A STUDY OF TWO DIMENSIONAL LINEAR DISCRIMINANTS FOR ASR

Sachin S. Kajarekar[1], *B. Yegnanarayana*[3] *and Hynek Hermansky*[1,2]

[1] Oregon Graduate Institute of Science and Technology , Oregon, USA
[2] International Computer Science Institute Berkeley, CA, USA
[3] Indian Institute of Technology, Department of Computer Science and Engineering, Madras, India

ABSTRACT

In this paper we study the information in the joint time-frequency domain using 1515 dimensional - 15 spectral energies and temporal span of 1 s - block of spectrogram as features. In this feature space, we first derive 20 joint linear discriminants (JLDs) using linear discriminant analysis (LDA). Using principal component analysis (PCA), we conclude that information in this block of the spectrogram can be analyzed independently across the time and frequency domains. Under this assumption, we propose a sequential design of two dimensional discriminants (CLDs), i.e., spectral discriminants followed by temporal discriminants. We show that these CLDs are similar to first few JLDs and the discriminant features derived from the CLDs outperform those obtained from JLDs in the continuous-digit recognition task.

1. INTRODUCTION

For speech recognition task, linear discriminant analysis (LDA) [1] has been used to discriminate between phones (as classes) and linear discriminants (LDs) have been obtained in the spectral, temporal and joint time-frequency domains. In spectral domain the discriminants were designed using short-term spectral energies as features[2, 3] and they were interpreted as spectral basis functions. These bases had decreasing resolution along the frequency axis which is similar to the widely used MEL/BARK warping[3].

In temporal domain, the discriminants were designed using time trajectories of the spectral energies as features, and they were interpreted as FIR RASTA filters[4]. These filters were similar across different across different frequency bands and higher discriminants were approximately derivatives of the lower discriminants. When obtained from the super-frames formed by concatenating several short-term spectral energy frames, discriminants were viewed as two-dimensional bases for projecting a block of speech spectrogram[5, 6, 7, 8].

The previous work on the joint discriminants has been limited to the temporal span of approximately 100 ms. We have, however, observed that information about current frame of speech spreads for approximately 250 ms around it [9]. In this work, we have studied discriminants in time-frequency domain with a temporal span of around 1 s. In the first approach, 1515 dimensional - 15 spectral energies and temporal span of 1 s - block of spectrogram has been used as feature vector. LDA was, then, used to obtain 20 linear discriminants (referred as JLDs) in this feature space and the discriminant features obtained from these LDs were evaluated on the continuous digit recognition task.

In the second approach, we have questioned the advantage of two dimensional discriminants over the one dimensional discriminants[1]. Conventionally, the joint discriminants are argued to benefit from the correlations across the time and frequency domains. In this work, we have used principal component analysis (PCA) to study the nature of information in the block of spectrogram. From the principal components, we have inferred that information in the spectrogram, as a first approximation, can be analyzed independently across time and frequency domains. Using this hypothesis, we have proposed an alternative scheme for obtaining the joint discriminants in the time-frequency domain. We show that 20 LDs obtained by combining the spectral and temporal LDs (referred as CLDs) outperform the 20 JLDs on the continuous digit recognition task.

The paper is organized as follows. Section 2 gives a brief overview of the statistical techniques used in this paper. Section 3 describes the experimental setup and the method of combining the one dimensional LDs. Finally, section 4 gives the results and conclusions from this work.

2. PCA AND LDA

This section gives a brief overview of the two statistical techniques used in this work. Interested reader can refer to [1] for a comprehensive description.

Principal component analysis (PCA) and linear discriminant analysis (LDA) are the statistical techniques used for feature selection, i.e., for finding the subspace in the original feature space that contains the most relevant information. PCA preserves the total energy of the original feature space. It assumes that the data has a unimodal Gaussian distribution. Hence the optimal n-dimensional subspace (such that $n < m$=dimension of the original subspace) using PCA is spanned by n leading eigenvectors. We refer to these eigenvectors as principal components (PCs).

LDA preserves the discriminant information of the original feature space. It assumes that the distribution of the features contains *n* classes such that the total variance is divided into 1) across-class variance (AC, space spanned by the means of the classes) and 2) within-class variance (WC, average variance within each class). The optimal subspace using LDA for *n* classes is spanned by *(n-1)* eigenvectors. These eigenvectors are obtained by solving generalize eigenvalue solution using AC and WC. We refer to these eigenvectors as linear discriminants (LDs).

In both these cases the new set of features (referred as discriminant features) are obtained by projecting the original features on the eigenvectors of the new subspace. Note that the eigenvectors from LDA are obtained by inverting the WC matrix and a well-conditioned WC matrix is needed for the proper solution. Due

[1] note that the relationship between one dimensional LDs and two dimensional LDs has not been studied yet

to high dimensional features (1515) used in our analysis, a large amount of data is needed for a proper conditioning of WC. We have used approximately 1 million frames for our analysis. In spite of that JLDs were noisy. This has adversely affected the recognition performance of the resulting discriminant features.

3. EXPERIMENTAL SETUP

For PCA and LDA we have used approximately 3 hours of hand-labeled speech data from the English part of the OGI Stories [10] database. The features were a set of logarithmic energies from 15 critical-band (BARK) filters that was estimated using 25 ms hamming window at 100 Hz frame rate. LDs were evaluated on continuous digit recognition task: vocabulary = 11 digits (0-9 and "oh"). These digits were modeled as a sequence of context independent monophones. We have used 23 monophones in the recognition experiments. Consequently these 23 monophones were used as classes in LDA.

The digit recognition experiments were performed on an independent database – part of OGI NUMBERS Database. The 23 context independent monophones were modeled using 5-state, 3-component HMMs. Approximately 2500 utterances were used for training HMMs and 12000 digits were used for testing.

3.1. Joint Linear Discriminants (JLDs)

For joint spectro-temporal LDA, a 1515 dimensional super-frame obtained from $15 * 101$ dimensional block of spectrogram was used as a feature vector. This block contained the phonetic context of 500 ms in the past and the future. The center frame in the block was referred as the current frame ($t = 0$) and the block was labeled by the phone label of center frame.

We used 20 leading eigenvectors as the JLDs for the recognition experiments. Each $15 * 101$ dimensional block of speech spectrogram was projected on these JLDs to obtain 20 discriminant features.

3.2. Combined Linear Discriminants (CLDs)

The discriminants in the time-frequency domain can also be designed using an alternative method that combines the spectral and temporal LDs. This method is explained as follows. First we obtain 8 linear discriminants in the spectral domain using spectral features described above. Then filter-bank energies are projected on these discriminants to generate a new 8 dimensional feature vector. Further, discriminants are designed on the temporal streams of these new features, only for the first stream. We use 101 dimensional vector from this stream for the LDA where each vector is labeled by the phone label of the center point. First three eigenvectors (or temporal discriminants) are used as the FIR RASTA filters. They are used to filter 8 spectral streams that results in 24 dimensional feature vector. This combination of the spectral and temporal LDs is referred as combined linear discriminants (CLDs).

The first 20 discriminant features from the CLDs were compared with the 20 discriminant features from JLDs on continuous digit recognition task. Note that the CLDs can also be obtained by deriving temporal LDs; using these discriminants to filter the spectral trajectories; and deriving spectral LDs from the resulting spectral vectors (referred as TSLDs, temporal followed by spectral LDs). On the continuous digit recognition tasks, we observed that the spectral LDs followed by temporal LDs (STLDs, referred in this paper as CLDs) outperform the TSLDs when the convolutive noise dominates the unseen noise (present task). However, in presence of colored additive noise - different noise characteristics in different spectral bands - the TSLDs were observed to outperform the STLDs.

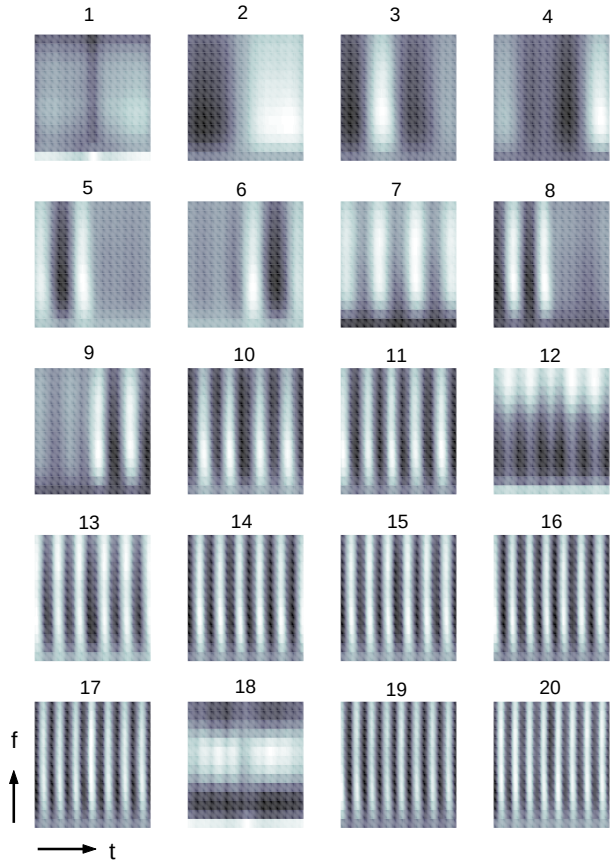

Fig. 1. The First 20 principal components from the joint time-frequency domain. Note that the PCs show activity either across time or frequency

4. RESULTS AND CONCLUSIONS

4.1. Principal Components

In this work, we have used PCA to study the structure of the information in time-frequency domain. Using principal components we have commented on the correlations across time and frequency domains.

Figure 1 shows the principal components (PCs) of feature space where each feature vector represents a $15*101 = 1515$ dimensional block of spectrogram. Figure 3 shows percentage of total variability captured by first n eigenvectors and Figure 2 shows the variance of each eigenvector across time and frequency. We observed that the first 20 principal components represented approximately 85 % of the total variability. The first PC was almost constant in the time-frequency plane and it represented approximately 35 % of the total variability. Thus the variation in the global mean of the spectrogram was the most significant source of variation. The 16 out of next 19

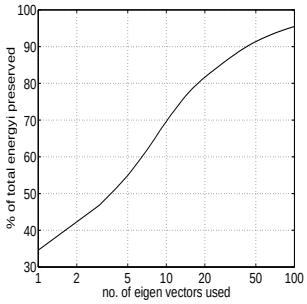

Fig. 2. % of total variance captured by the first n eigenvectors

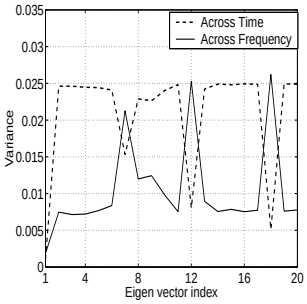

Fig. 3. Variance of the PCs along time and frequency axes

PCs were cosine-like basis across time and they are almost constant across different frequency bands. The 12th and 18th bases showed the variation across frequency and they were almost constant across time. The 7th PC was the only one that showed variation across both time and frequency.

These results showed that, in the $15 * 101$ dimensional block of spectrogram, the variation across time is dominant over variation across frequency. Further, it also showed that approximately 85% of the variation in the $15 * 101$ dimensional block of spectrogram can be explained by the bases that varied only across either time or frequency. Therefore it was concluded that the variability in speech spectrogram can be considered independently across time or frequency domain in the first approximation.

4.2. Linear Discriminants

Figure 4 and 5 show the first 8 spectral discriminants and the 3 temporal discriminants derived from the first spectral discriminant feature. For the description on the spectral discriminants refer to [3]. The temporal discriminants from the first spectral discriminant feature were symmetric or anti-symmetric band pass filters. These filters attenuated the 0 Hz frequency component (DC) and frequency components higher than 20 Hz. Note that the second and the third temporal LDs can be approximated by the derivative and the double-derivative of the first LD.

Figure 6 shows the joint discriminants and compares them to those obtained by combining the spectral and temporal discriminants. We observed that the first 4 joint discriminants can be approximated by combining the first spectral discriminant with the first 4 temporal discriminants, assuming that temporal discriminants are similar across different frequency bands. Higher JLDs are noisier and can not be compared to the CLDs. The noise in the JLDs was attributed to the insufficient data for estimating the $1515 * 1515$ dimensional within-class covariance matrix. We also observed that the JLDs were localized in time, i.e., unlike the PCs, their activity was limited to approximately 250 ms around the center frame.

4.3. Digit Recognition Results

Table 1 compares the digit recognition performance of different features. It shows that the JLDs (row 1) have performed worse than the CLDs (row 6). We hypothesized that the joint discriminants have performed worse because 1) they could not independently remove the mean from each filter-bank trajectory, and 2) they were noisy.

The convolutive (communication channel) noise in log spectral domain has been shown to affect mean of different frequency bands

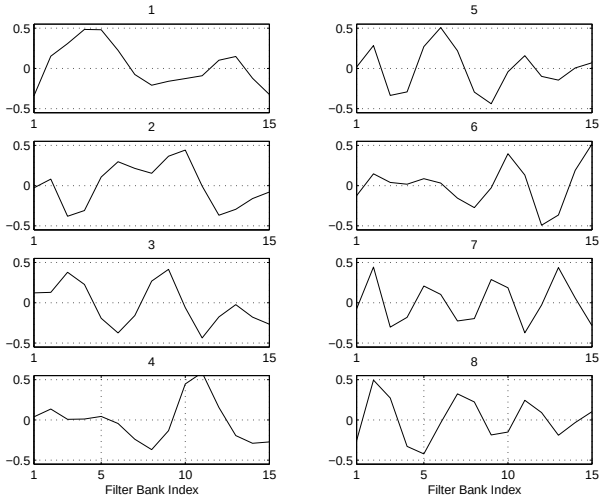

Fig. 4. The First 8 LDs from spectral domain

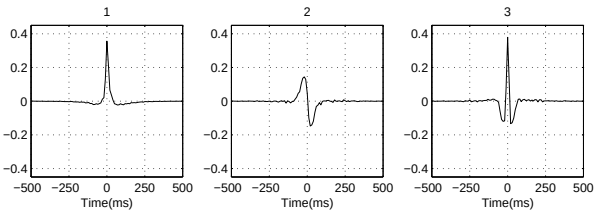

Fig. 5. The 3 LDs from temporal trajectory of first spectral discriminant feature

differently [9]. A widely used method to alleviate this noise removes the mean of each filter-bank over the utterance (also refereed as cepstral mean subtraction (CMS)). The JLDs, however, did not suppress the mean of each filter-bank trajectory separately. Instead they removed the mean over the block of the spectrogram which is a sub-optimal solution. To address this problem, the features were first preprocessed using CMS and the JLDs were derived on that data. The discriminant features from these JLDs significantly ($\alpha = 5\%$) improved the recognition performance (row 2 of Table 1) over those obtained from the JLDs without CMS (row 1 of Table 2).

The second problem, i.e., data scarcity, was addressed by reducing the time window of the features[2] after CMS from 1s (101 frames) to 0.09s (9 frames). It was observed that 20 discriminant features derived from the $15 * 51 = 765$ dimensional JLDs (row 3) outperformed those derived from the $15 * 101 = 1515$ dimensional JLDs (row 2). The 20 discriminant features from $15 * 25 = 375$ dimensional JLDs (row 4), performed comparable to the former and discriminant features from $15 * 9 = 135$ dimensional JLDs performed the best among all (row 5).

4.4. Conclusion

In this paper we have compared two approaches for obtaining discriminants in time-frequency domain. In first approach JLDs were

[2] a block of spectrogram

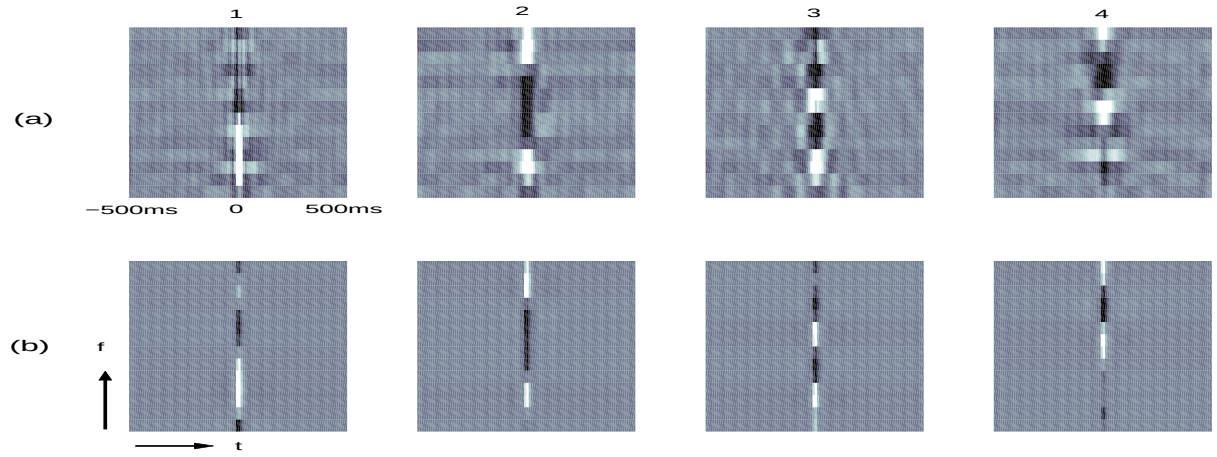

Fig. 6. (a) The First 4 linear discriminants from the data and (2) the 4 discriminants by combining the spectral and temporal discriminants

obtained using 1515 dimensional block of spectrogram as features. In the second approach spectral discriminants were obtained first and the temporal discriminants were derived after projecting the spectral energies on the spectral bases (CLDs). The discriminant features from JLDs outperformed those obtained from JLDs in the continuous digit recognition experiments. This was attributed to the fact that 1) JLDs did not remove the mean from each filterbank trajectory independently and 2) JLDs were noisy due to lack of sufficient data for their estimation. To alleviate these shortcomings, the data was preprocessed using CMS and the temporal window used for the estimation for JLDs was reduced. This improved the performance of the JLDs. However it did not show any significant improvement over the CLDs.

We also investigated into the nature of information in the 1515 dimensional feature space spanned by 15 spectral energies and 50 frames of context in the past and the future using PCA. The principal components corresponding to approximately 85 % of total variability spanned either the time or frequency domains independently. This showed that the variability in speech can be considered independently across time and frequency domains. It also supported the design of CLDs and the performance improvement obtained using CLDs. Thus we can conclude that independent spectral and temporal LDs might be the most practical solution for the connected digit recognition task.

5. ACKNOWLEDGMENT

The research was supported by NSA/DoD under MDA904-00-C-2089, NSF under IRI-9712579 and by an industrial grant from Intel Corporation.

6. REFERENCES

[1] Keinosuke Fukunaga, *Statistical Pattern Recognition*, Academic Press, San Diego, 1990.

[2] M. Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *ICASSP'89*, Glasgow, Scotland, 1989, IEEE, pp. 262–265.

[3] Hynek Hermansky and Narendranath Malayath, "Spectral basis functions from discriminant analysis," in *Proc. of ICSLP*, Sydney, 1998.

[4] C. Avendano, S. van Vuuren and H. Hermansky, "Data-Based RASTA-Like Filter Design for Channel Normalization in ASR," in *ICSLP'96*, Philadelphia, PA, USA, Oct. 1996, vol. 4, pp. 2087–2090.

[5] P. Brown, *The Acoustic-Modelling Problem in Automatic Speech Recognition*, Ph.D. thesis, Computer Science Department, Carnegie Mellon University, 1987.

[6] L. Bahl et. al., "Robust methods for using contex-dependent features and models in a continuous speech recognizer," in *Proc. of ICASSP*, 1994, pp. 533–536.

[7] P. McMohan et. al., "Discriminative spectral-temporal multiresolution features for speech recognition," in *Proc. of ICSLP*, Phoenix, Arizona, 1999.

[8] Tsuneo Nitta, "A novel feature-extraction for speech recognition based on multiple acoustic-feature planes," in *Proc. of ICASSP*, 1998, pp. 29–32.

[9] Sachin Kajarekar, Naren Malayath and Hynek Hermansky, "Analysis of sources of variability in speech," in *Proc. of EUROSPEECH*, Budapest, Hungary, 1999, pp. 343–346.

[10] R. Cole, M. Noel, and T. Lander, "Telephone speech corpus development at CSLU," *Proc. ICSLP 94*, September 1994.

Table 1. Connected digit recognition performance using 24 discriminant features

Original Dimension	LD type	% error
15*101	JLD without MS	11.5
15*101	JLD with MS	10.0
15*51	JLD with MS	7.5
15*25	JLD with MS	8.0
15*9	JLD with MS	6.8
15*101	CLD without MS	5.8