as follows. Assume the within- and between-class scatter matrices B and W and the left and right eigenvectors y_i and x_i, and corresponding eigenvalue λ_i, respectively:

$$Bx_i = \lambda_i W x_i, \qquad y_i^H B = \lambda_i y_i^H W.$$

Reducing (B, W) to the generalized upper *Hessenberg* form, a generalized *Schur* decomposition results in the upper triangular matrix pair (S, T). The left and the right eigenvectors are computed from (S, T). The corresponding eigenvalues λ_i are calculated from the diagonal elements of (S, T):

$$\lambda_i \quad = \quad \frac{S_{ii}}{T_{ii}} = \frac{\alpha_i}{\beta_i}. \tag{1}$$

Note that before calculating the eigenvalues, (near) singular cases can be found by checking (α_i, β_i) for values close to zero.

Assume estimated (α', β') leading to a real eigenvalue λ of the perturbed matrix pair $(B + E, W + F)$ with $||(E, F)|| = \epsilon||(B, W)||_1$ and ϵ is the 64bit machine precision. In perturbation theory, for generalized eigenvalues the *chordal distance* between the corresponding unperturbed (α, β) and the perturbed (α', β') is defined as

$$\mathcal{X}((\alpha, \beta), (\alpha', \beta') = \frac{|\alpha\beta' - \alpha'\beta|}{\sqrt{|\alpha|^2 + |\beta|^2}\sqrt{|\alpha'|^2 + |\beta'|^2}}. \tag{2}$$

Now, instead of discussing perturbations to a possibly singular eigenvalue, perturbations to α and β are addressed in relation to a perturbation of the scatter matrices. Then, an *asymptotic upper bound* for the error between the real and the estimated eigenvalues is given by:

$$\mathcal{X}((\alpha, \beta), (\alpha', \beta') \quad \leq \quad \frac{\epsilon||(B, W)||_1}{S(\lambda)}, \tag{3}$$

where $S(\lambda)$ is called the *reciprocal condition number* of the eigenvalue λ. Small values of $S(\lambda)$ indicate ill-conditioned eigenvalues, since a small perturbation of the matrix pair (B, W) results in a large difference between the estimated and the real eigenvalues. Similar to eigenvalues, the asymptotic error bound and the reciprocal condition number can also be derived for eigenvectors. For details cf. [10].
We now present speech recognition experiments on the *EPPS* corpus to investigate the relationship between WER and asymptotic error bounds and reciprocal condition numbers delivered by the *dggevx* algorithm, respectively. We present the error bounds and the reciprocal condition numbers averaged over all eigenvalues respectively eigenvectors. Table 5 summarizes the results for combination of different sets of features. The first line gives the baseline results applying LDA on the MFCC feature only. In the next experiment, a singularity was introduced artificially by repeating the first MFCC coefficient to simulate a strongly correlated additional one-dimensional feature. Although the information contained in the acoustic features has not changed, WERs increased considerably. Simple methods, like explicitly excluding eigenvectors from the projection matrix which belong to low eigenvalues $(\alpha, \beta) < \mu$ has not improved the results. Increasing values of μ have been tested which lead to excluding an increasing amount of eigenvalues close to singularity. The best recognition result has been obtained by $\mu = 3 \times 10^{-8}$. The average condition numbers have dropped rather heavily, indicating weak estimates of the eigenvalues and eigenvectors, which might explain this degradation. Furthermore, the low average of the condition numbers indicates that a strong singularity effects not only the conditioning of the singular eigenvalue but also all the rest of the eigenvalue estimates. The third and fourth lines of the table show results of experiments using the combination of different features. As expected from the recognition performance, the experiment combining the MFCC, VTLN, V, and SD features has not resulted in large differences in condition numbers compared to the baseline experiment. Nevertheless, the conditioning of the eigenvectors decreased more strongly, which needs to be further investigated. Finally, we have calculated the average reciprocal condition numbers for the combination of the MFCC, MF-PLP, and PLP features. Our goal was to find an explanation for the unexpected increase in WER compared to using the single MFCC feature. Although the degradation in WER is comparable with the experiment repeating the first MFCC coefficient, the conditioning has not decreased as heavily as in the second line of the table. Further analysis of this problem is required to verify if the small reduction in the eigenvalue condition number and the increase in the asymptotic error bounds explain the degradation in WER. Table 6, summarizes results obtained by increasing the LDA window length i.e. the number of successive concatenated feature vectors. Firstly, the

Table 5: Average reciprocal condition numbers and asymptotic error bounds of eigenvalues and eigenvectors on the *EPPS* corpus obtained by LDA-based feature combination tests yielding improvements and degradations in WER.

acoustic features	WER [%] dev	WER [%] eval	#LDA input	avr. recip. cond. num. eigenvalue	avr. recip. cond. num. eigenvector	avr. asym. error bound eigenvalue	avr. asym. error bound eigenvector
MFCC	14.7	15.3	144	4.0	2.7×10^{-2}	2.5×10^{-13}	3.7×10^{-8}
MFCC+Repeated-1st-Coeff	15.6	16.2	153	1.7×10^{-8}	2.0×10^{-19}	8.3×10^{2}	$> \pi$
MFCC+VTLN+V+SD	13.8	14.1	306	1.0	5.2×10^{-4}	2.3×10^{-11}	1.0×10^{-4}
MFCC+MF-PLP+PLP	15.8	16.2	432	1.5×10^{-1}	6.8×10^{-4}	3.0×10^{-11}	2.5×10^{-3}

Table 6: Average reciprocal condition numbers (CN) and asymptotic error bounds (EB) of eigenvalues (EVL) and eigenvectors (EVC) on the *EPPS* corpus obtained by using increasing LDA window lengths.

input window length	WER [%] dev	WER [%] eval	LDA input dim.	CN EVL	CN EVC $[10^{-2}]$	EB EVL $[10^{-13}]$	EB EVC $[10^{-8}]$
5	15.5	16.8	80	4.3	3.4	1.4	0.25
7	15.0	15.3	112	4.2	3.2	2.0	0.99
9	14.7	15.3	144	4.0	2.7	2.5	3.7
11	15.0	15.5	176	4.0	2.1	3.5	12
13	15.1	15.6	208	4.1	2.2	3.9	20
17	15.4	15.8	272	3.8	2.1	20	1200

condition numbers do not change significantly when increasing the LDA window length. Although the asymptotic error bounds do show a tendency to increase, their relation to WER is not obvious. At hardly changing condition numbers, the increasing error bounds must be caused by the increasing matrix norm $||(B, W)||_1$ which is most probably caused by the increasing dimension. Therefore, the increasing size of the scatter matrices seems not to lead to ill-conditioned eigenvalue problems. For an explanation of the degradation also the relation between the number of input features and the amount of training data available might have to be considered.

6. Summary

The results presented for LDA-based combination of multiple acoustic features show improvements in WER of up to 5% relative to the best single-feature system. Yet, in some cases LDA-based feature combination leads to unexpected degradations in WER. Experiments with additional random components indicate that the LDA input dimension is not the bottleneck. Adding about 1000 independent random features did not alter the WER. On the other hand, adding strongly correlated acoustic features lead to degradations in WER due to unstable estimates of the projection matrix. The stability of the numerical estimation of LDA was analyzed by means of perturbation theory. Nevertheless, degradations in WER when increasing the LDA window length and when combining MFCC, MF-PLP, and PLP features could not be explained by this analysis. Therefore, careful preselection of features to be combined still is necessary. Consequently, in future work algorithms specifically developed for singular pencils [11, 12] will be considered to improve the stability of the eigenvalue and eigenvector estimates.

Acknowledgments This work was partly funded by the DFG (Deutsche Forschungsgemeinschaft) under the post graduate program "Software für Kommunikationssysteme" and by the European Commission under the project TC-Star (FP6-506738).

7. References

[1] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, San Francisco, CA, Mar. 1992, vol. 1, pp. 13 – 16.

[2] R. Haeb-Umbach and M. Loog, "An investigation of cepstral parameterisations for large vocabulary speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sept. 1999, vol. 3, pp. 1323 – 1326.

[3] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 457 – 460.

[4] L. Welling, *Merkmalsextraktion in Spracherkennungssystemen für grossen Wortschatz*, Ph.D. thesis, RWTH Aachen University, 1999.

[5] M. Katz, H-G. Meier, H. Dolfing, and D. Klakow, "Robustness of linear discriminant analysis in automatic speech recognition," in *Proc. Int. Conf. on Pattern Recognition*, Québec, Canada, Aug. 2002, vol. 3, pp. 30371 – 30374.

[6] D. Kocharov, A. Zolnay, R. Schlüter, and H. Ney, "Articulatory motivated acoustic features for speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, Lisboa, Portugal, Sept. 2005, vol. 2, pp. 1101 – 1104.

[7] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738 – 1752, June 1990.

[8] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numeric Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 2nd edition, 1992.

[9] E. Anderson, Z. Bai, J. Bishop, J. Demmel, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK User's Guide*, SIAM, Philadelphia, PA, 3rd edition, 1999.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, 3rd edition, 1996.

[11] J. Demmel and B. Kagström, "The generalized schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications. part i: Theory and algorithms.," *ACM Transactions on Mathematical Software*, vol. 19(2), pp. 160 – 174, 1993.

[12] B. N. Parlett and H. C. Chen, "Use of indefinite pencils for computing damped natural modes," *Journal of Linear Algebra and Its Applications*, pp. 140:53 – 88, 1990.