SPECTRAL BASIS FUNCTIONS FROM DISCRIMINANT ANALYSIS

Hynek Hermansky[1,2] *and Narendranath Malayath*[1]

[1]Oregon Graduate Institute of Science and Technology,
Portland, Oregon , USA.
[2]International Computer Science Institute,
Berkeley, California, USA.
Email: hynek,naren@ece.ogi.edu

ABSTRACT

The work examines Karhunen-Loeve Transform and Linear Discriminant Analysis as means for designing optimized spectral bases for the projection of the critical-band auditory-like spectrum.

1. INTRODUCTION

1.1. The state-of-art

Typical large vocabulary automatic recognition of speech (ASR) consists of three main components: feature extraction, pattern classification, and language modeling. The feature extraction attempts to reduce the information rate of raw speech data by alleviating irrelevant variability such as speaker characteristics or environmental noise, the pattern classification further reduces information rate by classifying each time instant into one of (phoneme-like) subword-unit classes, and language modeling compensates for possible errors of classification by emphasizing more likely word combinations.

Over the past two decades we witnessed the introduction of stochastic approaches in both the pattern classification and the language modeling modules. Stochastic techniques typically use only minimal a priori assumptions about the nature of the problem and derive their structure mostly directly from the data. Replacing the hardwired prior knowledge by the knowledge derived from the data turned out to be one of most significant advances in ASR research.

1.2. Motivation for the current work

Data-driven approaches are still largely absent in the analysis module. Only recently, some emerging efforts in deriving temporal RASTA processing in analysis from the data [2, 11, 5] started to appear. The current work attempts to extend such data-driven techniques into optimization of spectral bases in speech analysis.

The analysis module in ASR typically consists of a series of processing steps, some of which are inherited from speech coding, and some justified by perceptual or pattern matching arguments. A currently dominant speech representation is the auditory-like cepstrum [10, 4]. This cepstrum represents an appropriately modified (through auditory-like frequency and amplitude warping and critical-band smoothing) short-term spectrum of speech, projected onto the cosine basis (see Fig. 1). The short-term spectrum is derived from about 20 ms long consecutive segments of the speech signal. The spectral modifications are justified by properties of human hearing [4], and the cosine projection by the need for partial decorrelation of features [9] used in the subsequent pattern classification.

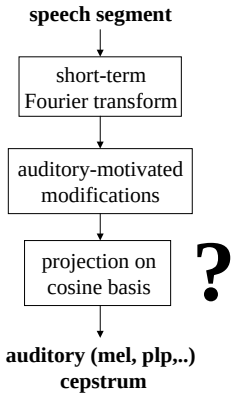

Figure 1: Generic form of dominant speech representation in ASR. The short-term speech spectrum is modified by auditory-motivated processing. This may include warping along its frequency axis by $log(const + x)$-like nonlinearity and smoothing. Modifications along the amplitude axis may include some form of emphasis of higher frequencies (6dB/oct preemphasis, simulated equal loudness curve,..) and logarithmic warping. The modified spectrum is then projected on cosine basis.

This work investigates the suitability of the cepstral representation and attempts to derive an alternative bases for projection of the auditory-like critical-band spectrum. The technique we use is based on analyzing the variance of about 2 hours from the OGI Stories database [1]. Stop-consonants were excluded from the basis derivation reported in this paper. However, all classes were evaluated in the phoneme classification experiment reported in the Section 3.

[1]This portion of OGI Stories database consists of phoneme hand-labeled fluent telephone-quality speech from 208 adults of both genders, each asked to speak on an arbitrary topic for about one minute.

2. DATA-DRIVEN SPEECH ANALYSIS

2.1. Linear discrimination analysis

In the past few years we have been experimenting with ways of utilizing large amounts of speech data for improving the speech feature extraction module. The main tool we use is the linear discrimination analysis (LDA) technique. LDA is a well known technique which attempts to find a linear transformation of the feature space which would optimize linear separability of classes. Typically, only a few eigenvectors of the transformation matrix are of interest [2].

LDA is not new to speech processing. To our knowledge, its use has been first studied by Hunt [6] who later used it for discriminative dimensionality reduction [7]. Brown [3] was first to apply LDA to several concatenated feature vectors, thus addressing both temporal and spectral dimensions. Earlier, we have used LDA for deriving temporal RASTA filters for processing time trajectories of critical-band spectral energies [2, 11, 5]. In our present work we use LDA to derive spectral weighting functions (spectral basis) as an alternative to the cosine basis in the conventional Mel cepstral analysis.

2.2. Spectral basis from KLT

In the past we have used data-driven techniques for designing temporal RASTA filters for enhancement of noisy speech [1] and for robust ASR [2, 11]. In the current work we attempt to derive optimized spectral bases functions for ASR.

The projection onto a cosine basis approximately decorrelates the spectral vector space [9]. This is illustrated in Fig. 2. The proper way to decorrelate the vector space is through the data-dependent Karhunen-Loeve transform (KLT). This is illustrated in Fig. 3 which shows the first six eigenvectors of the covariance matrix of the 14-dimensional critical-band spectral space (the first six elements of the KLT basis) derived from the 2 hours of OGI Stories database (stops excluded).

The basis vectors are reminiscent of cosine functions of the Mel cepstral analysis with the first vector evaluating the spectral energy and the consecutive higher ones are sensitive to cosine spectral ripples with decreasing period. The KL transformed covariance matrix is of course diagonal.

2.3. Spectral basis from LDA

The KL transform projects on the directions of maximum variability. As known, there are many sources of variability in speech, many of them harmful for phonetic classification [5]. It may be better to project the space on the direction of maximum separability rather than on the direction of maximum variance.

The transform which projects on the directions of maximum separability is the LDA. The basis derived by the LDA transform may be different from the basis derived by

[2] The matrix is of rank (N-1) where N is a number of classes in the classification problem.

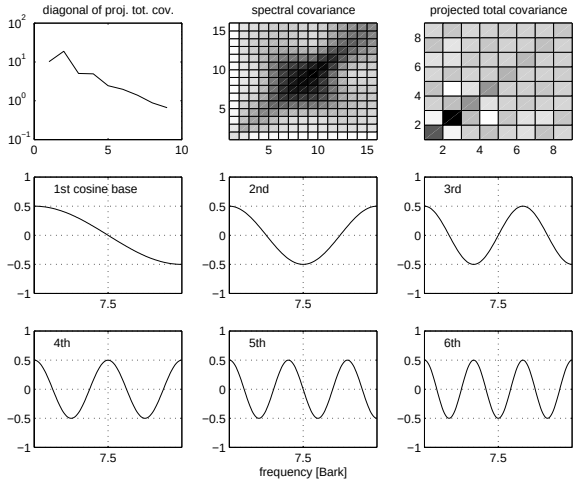

Figure 2: Upper left: Diagonal of the total covariance matrix projected on the cosine basis. Upper center: Covariance matrix of the original critical-band auditory spectral space derived from about 2 hours of hand-segmented OGI Stories database. As seen, the spectral covariance matrix is far from diagonal. Upper right: The total spectral covariance matrix, projected on the first 8 vectors of the cosine basis, is partially diagonalized. The figure also shows below the correlation matrixes the first 6 cosine basis functions (the zeroth base function, i.e. constant evaluating the energy of the spectrum is excluded) used in the ceptral projection.

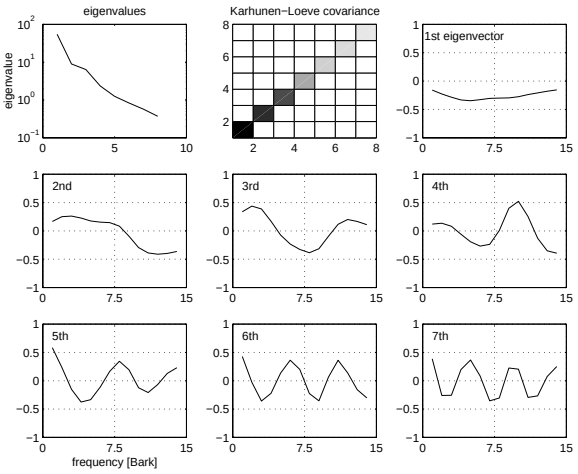

Figure 3: Upper left: Eigenvalues of the KLT basis. Upper center:The total spectral covariance matrix projected on the first 8 basis vectors of the KLT basis. Upper left: Eigenvalues of the first 8 KLT vectors. The first 6 KL spectral basis functions derived by PCA analysis of the critical-band spectral space are also shown.

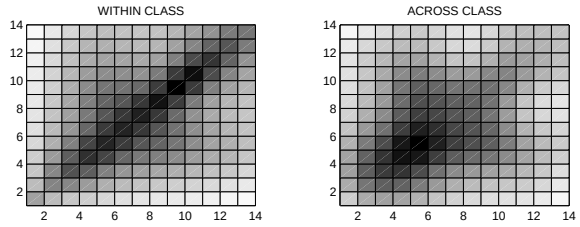

Figure 4: Within-class and between-class covariance matrixes for the critical-band spectrum of phonetically-labeled OGI Stories database.

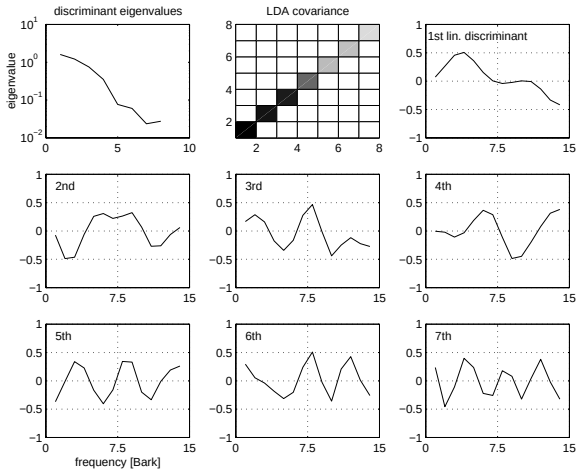

Figure 5: Upper left: Eigenvalues of the LDA-derived basis. Upper center: The total critical-band spectral correlation matrix, projected on the first 8 basis vectors of the LDA-derived basis. Upper left: Eigenvalues of the first 8 LDA-derived eigenvectors. The first 7 LDA-derived spectral basis functions of the critical-band spectral space are also shown.

the KL transform.

Computing the LDA projection involves computing principal components of the so called Fisher covariance matrix [8]

$$S_{wb} = S_w^{-1} S_b \tag{1}$$

where S_w refers to the matrix of the mean of the within-class variances and S_b the matrix of the variance of the means of the classes (Fig. 4). The S_w computed directly from the original data is not well conditioned and this may lead to difficulties in computing the S_{wb}. In this work we alleviated this problem by first smoothing the critical-band energy space using truncated KLT which preserved 95% of the original variance in the data.

The result is illustrated in Fig. 5 which shows first six eigenvectors (linear discriminants) of the KLT-smoothed (95% of variance) Fisher discriminant matrix [8] of the stops-excluded OGI stories critical-band spectral space.

Unlike the KLT transform, the total energy of the spectrum is no longer emphasized. The first discriminant appears to evaluate spectral energy in the first formant region and could be primarily discriminating between sonorant and non-sonorant sounds. The second and third discriminants are focusing on spectral ripples [12] in the central part of the critical-band spectrum, the second one being more sensitive to larger ripples than the third one. The 4th one evaluates spectral slope above 3 Bark and the 5th one is sensitive to about 5 Bark spectral ripple through the whole available spectrum. Higher discriminants with rather small eigenvalues are perhaps of lesser importance. Just as in the KLT case, the LDA transformed covariance matrix is also diagonal.

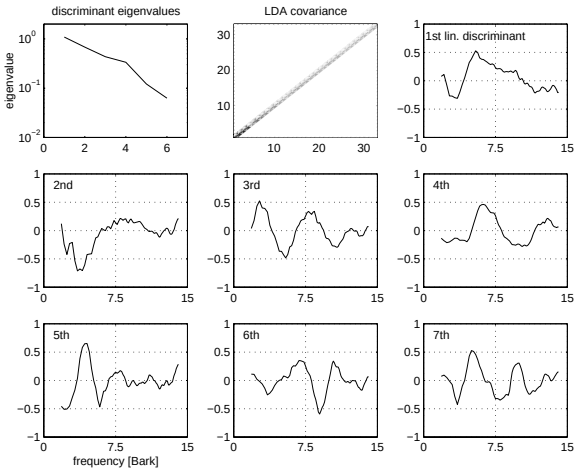

Figure 6: Upper left:Eigenvalues of the LDA-derived basis. Upper center: The total short-term spectrum (FFT-derived) correlation matrix, projected on the first 8 basis vectors of the LDA-derived basis. The first 6 LDA-derived spectral basis functions derived by LDA analysis of the short-term power-spectrum FFT space and displayed with the Bark-warped spectral axis are also shown.

2.4. Optimality of Bark scale?

It appears that zero-crossings of the LDA-derived spectral basis are reasonably uniformly spaced on the Bark scale of the auditory-like critical-band spectrum. Thus, it appears that the Bark frequency scale allows for use of a simple bases in phoneme classification. This would support optimality of the Bark (or Mel) scale for phoneme discrimination observed earlier [13].

To test this observation further we have also obtained LDA-derived spectral basis directly from the unsmoothed 129-point FFT power spectrum. The zero-crossings of the significant linear discriminants are typically more dense at lower frequencies. When displayed on the Bark frequency scale (see Fig. 6), the spectral bases resemble the bases from the critical-band analysis shown in Fig. 5. The most noticeable difference is reduced emphasis on higher frequencies.

3. PHONEME CLASSIFICATION EXPERIMENTS

To asses the effectiveness of the data-derived spectral bases we ran a phoneme-classification experiment. The task was to classify all frames of the test set which consisted of 29 phonemes present in the hand-labeled OGI Numbers database. MLP-based classifier achieves about

45% error on this task.

The classification was based on a single spectral frame. Logarithmic spectral means were subtracted from each file to partially compensate for communication channel differences. Speech from about 800 files was used in the training of a simple single-density, diagonal-covariance Gaussian classifier. Each file contains an utterance of natural number (zip codes, telephone numbers,..) by a single speaker. Some speakers utter more than one utterance. 50000 spectral frames from around 300 files were used in the test. That makes differences of about 0.5% significant at the 95% level according to a binomial test.

Three different spectral bases were evaluated:

1. MEL cosine spectral basis (8 Mel cepstrum coefficients, zeroth excluded).
2. KLT-derived spectral basis (first 8 coefficients).
3. LDA-derived spectral basis (first 8 coefficients).

Results of this experiment are shown in Table.1.

MATRIX	COS	PCA	LDA
FULL	53.7%	53.4%	53.4%
DIAG	57.3%	57.6%	56.6%

Table 1: Phoneme classification error on the OGI Stories corpus - full training

The full covariance classifier results are practically identical (they should be identical if there was no truncation of higher basis functions since such classifier is invariant under linear projections). For the diagonal covariance case, which is probably of most interest for HMM classification, the LDA-derived basis appears to be better than both the cepstrum and the KLT-derived basis. However, the difference (although significant according to the binomial test) is not large.

4. DISCUSSION AND CONCLUSIONS

We have shown that optimal spectral basis for projecting onto the direction of the maximum variability (KLT-derived) are different from the optimal spectral basis for projecting onto the direction of maximum phoneme separability (LDA-derived). The KLT-derived basis are similar to the conventional cosine basis used in cepstral analysis, the LDA-derived basis differ. Periodicity of the optimized spectral bases in Bark domain supports usefulness of Bark-like spectral warping in phoneme classification.

The alternative spectral bases so far do not seem to offer significant advantage in phoneme classification of spectral vectors. One would hope that there is more to gain from optimized projections of the spectral vector space. This result may suggest limited utility of single-frame spectral representations in classification of phonemes from fluent speech. Further experiments are pending to substantiate this conclusion.

ACKNOWLEDGMENTS

We thank Pratibha Jain for her significant help in running the classification experiments and Sarel van Vuuren for sharing his experience with discriminant analysis. Initial preliminary experiments with LDA in deriving the alternative spectral bases were carried out with Terri Kamm at the 1997 Summer Workshop at Johns Hopkins University. The work was supported by DoD (MDA904-98-1-0521, MDA904-97-1-0007), NSF (IRI-9712579) and by industrial grants from Intel and Texas Instruments to Anthropic Signal Processing Group at OGI.

5. REFERENCES

1. Avendano, C. and H. Hermansky (1997), "On the properties of temporal processing for speech in adverse environments", *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York.
2. Avendano, C. S. van Vuuren and H. Hermansky: Data based filter design for RASTA-like channel normalization in ASR, Proc. ICSLP-96, Philadelphia, October 1996.
3. Brown, P.: The Acoustic-Modeling Problem in Automatic Speech Recognition, PhD Thesis, Carnegie Mellon University, 1987.
4. Hermansky, H. (1990), "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752.
5. Hermansky, H.: Should recognizers have ears?, Speech Communication, Vol. 27, No. 1-3, pp. 3-27, September 1998
6. Hunt, M.: A Statistical Approach to Metrics for Word and Syllable Recognition, J. Acoust. Soc. Am. *66*(S1), S35(A), 1979.
7. Hunt, M. and Lefebvre, C. (1989), "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech", *Proc. Internat. Conf. Acoust. Speech Signal Processing*, Glasgow, Scotland, pp. 262-265.
8. Kil, D.H. and F.B. Shin: Pattern recognition and prediction with applications to sognal characterization, American Institute of Physics, 1996.
9. Merhav, N. and Chin-Hui Lee: On the asymptotic statistical behavior of empirical cepstral coefficients, IEEE Trans. Signal Processing, Vol. 41, No 3, May 1993.
10. Mermelstein, P. (1976), "Distance measures for speech recognition, psychological and instrumental", in *Pattern Recognition and Artificial Intelligence*, R.C.H. Chen, ed., Academic Press: New York, pp. 374-388.
11. Sarel van Vuuren and Hynek Hermansky: Data-driven design of RASTA-like filters, Proc. EUROSPEECH 97, Rhodos , Greece, September 1997.
12. Wang, K. and S.S. Shamma (1995), "Spectral Shape Analysis in the Central Auditory System", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 5, pp. 382-394.
13. Umesh, S., L. Cohen, and D. Nelson: Frequency-Warping and Speaker-Normalization, Proc. ICASSP-97, pp. 983-987, Munich, Germany, April 1997.