IDIAP RESEARCH REPORT

Discriminant linear processing of time-frequency plane

Fabio Valente [a] Hynek Hermansky [a c]

IDIAP–RR 06-20

June 2006

published in
ICSLP 2006

[a] IDIAP Research Institute, Martigny, Switzerland
[c] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-20

Discriminant linear processing of time-frequency plane

Fabio Valente Hynek Hermansky

June 2006

published in
ICSLP 2006

Abstract. Extending previous works done on considerably smaller data sets, the paper studies linear discriminant analysis of about 30 hours of phoneme-labeled speech data in the time-frequency domain. Analysis is carried both independently in time and frequency and jointly. Data driven spectral basis show similar frequency sensitivity as human hearing. LDA-derived temporal FIR filters are consistent with temporal lateral inhibition. Considerable improvement is obtained using first temporal discriminant.

1 Introduction

Data-driven design of analysis module in automatic recognizer of speech (ASR) has been shown to be powerful means for improvement of performance in a number of well accepted ASR tasks [1],[2]. It allows for extracting speech-specific knowledge from large amounts of labelled speech data and for efficient use of this knowledge on new ASR tasks [3]. While the most effective in its nonlinear form [4], its linear version represented by linear discriminant analysis (LDA) has a distinct advantage in data-driven feature module design since its output can be readily interpreted in terms of linear systems [5]. It has been used for design of spectral basis [6], FIR RASTA filters [11] and 2-D spectro-temporal basis [12],[18],[14]. The LDA-derived spectral basis yield spectral analysis with auditory-like spectral resolution and LDA-derived FIR band-pass filters with symmetrical (approximately zero-phase) impulse responses emphasize dominant speech components in the modulation spectrum of the signal. 2-D discriminants derived from a relatively short segments (about 100 ms) of speech in [12],[18] were found beneficial but their properties were not explicitly reported.

Informal evidence suggests that the 2-D LDA basis can be well approximated by an outer product of LDA-derived spectral basis and LD-derived FIR filters [21], [14]. Kajarekar et al. work is of a particular interest since they report forms of their 2-D discriminants, using longer speech segments (about 1000 ms) that were reported necessary for sufficient classification of phonemes in speech [15], and explicitly aimed at comparisons of 1-D and 2-D discriminants. However in [14], the authors had at their disposal only about 3 hours of phoneme-labeled data, and reported that use of longer speech segments could have created problems in 2-D LDA design since the necessary covariance matrixes are rather large. Further, their speech database (OGI Stories), though reasonably realistic, consists of monologues and does not fully represent true conversational speech.

As the popular wisdom goes: "There is no data like more data". We are convinced it is beneficial to revisit the LDA using large amounts of phoneme-labeled realistic conversational speech that became recently available from SRI International in order to understand if same conclusion holds across databases, and this is the topic of our paper. The paper is organized as follows: in section 2, we review LDA based techniques for degenerate matrix we use in experimental part, in section 3 we give a brief description of the database, in sections 4,5 we describe experiments in temporal, spectral and joint spectro-temporal domain and in section 6 we present ASR experiments.

2 Linear Discriminant Analysis

Given a data set $\{X_i^k\}$ where X_i^k is an n-dimensional feature vector that represents the ith sample of kth class, LDA aims at simultaneously minimizing the within-class covariance matrix S_b and the across-class covariance matrix S_w defined as:

$$S_b = \sum_{k}^{K}(\bar{X}^k - \bar{X})(\bar{X}^k - \bar{X})^T \quad S_w = \sum_{k=1}^{K}\sum_{i=1}^{M_k}(X_i^k - \bar{X}^k)(X_i^k - \bar{X}^k)^T \tag{1}$$

where K is the number of classes, M_k is the number of elements in the kth class, $\bar{X}^k$ is the class mean defined as $\bar{X}^k = 1/M_k \sum_{i=1}^{M_k} X_{i=1}^k$ and $\bar{X}$ is the global mean of the data $\bar{X} = 1/(\sum_{k=1}^{K} M_k)\sum_{k=1}^{K}\sum_{i=1}^{M_k} X_i^k$. LDA aims at finding a transformation $Y = AX$ such that the Fisher criterion defined in terms of S_b and S_w is maximized [7] i.e.

$$max_A(trace(A^T S_w A)^{-1}(A^T S_b A)). \tag{2}$$

Solving equation (2) is equivalent to the generalized eigenvalue problem $S_b x = \lambda S_w x$ for $\lambda \neq 0$ and solution can be obtained by applying eigen-decomposition to the matrix $S_w^{-1} S_b$ if S_w is not singular.

However in real situations, the matrix S_w can be singular. A common technique for overcoming this problem is the PCA-LDA in which the space is first smoothed preserving the principal components of total covariance matrix and then LDA is applied to non degenerate matrix (as in [17]). PCA may not be compatible with LDA; in [8] it is shown that the most important space for linear discrimination is the null space of matrix S_w[1]. In fact if $S_w A = 0$ and $S_b \neq 0$, the ratio (2) is maximized and perfect classification is achieved. However, use of PCA may directly eliminate the null space of S_w, eliminating the most discriminative information. A possible solution proposed in [9] computes first $V^T S_b V = \Lambda$, eliminating the null space of S_b (which is useless for discrimination) and projecting S_w in this space. In this way the null space of S_w is preserved together with most important discriminants (for details see [9]).

In real data situations, many eigenvalues are generally very small but different from zero ; this induces the problem of the number of components that should be preserved while smoothing S_b. We apply cross validation over a separate data set using as validation criterion the ratio (2) in order to decide the valid number of discriminants.

3 Database description

Experiments are run using 30 hours of speech obtained from the CTS (Conversational Telephone Speech) database. CTS database is a collection of narrowband speech data from many different previous databases (Switchboard, Fisher databases, etc.). Data are labeled into 40 phonetic classes and labels provided by SRI are automatically obtained using forced alignment. This amount of data is significantly larger than the amount of data previously used for this task, allowing robust estimation of Spectro-Temporal discriminants.

4 Data driven temporal and spectral discriminants

Temporal filters were introduced as means for alleviating effects of linear distortion of the signal [10]. In [11] temporal filters were derived using LDA on time trajectory of critical band energies; data driven discriminants had similar magnitude frequency response as RASTA filter [10]. LDA is applied here on vectors composed by a sequence of critical band energies with a total duration of 1010ms (101 frames) and labeled as the phonetic class in the center of the vector.

In this first set of experiments we apply LDA as described in section (2). Until now only first three discriminants have been studied; the use of large amount of data allows the robust estimation of higher discriminants. If data are split in sentences as it is usually done in speech recognition, discriminants exhibit artifacts at the end and at the begin, otherwise if they are processed with full context they show significant non-zero values only in the center. This suggest that the procedure of splitting the data in blocks may be detrimental for such temporal processing of data.

Figure 1 shows four discriminants obtained from the fifth critical band. Discriminants at other frequencies are very similar. Width of those filters suggests that information about phonemes is spread in time over an interval of around 500ms around the center of the phoneme. First discriminant is qualitatively similar to RASTA filter while higher order discriminants describe more details of signal dynamics. In frequency domain they corresponds to pass band filters that pass lower frequencies of the modulation spectrum. Width of temporal discriminants progressively increase, suggesting the use of different time resolutions. This result directly supports the intuition in [13] where a filter bank of multi-resolution RASTA filters is used for extracting temporal informations.

In [19],[17] spectral discriminants are derived using LDA, however PCA is used for smoothing S_b and S_w. According to the discussion of section 2 this is a suspect method that can significantly affect the result. We repeat the same experiment using 30 hours of speech and the previously discussed LDA technique for singular matrices. Hamming window shifted by 10ms step is used to obtain 129 points of 12th order LPC logarithmic power spectrum.

Cross validation experiments select out of the possible 39 discriminants only 25 (that are enough for covering the discriminative space). In figure 2 first four temporal discriminants are plotted on a linear

[1]The null space of a matrix S is defined as $\{x|Sx = 0, x\epsilon R^n\}$

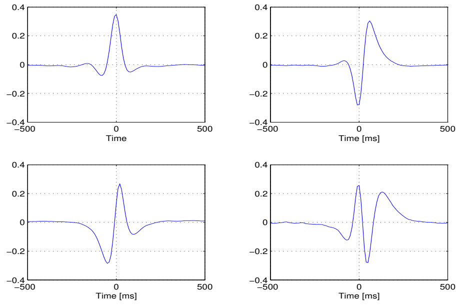

Figure 1: Four discriminants obtained using LDA on Temporal Patterns for the fifth critical band

scale. Linear discriminants show a higher oscillation frequency at low frequency and progressively lower oscillation frequency in the higher part of the spectrogram, suggesting different frequency resolution at different parts of the spectrum.

To further investigate this issue we performed sensitivity analysis as described in [19]. Sensitivity of a given bases is computed as the Euclidean distance between a gaussian shape centered at a given frequency and the same shape shifted by a certain value, projected on the bases. In other words if $g(f)$ is a gaussian shape centered at frequency f and W is the LDA basis, the sensitivity $S(f)$ is defined as $S(f) = ||g(f) \cdot W - g(f+\mu) \cdot W||$ where μ is a shift. Figure 3 (up) plots sensitivity to a constant shift $\mu = 25Hz$ on a linear scale; in this case LDA basis are more sensitive at lower frequencies. Figure 3 (down) plots sensitivity to a constant shift $\mu = 0.8$ Bark on a Bark scale; sensitivity is now constant, suggesting that LDA discriminants emulate the Bark scale with higher resolution at low frequencies and lower resolution at higher frequencies. The Bark scale was derived by perceptual experiments, LDA spectral basis are completely data-driven supporting [19].

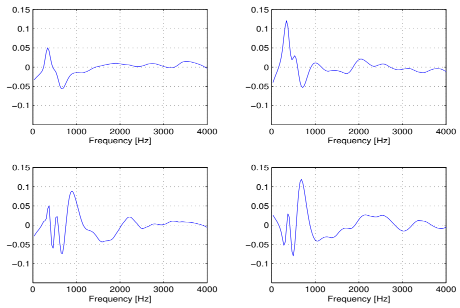

Figure 2: Four spectral discriminants on a linear scale.

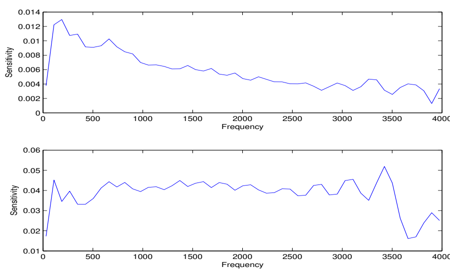

Figure 3: Sensitivity of LDA linear basis versus constant shift on a linear scale (up) and on a bark scale (down).

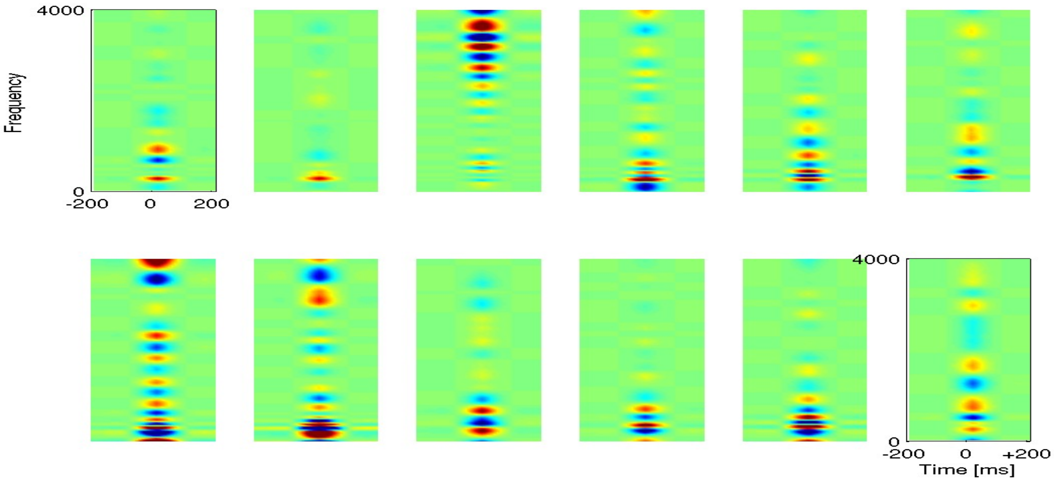

Figure 4: First twelve spectro temporal discriminant

5 Joint spectro-temporal analysis

Analysis of spectrum and of temporal trajectories has been done in previous sections independently. We want to investigate now discrimination in joint spectro-temporal domain using discriminant analysis. A $T \times F$ matrix where T is the temporal context and F is the number of frequency components is labeled according to the phoneme in its center. The matrix can be represented as a vector of size $T \times F$ and classical LDA can be applied. This approach was applied in [14] in critical band domain with a context of 101 frames on small amount of data. Conclusion was that the amount of data was not sufficient for robust estimation of discriminants. Better results were obtained if analysis was carried independently in time and in frequency and discriminants recombined. We are interested here in using the LPC power spectrum (dimension 129 points) in combination with a temporal context of 101 frames. If matrices are represented as vectors, S_w and S_b have a dimension of 13000 $\times$ 13000 which is unsuitable for computational reasons. A way to overcome this dimension problem is doing discriminant analysis directly in the matrix space represented as a tensor. In other words, operations on vectors are replaced by operations on tensors and final discriminant space is a tensorial space. If X is a matrix of dimension $T \times F$, we seek the space transformation that reduces $T \times F$ into a space of dimensions $l_1 \times l_2$; this space is obtained by the tensor product of a subspace L of dimension $T \times l_1$ and R of dimension $F \times l_2$. Projection of an element X in this space is given by the product $Y = L^T X R$ with final dimension $l_1 \times l_2$. In the tensorial space, the Frobenious norm can be used to derive within and across class matrices S_w and S_b defined as:

$$S_w = \sum_{i=1}^{k} \sum_{X \epsilon C_i} ||X - M_i||_F^2 \quad S_b = \sum_{i=1}^{k} n_i ||M_i - M||_F^2 \tag{3}$$

where M_i is the class mean matrix and M is the global mean matrix. Using the Frobenious norm property $trace(MM^T) = ||M||_F^2$ and applying the transform $Y = L^T X R$ expression (3) reduces to:

$$\bar{S}_w = trace(\sum_{i=1}^{k} \sum_{X \epsilon C_i} L^T (X - M_i) R R^T (X - M_i)^T L) \tag{4}$$

$$\bar{S}_b = trace(\sum_{i=1}^{k} n_i L^T (M_i - M) R R^T (M_i - M)^T L). \tag{5}$$

Optimal transforms L and R can be found iteratively fixing one of them, projecting tensor in their space and solving the generalized eigenvalue problem (see [16]). This result in eigen-decomposition of matrix 101×101 and 129×129 instead of 13000×13000. In other words rows are projected on matrix RR^T and columns on matrix $L^T L$ that span the linear discriminant space of rows and columns of X respectively. Linear discriminants obtained using 2DLDA have similar shape as outer product of discriminants obtained processing independently the temporal and spectral domain, suggesting that those domains can be processed independently[2]. Figure 4 shows twelve 2D discriminants in the time frequency domain. Some of them show strong localization properties both in time and in spectral domains as if they were sensitive to a particular region in the plane; on the other hand we can notice as well discriminants with a sensitivity more spread over the frequency domain.

6 Experiments

In this section we describe results obtained running recognition experiments using spectral and temporal discriminants. LPC power spectrum is projected on spectral and temporal basis and results are compared with LPC power spectrum projected on DCT basis and with PLP cepstral coefficients. For experiments we used a database that is different from the one used for deriving linear discriminants assuming that those findings are universal properties of speech and not task dependent. Recognition results are run on the OGI-digits database. Table 1 shows results obtained using the following set of 13 features: (a) LPC power spectrum projected on 13 DCT basis (b) PLP (c) LPC power spectrum projected on 13 spectral linear discriminants (d) LPC power spectrum projected on 13 spectral linear discriminants and filtered with one temporal discriminant. If LDA basis are used instead of

13 LPCC	13 PLP	13 spec.	13 spec. × 1 temp.
85.9	86.5	86.5	90.9

Table 1: Accuracy for different sets of 13 features on OGI-digits

DCT basis, an improvement of 4% (relative) is obtained. DCT basis has a uniform spectral sensitivity while LDA has a higher sensitivity at lower frequencies (emulating somehow the bark scale) where the most important information for recognition is contained. Spectral basis designed from data yield similar performance as PLP features designed according to auditory principles [10]. If spectral features are filtered with first temporal discriminant a considerable improvement of 35% (relative) w.r.t. the LPC baseline and 32% (relative) over PLP is obtained indicating the effectiveness of larger temporal context, imposed by the temporal filtering. In table 2 we compare results for 39 features, i.e. LPCC features plus delta and double delta with power spectrum projected on 13 spectral basis and 3 temporal basis. In this case LDA spectral and temporal discriminants outperform LPCC plus delta

39 LPCC + deriv.	39 PLP + deriv.	13 spect. × 3 temp.
94.0	94.6	94.7

Table 2: Accuracy for different sets of 39 features on OGI-digits

and double delta by 11% (relative) while only very small improvement w.r.t. PLP and derivatives is found. Again data guided features yield equivalent results as currently often used PLP static and dynamic features.

[2]This is similar to the Combined Discriminant Analysis proposed in [14] but here final estimation is achieved trough an iterative algorithm that converges to two discriminant subspaces

7 Conclusions

In this work we have extended previous LDA analysis of spectro-temporal domain done on smaller data sets. A tensorial LDA is proposed for processing long time-frequency slices and a revisited LDA is used for dealing with singular covariance matrices. Temporal basis have similar magnitude frequency characteristic as RASTA filters but differ in phase, spectral bases have similar frequency sensitivity as the Bark scale of hearing and obtained 2D filters show localization properties both in time and frequency. Those conclusions are qualitatively consistent with what was presented earlier in literature [6],[11],[14] on smaller databases. We found a large improvement in the use of data driven front-end when only 13 features are used. In this case the most important gain in performances is obtained when time trajectories are filtered with first temporal discriminant. On the other side only small improvements are obtained when dynamic features are added. The fact that results were carried over different databases supports the universal (speech specific and not task specific) nature of our findings.

8 Acknoledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 and by the EU under the grant DIRAC IST 027787. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Authors would like to thank Andreas Stolcke for providing data labels and Petr Fousek for his help.

References

[1] A. Adami and L. Burget and S. Dupont and H. Garudadri and F. Grezl and H. Hermansky and P. Jain, S. Kajarekar and N. Morgan and S. Sivadas: QUALCOMM-ICSI-OGI Features for ASR, Proc. ICSLP 2002, Denver, Colorado, USA, Sep, (2002).

[2] Q Zhu, A Stolcke, BY Chen, N Morgan - Using MLP Features in SRI Conversational Speech Recognition System Proc. Interspeech, Lisbon (2005).

[3] Sunil Sivadas and Hynek Hermansky: On Use of Task Independent Training Data in Tandem Feature Extraction, Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-04, 2004.

[4] H. Hermansky and D.P.W. Ellis and S. Sharma, "Connectionist Feature Extraction for Conventional HMM Systems", in ICASSP'00, Istanbul, Turkey, 2000.

[5] Hermansky: Human Speech Perception: Some Lessons from Automatic Speech Recognition, Text, Speech and Dialogue 2001, Matousek et al. Eds. Springer 2001.

[6] H. Hermansky and N. Malayath, "Spectral Basis Functions from Discriminant Analysis", in ICSLP'98, Sydney, Australia, 1998.

[7] Fukunaga K.: Introduction to Statistical Pattern Recognition. Academic Press, New York, 1991.

[8] Chen L., Liao H.,Ko M.,Lin J. and Yu G.: A new lda-based face recognition system which can solve the small sample size problem. Pattern Recognition, 33(10):1713-1726, Oct 2000

[9] H. Yu and J. Yang: A direct LDA algorithm for high-dimensional data with application to face recognition, Pattern Recognition, vol. 34, pp. 2067-2070, 2001.

[10] Hermansky H. and Morgan N.: RASTA processing of speech. IEEE Trans. on Speech and Audio Processing, vol 2, num 4, pp 578-589, Oct 1994.

[11] van Vuren S., Hermansky H.: Data-Driven Design of RASTA-Like Filters, Proc. of Eurospeech 87, Rhodes, Greece, 1997.

[12] Peter F. Brown. The Acoustic-Modelling Problem in Automatic Speech Recognition. PhD thesis, School of Computer Science, Carnegie Mellon University, 1987.

[13] Hermansky H. and Fousek P.: Multi-resolution RASTA filtering for TANDEM-based ASR, Proc. of Interspeech 2005, Lisboa, 2005.

[14] Kajarekar A., Yegnanarayana B. and Hermansky H., "A Study of Two Dimensional Linear Discriminants for ASR", Proc. of ICASSP'01, Salt Lake City, Utah, USA, May, 2001

[15] Kozhevnikov and L. Chistovich, "Speech, Articulation and Perception", translated by Joint Publications Research Service, Washington, 1965

[16] Jieping Y., Ravi J. and Qi L.: Two-Dimensional Linear Discriminant Analysis, Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA, pp. 1569-1576, 2005.

[17] Hermansky H, and Malayath N. : "Spectral Basis Functions from Discriminant Analysis", Proc. of ICSLP'98, Sydney, Australia, 1998

[18] Haeb-Umbach, R., and Ney, H. :"Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition, in Proc. ICASSP 1992, San Francisco, CA, March 1992, pp I.13-I.16.

[19] Malayath, "Data-Driven Methods for Extracting Features from Speech", OGI, Portland, USA, Jan, 2000.

[20] Fletcher H., "Speech and hearing in communication" Van Nostrand 1953.

[21] Picheny M., Personal communications 2000.