APPLICATION OF CLASSIFIER-OPTIMAL TIME-FREQUENCY DISTRIBUTIONS TO SPEEECH ANALYSIS

J. Droppo and L. Atlas

Department of Electrical Engineering, Box 35250
University of Washington, Seattle, WA 98195-2500, USA

ABSTRACT

Discrete operator theory maps each discrete time signal to a multitude of time-frequency distributions, each uniquely specified by a kernel function. This kernel function selects some details to emphasize and other details to smooth. Traditionally, kernels are chosen to impart specific properties to the resulting distributions, such as satisfying the marginals or reducing cross-terms.

Given a labeled set of data from several classes, we seek to generate a kernel function that emphasizes classification relevant details present in the distribution. In this paper, we extend our previous work on class dependent time-frequency distributions. The new kernel formulation is similar to [1], with one modification. Previously, the discriminant function did not consider the within-class to between-class variance of coefficients, and was vulnerable to choosing very "noisy" features.

1. INTRODUCTION

Discrete operator theory establishes a direct link between a discrete signal and its discrete-time, discrete-frequency time-frequency representation (TFR) [2]. The discrete Rihaczek distribution [3] contains all correlations of a signal with time and frequency shifted versions of itself. This representation includes the traditional autocorrelation sequence, the discrete Fourier transform (DFT) of the signal's instantaneous energy, and everything in between.

The discrete Rihaczek distribution contains many non-smooth features. There is a significant amount of detail present that may or may not be useful for any given task. Our goal is to find a kernel function that maps the Rihaczek distribution onto a new time-frequency distribution that has been optimized for classification.

2. KERNEL GENERATION

Given a discrete-time signal $x[n]$, the discrete-time discrete-frequency Rihaczek distribution is

$$P_R[n,k] = x^*[n]X[k]\exp(j2\pi nk/N). \qquad (1)$$

By taking the two dimensional discrete Fourier transform, this is expressed in the ambiguity plane as

$$M_R[\eta,\tau] = \sum_n x^*[n]x[n+\tau]\exp(j2\pi n\eta/N). \qquad (2)$$

A kernel function can be specified in either plane. $\Phi[n,k]$ operates convolutionally on P_R in the time-frequency plane, and $\Phi[\eta,\tau]$ operates as a mask on M_R in the ambiguity plane.

3. DIMENSIONALITY REDUCTION

For a signal length N, the ambiguity plane contains N^2 coefficients. The coefficients along the $\eta = 0$ axis are the autocorrelation of the sequence, and contain information about the stationary spectrum of the signal.

$$M_R[0,\tau] = \sum_n x^*[n]x[n+\tau] \qquad (3)$$

The $\tau = 0$ axis contains the discrete Fourier transform of the instantaneous signal energy. A non-stationary signal manifests itself along this area.

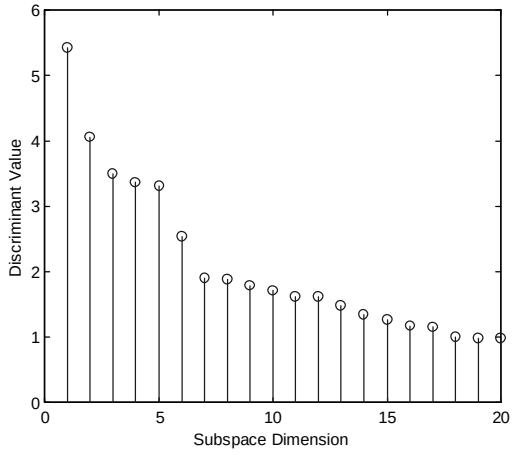

Figure 1: Fishers linear discriminant value for the first 20 dimensions of the phones "aa" and "uw".

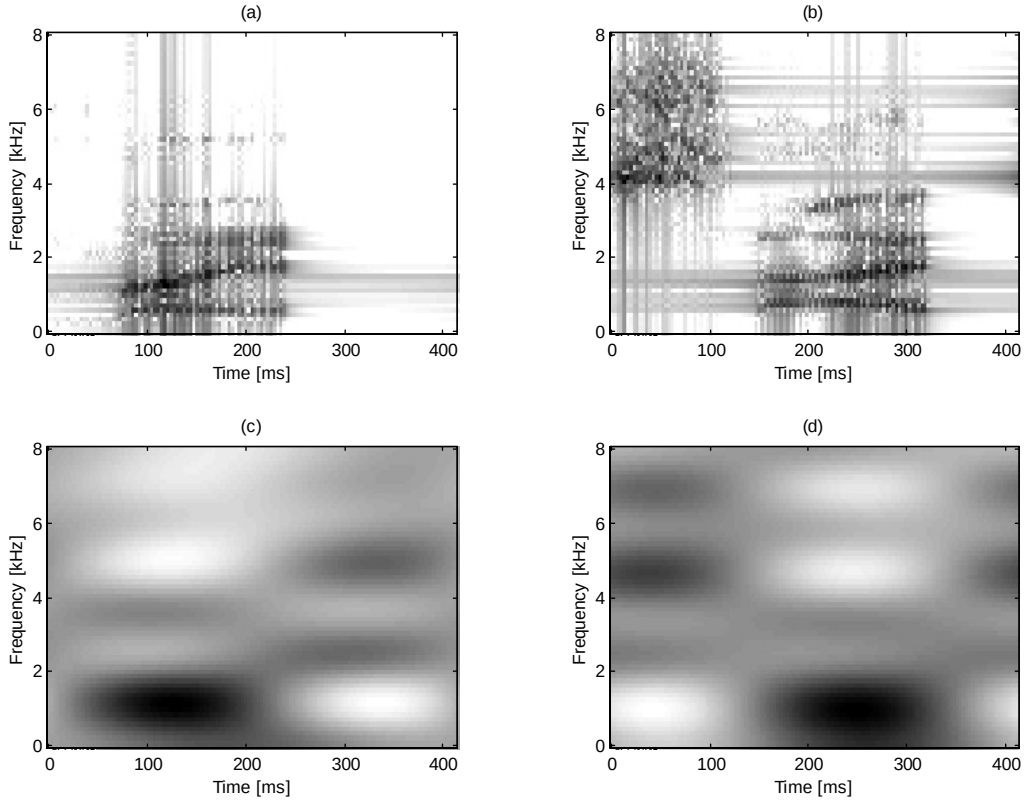

Figure 2: Class-conditional kernel trained on the sounds "may" and "say".
(a) and (b) log-magnitude Rihaczek distributions for two test signals.
(c) and (d) class-conditional time-frequency representations

$$M_R[\eta,0] = \sum_n x^*[n]x[n]\exp(j2\pi n\eta / N) \qquad (4)$$

Somewhere among these N^2 coefficients is the information most relevant to the task at hand, namely classification. If stationary spectral information is important, for example, the kernel should concentrate itself along the $\eta = 0$ axis. The resulting time-frequency distribution shows almost no time variation, regardless of the nature of the input signal. The spectral information remains relatively untouched.

Conversely, if time-modulation information is important for the classification task, the kernel should concentrate itself along the $\tau = 0$ axis. The spectral information is disregarded, and the modulation is emphasized.

Each coefficient in the ambiguity plane is taken as an independent dimension in the feature space. Then, given training data from a set of known classes, a linear Fisher's discriminant is used to rank-order the importance of these dimensions.

Figure 1 contains a plot of the discriminant value for the first twenty significant dimensions of the data.

Standard practice dictates that only the coefficients with high numerical values are kept, and the rest are not considered. A kernel that passes only the desirable points in the ambiguity plane and masks the rest is then generated. It is given a constant value at $[\eta, \tau]$ values that correspond to a large linear Fisher's discriminant, and zero everywhere else.

4. CLASSIFICATION

The classifier described in [1] found the training class whose centroid had a minimum distance to the test signal in the subspace defined by the kernel. A distance metric, $D_n(x)$, is computed for each class, n.

$$D_n(x) = (x-\mu_n)^T(x-\mu_n) \qquad (5)$$

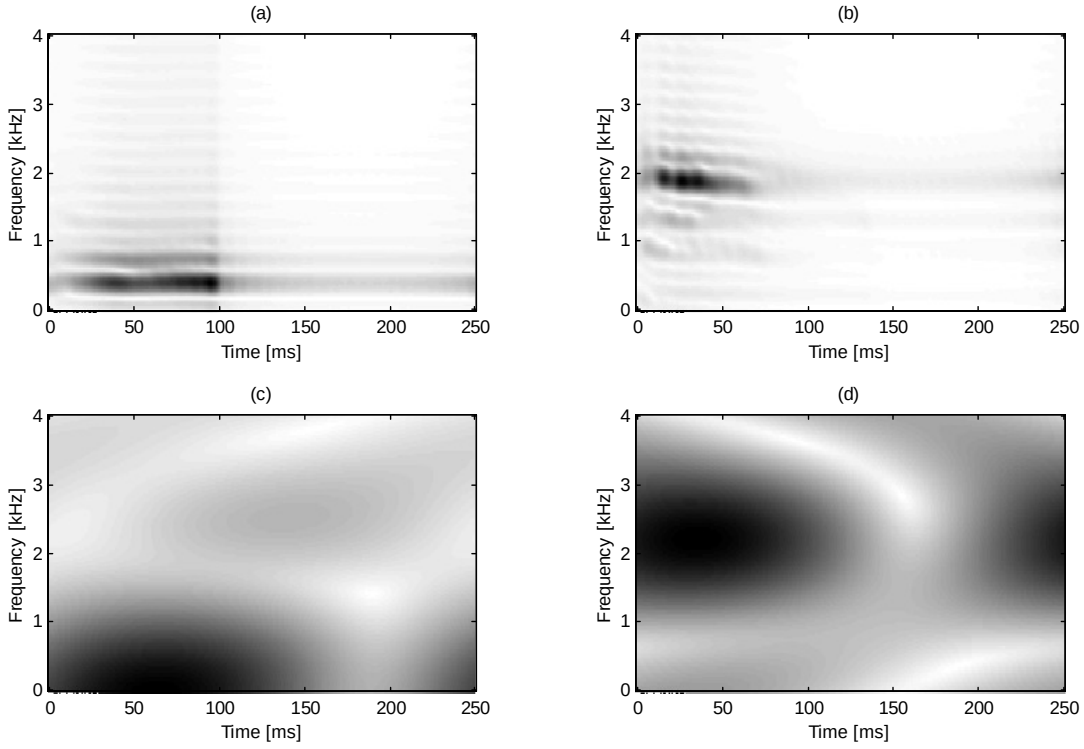

Figure 3: Class-conditional kernel trained on the sounds "aa" and "uw".
(a) and (b) log-magnitude Rihaczek distributions for two test signals.
(c) and (d) class-conditional time-frequency representations

Under the assumption that each dimension of each class is Gaussian, statistically independent, and has the same variance, this is equivalent to a Bayes-optimal classifier.

Relaxing this assumption, we can assume that each training class has a Gaussian distribution in the subspace defined by the kernel. The parameters of this distribution, namely the mean and covariance matrix, can then be estimated from the training data, and used in classification. The minimum centroid distance is replaced with a minimum Mahalanobis distance, resulting in a more general Bayes-optimal classifier.

$$D_n(x) = \left(2\pi|\Sigma_n|\right)^{-1/2} \exp\left((x-\mu_n)^T \Sigma_n^{-1}(x-\mu_n)\right) \quad (6)$$

5. RESULTS

State of the art speech processing uses streams of short-time estimates of long-term spectral information, such as cepstral coefficients. Each vector of coefficients typically represents 25ms of signal, and comes at a rate of one every 10ms. Recognition systems introduce delta cepstral coefficients and Markov models to model a signal's time variation, but it is not clear that either is an optimal way of encoding the information.

Long term features, which implicitly incorporate time variation and spectral information, introduce the possibility of simplifying and improving the state of the art. Preliminary tests with consonant-vowel pairs and vowels are promising. The kernel function correctly determines which spectral and time information is necessary for discrimination.

5.1 Consonant-Vowel Pair Discrimination

Figure 2 is typical of our class-conditional result. Ten examples each of the words "may" and "say" from different speakers in the TIMIT database were used to generate a kernel that could differentiate these two utterances. In both Rihaczek distributions, the formant structure and glottal excitation are clearly present. For the utterance "say," the initial fricitave is also apparent.

After applying the class-conditional kernel, the formant and glottal structure has been disregarded; it is not useful for this classification task. The "may" utterance is

characterized by energy below 2kHz early in the signal, and the “say” utterance is characterized by its initial fricitave energy and later onset of voiced energy. The kernel has extracted this core piece of information, which distinguishes the two utterances.

5.2 Vowel Discrimination

Figure 3 shows the time-frequency distributions generated to discriminate the vowel sounds "aa" and "uw". Again, the Rihaczek distributions show clear time and frequency structures. The class-conditional kernel removes the excitation structures, and smoothes the formant structures just enough such that the two are distinguishable.

6. CONCLUSIONS

We have applied the class-conditional kernel generation method to two speech discrimination tasks.

The Rihaczek distribution of a signal contains all correlations of a signal with frequency modulated versions of itself, which yields a N^2-dimensional feature space. The class conditional kernel selects a subspace that is suited for the classification task.

After the kernel is applied, there is still enough structure remaining to discriminate between the two classes, and extraneous information has been removed.

7. REFRENCES

[1] Les Atlas, James Droppo and Jack McLaughlin, “Optimizing Time-Frequency Distributions for Automatic Classification,” Proc. SPIE 1997, Volume 3162.

[2] Jack McLaughlin, James Droppo and Les Atlas, “Class-Dependent, Discrete Time-Frequency Distributions via Operator Theory,” Proceedings of the 1997 IEEE ICASSP, vol. 3, p. 2045-8, 1997.

[3] A. W. Rihaczek, “Signal Energy Distribution in Time and Frequency,” IEEE Trans. Info. Theory, vol. 14, pp. 369-74, 1968.