The PDF Projection Theorem and the Class-Specific Method

Paul M. Baggenstoss, *Member, IEEE*

***Abstract*—In this paper, we present the theoretical foundation for optimal classification using class-specific features and provide examples of its use. A new probability density function (PDF) projection theorem makes it possible to project probability density functions from a low-dimensional feature space back to the raw data space. An M-ary classifier is constructed by estimating the PDFs of class-specific features, then transforming each PDF back to the raw data space where they can be fairly compared. Although statistical sufficiency is not a requirement, the classifier thus constructed will become equivalent to the optimal Bayes classifier if the features meet sufficiency requirements individually for each class. This classifier is completely modular and avoids the dimensionality curse associated with large complex problems. By recursive application of the projection theorem, it is possible to analyze complex signal processing chains. We apply the method to feature sets including linear functions of independent random variables, cepstrum, and MEL cepstrum. In addition, we demonstrate how it is possible to automate the feature and model selection process by direct comparison of log-likelihood values on the common raw data domain.**

***Index Terms*—Bayesian classification, class-dependent features, classification, class-specific features, hidden Markov models, maximum likelihood estimation, pattern classification, PDF estimation, probability density function.**

I. Introduction

A. Overview and Outline

In this paper, we introduce a theorem that can be applied to any statistical approach, which makes use of likelihood comparisons, such as detection, classification, and statistical modeling. The theorem allows likelihood comparisons to be made in the common raw data domain while the difficult task of probability density function (PDF) estimation can be made in class (or state) dependent low-dimensional feature spaces. Because each feature set can be designed without regard to other classes (or states), it can be of much lower dimension than a common feature set that must account for all classes, effectively avoiding the curse of dimensionality. The transformation of feature PDFs to the raw data domain, which we term "PDF projection," is accomplished by deriving a correction term that amounts to a generalized Jacobian of the feature transformation. This correction term depends only upon the feature transformation and a hand-picked class (or state) dependent statistical reference hypothesis. When combined with the feature likelihood value, it results in a raw data likelihood function which is guaranteed by the theorem to be a PDF on the raw data space. Examples of the method involving commonly used autoregressive and cepstrum features are provided.

A few words about the chronology of development are in order. This paper is based on previous work in class-specific features by the author and by Prof. S. Kay at the University of Rhode Island. The first two papers on the subject [1], [2] describe the original form of the class-specific method, which was based on a common fixed reference hypothesis and the properties of sufficient statistics. Although the present method is based on this previous work, we say little about it in this paper. This is because the present work is best understood from the viewpoint of PDF projection, and it would confuse the readers to begin with sufficient statistics. The interested reader is encouraged to examine this previous work, especially [2].

In Section I, we review classical Bayesian classification and discuss the dimensionality problem. In Section II, we introduce the PDF projection theorem (PPT) and the associated chain rule. In Section III, we discuss various methods of calculating the PPT correction term. In Section IV, we discuss how to apply the PPT to classification. In Section V, we apply the method to feature transformations involving linear functions of independent random variables (RVs). In Section VI, we apply the method to cepstrum and MEL cepstrum. In Section VII, we present a method of automatic feature selection.

B. Classical Classification Theory and the Dimensionality Problem

The so-called M-ary classification problem is that of assigning a multidimensional sample of data $\mathbf{x} \in \mathcal{R}^N$ to one of M classes. The statistical hypothesis that class j is true is denoted by H_j, $1 \le j \le M$. The statistical characterization of $\mathbf{x}$ under each of the M hypotheses is described completely by the PDFs, which are written $p(\mathbf{x}|H_j)$, $1 \le j \le M$. Classical theory as applied to the problem results in the so-called Bayes classifier, which simplifies to the Neyman–Pearson rule for equiprobable prior probabilities

$$j^* = \arg\max_j p(\mathbf{x}|H_j). \tag{1}$$

Because this classifier attains the minimum probability of error of all possible classifiers, it is the basis of most classifier designs. Unfortunately, it does not provide simple solutions to the dimensionality problem that arises when the PDFs are unknown and must be estimated. The most common solution is to reduce the dimension of the data by extraction of a small number of

Manuscript received March 20, 2002; revised October 17, 2002. This work was supported by the Office of Naval Research. The associate editor coordinating the review of this paper and approving it for publication was Dr. Hamid Krim.

The author is with the Naval Undersea Warfare Center, Newport, RI 02841 USA (e-mail: p.m.baggenstoss@ieee.org).

Digital Object Identifier 10.1109/TSP.2002.808109

information-bearing features $\mathbf{z} = T(\mathbf{x})$, then recasting the classification problem in terms of $\mathbf{z}$:

$$j^* = \arg\max_j p(\mathbf{z}|H_j). \tag{2}$$

This leads to a fundamental tradeoff: whether to discard features in an attempt to reduce the dimension to something manageable or to include them and suffer the problems associated with estimating a PDF at high dimension. Unfortunately, there may be no acceptable compromise. Virtually all methods which attempt to find decision boundaries on a high-dimensional space are subject to this tradeoff or "curse" of dimensionality. For this reason, many researchers have explored the possibility of using class-specific features [3]–[9].

The basic idea in using class-specific features is to extract M class-specific feature sets $\mathbf{z}_j = T_j(\mathbf{x}), \quad 1 \le j \le M$, where the dimension of each feature set is small, and then to arrive at a decision rule based only upon functions of the lower dimensional features. Unfortunately, the classifier modeled on the Neyman–Pearson rule

$$j^* = \arg\max_j p(\mathbf{z}_j|H_j) \tag{3}$$

is invalid because comparisons of densities on different feature spaces are meaningless. One of the first approaches that comes to mind is to computes for each class a likelihood ratio against a common hypothesis composed of "all other classes." While this seems beneficial on the surface, there is no theoretical dimensionality reduction since for each likelihood ratio to be a sufficient statistic, "all features" must be included when testing each class against a hypothesis that includes "all other classes." A number of other approaches have emerged in recent years to arrive at meaningful decision rules. Each method makes a strong assumption (such as that the classes fall into linear subspaces) that limits the applicability of the method or else uses *ad hoc* method of combining the likelihoods of the various feature sets.

1) A method used in speech recognition [3] uses phone-specific features. While, at first, this method appears to use class-specific features, it is actually using the same features extracted from the raw data but applyoing different models to the time evolution of these features.
2) A method of image recognition [10] uses class-specific features to detect various image "fragments." The method uses a nonprobabilistic means of combining fragments to form an image.
3) A method has been proposed that tests all pairs of classes [4]. To be exhaustive, this method has a complexity of $O(M^2)$ different tests and may be prohibitive for large M. A hierarchical approach has been proposed based on a binary tree of tests [5]. Implementation of the binary tree requires initial classification into meta-classes, which is an approach that is suboptimal because it makes hard decisions based on limited information.
4) Methods based on linear subspaces [6], [7] are popular because they use the powerful tool of linear subspace analysis. These methods can perform well in certain applications but are severely limited to problems where when the classes are separable by linear processing.
5) Support vectors [8] are a relatively new approach that is based on finding a linear decision function between every pair of classes.

As evidenced by the various approaches, there is a strong motivation for using class-specific features. Unfortunately, classical theory as it stands requires operating in a common feature space and fails to provide any guidance for a suitable class-specific architecture. In this paper, we present an extension to the classical theory that provides for an optimal architecture using class-specific features.

II. PDF Projection Theorem

It is well known how to write the PDF of $\mathbf{x}$ from the PDF of $\mathbf{z}$ when the transformation is 1:1. This is the change of variables theorem from basic probability. Let $\mathbf{z} = T(\mathbf{x})$, where $T(\mathbf{x})$ is an invertible and differentiable multidimensional transformation. Then

$$p_x(\mathbf{x}) = |\mathbf{J}(\mathbf{x})|\, p_z(T(\mathbf{x})) \tag{4}$$

where $|\mathbf{J}(\mathbf{x})|$ is the determinant of the Jacobian matrix of the transformation

$$\mathbf{J}_{ij} = \frac{\partial z_i}{\partial x_j}.$$

What we seek is a generalization of (4), which is valid for many-to-1 transformations. Define

$$\mathcal{P}(T, p_z) = \{p_x(\mathbf{x}) : \mathbf{z} = T(\mathbf{x}) \text{ and } \mathbf{z} \sim p_z(\mathbf{z})\}$$

that is, $\mathcal{P}(T, p_z)$ is the set of PDFs $p_x(\mathbf{x})$, which, through $T(\mathbf{x})$, generate PDF $p_z(\mathbf{z})$ on $\mathbf{z}$. If $T(\ )$ is many-to-one, $\mathcal{P}(T, p_z)$ will contain more than one member. Therefore, it is impossible to uniquely determine $p_x(\mathbf{x})$ from $T(\ )$ and $p_z(\mathbf{z})$. We can, however, find a particular solution if we constrain $p_x(\mathbf{x})$. In order to apply the constraint, it is necessary to make use of a reference hypothesis H_0 for which we know the PDF of both $\mathbf{x}$ and $\mathbf{z}$. If we constrain $p_x(\mathbf{x})$ such that for every transform pair $(\mathbf{x}, \mathbf{z})$ we have

$$\frac{p_x(\mathbf{x})}{p_x(\mathbf{x}|H_0)} = \frac{p_z(\mathbf{z})}{p_z(\mathbf{z}|H_0)} \tag{5}$$

or that the likelihood ratio (with respect to H_0) is the same in both the raw data and feature domains, we arrive at a satisfactory answer. We cannot offer a justification for this constraint other than it is a means of arriving at an answer; however, we will soon show that this constraint produces desirable properties. The particular form of $p_x(\mathbf{x})$ is uniquely defined by the constraint itself, namely

$$p_x(\mathbf{x}) = \frac{p_x(\mathbf{x}|H_0)}{p_z(\mathbf{z}|H_0)} p_z(\mathbf{z}); \text{ where } \mathbf{z} = T(\mathbf{x}). \tag{6}$$

Theorem 1 states that not only is $p_x(\mathbf{x})$ a PDF but that it is a member of $\mathcal{P}(T, p_z)$.

Theorem 1—PDF Projection Theorem: Let H_0 be some fixed reference hypothesis with known PDF $p_x(\mathbf{x}|H_0)$. Let $\mathcal{X}$ be the region of support of $p_x(\mathbf{x}|H_0)$. In other words, $\mathcal{X}$ is the set of all points $\mathbf{x}$, where $p_x(\mathbf{x}|H_0) > 0$. Let $\mathbf{z} = T(\mathbf{x})$ be a many-to-one transformation. Let $\mathcal{Z}$ be the image of $\mathcal{X}$ under

the transformation $T(\mathbf{x})$. Let the PDF of $\mathbf{z}$ when $\mathbf{x}$ is drawn from $p_x(\mathbf{x}|H_0)$ exist and be denoted by $p_z(\mathbf{z}|H_0)$. It follows that $p_z(\mathbf{z}|H_0) > 0$ for all $\mathbf{z} \in \mathcal{Z}$. Now, let $p_z(\mathbf{z})$ be any PDF with the same region of support $\mathcal{Z}$. Then, the function (6) is a PDF on $\mathcal{X}$, and thus

$$\int_{\mathbf{x}\in\mathcal{X}} p_x(\mathbf{x})dx = 1.$$

Furthermore, $p_x(\mathbf{x})$ is a member of $\mathcal{P}(T, p_z)$.

Proof: These assertions are proved in [11].

A. Usefulness and Optimality Conditions of the Theorem

The theorem shows that, provided we know the PDF under some reference hypothesis H_0 at both the input and output of transformation $T(\mathbf{x})$, if we are given an arbitrary PDF $p_z(\mathbf{z})$ defined on $\mathbf{z}$, we can immediately find a PDF $p_x(\mathbf{x})$ defined on $\mathbf{x}$ that generates $p_z(\mathbf{z})$. Although it is interesting that $p_x(\mathbf{x})$ generates $p_z(\mathbf{z})$, there are an infinite number of them, and it is not yet clear that $p_x(\mathbf{x})$ is the best choice. However, suppose we would like to use $p_x(\mathbf{x})$ as an approximation to the PDF $p_x(\mathbf{x}|H_1)$. Let this approximation be

$$\hat{p}_x(\mathbf{x}|H_1) \triangleq \frac{p_x(\mathbf{x}|H_0)}{p_z(\mathbf{z}|H_0)}\hat{p}_z(\mathbf{z}|H_1) \text{ where } \mathbf{z} = T(\mathbf{x}). \quad (7)$$

From Theorem 1, we see that (7) is a PDF. Furthermore, if $T(\mathbf{x})$ is a sufficient statistic for H_1 versus H_0, then as $\hat{p}_z(\mathbf{z}|H_1) \to p_z(\mathbf{z}|H_1)$, we have

$$\hat{p}_x(\mathbf{x}|H_1) \to p_x(\mathbf{x}|H_1).$$

This is immediately seen from the well-known property of the likelihood ratio, which states that if $T(\mathbf{x})$ is sufficient for H_1 versus H_0

$$\frac{p_x(\mathbf{x}|H_1)}{p_x(\mathbf{x}|H_0)} = \frac{p_z(\mathbf{z}|H_1)}{p_z(\mathbf{z}|H_0)}. \quad (8)$$

Note that for a given H_1, the choice of $T(\mathbf{x})$ and H_0 are coupled so that they must be chosen *jointly*. In addition, note that the sufficiency condition is required for optimality, but is not necessary for 7 to be a valid PDF. Here, we can see the importance of the theorem. The theorem, in effect, provides a means of creating PDF approximations on the high-dimensional input data space without dimensionality penalty using low-dimensional feature PDFs and provides a way to optimize the approximation by controlling both the reference hypothesis H_0 as well as the features themselves. This is the remarkable property of Theorem 1: that the resulting function remains a PDF whether or not the features are sufficient statistics. Since sufficiency means optimality of the classifier, approximate sufficiency means PDF approximation and approximate optimality.

Theorem 1 allows maximum likelihood (ML) methods to be used in the raw data space to optimize the accuracy of the approximation over T and H_0 as well as $\boldsymbol{\theta}$. Let $\hat{p}_z(\mathbf{z}|H_1)$ be parameterized by the parameter $\boldsymbol{\theta}$. Then, the maximization

$$\max_{\boldsymbol{\theta},T,H_0} \left\{ \frac{p_x(\mathbf{x}|H_0)}{p_z(\mathbf{z}|H_0)}\hat{p}_z(\mathbf{z}|H_1;\boldsymbol{\theta}),\ \mathbf{z} = T(\mathbf{x}) \right\} \quad (7a)$$

is a valid ML approach and can be used for model selection (with appropriate data cross-validation).

Example 1: In this simple example, we demonstrate the applicability of Theorem 1. We consider the case of independent Gaussian RV's and two hypotheses concerning the mean. Let $\mathbf{x} = [x_1 \dots x_N]'$. Let the feature transformation be

$$z = T(\mathbf{x}) = \sum_{i=1}^{K} x_i$$

where $1 \le K \le N$. Let $\mathcal{E}(x_i) = 0$ under H_0 and $\mathcal{E}(x_i) = 1$ under H_1, where $\mathcal{E}(\ )$ is the expectation operator. Because H_0 and H_1 are hypotheses concerning the mean of Gaussian RV's with fixed variance, z is a sufficient statistic for the mean when $K = N$. The Gaussian PDF of $\mathbf{x}$ under H_0 may be written

$$p_x(\mathbf{x}|H_0) = (2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N} x_i^2\right\}.$$

Under H_0, z will be Gaussian zero-mean with variance $K\sigma^2$, and thus

$$p_z(z|H_0) = (2\pi K\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2K\sigma^2}z^2\right\}.$$

We let $p_z(z)$ be the Gaussian PDF

$$p_z(z|H_1) = (2\pi K\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2K\sigma^2}(z-K)^2\right\}.$$

By the projection theorem

$$p_x(\mathbf{x}) = \frac{p_x(\mathbf{x}|H_0)}{p_z(z|H_0)}p_z(z|H_1).$$

Thus

$$\begin{aligned}
p_x(\mathbf{x}) &= (2\pi\sigma^2)^{-N/2}(2\pi K\sigma^2)^{1/2}(2\pi K\sigma^2)^{-1/2} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N} x_i^2 - \frac{z^2}{K} + \frac{1}{K}(z-K)^2\right]\right\} \\
p_x(\mathbf{x}) &= (2\pi\sigma^2)^{-N/2} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N} x_i^2 - \frac{z^2}{K} + \frac{z^2}{K} - 2z + K\right]\right\} \\
p_x(\mathbf{x}) &= (2\pi\sigma^2)^{-N/2} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N} x_i^2 - 2z + K\right]\right\} \\
&= (2\pi\sigma^2)^{-N/2} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{K}(x_i-1)^2 + \sum_{i=K+1}^{N} x_i^2\right]\right\}
\end{aligned}$$

where we have made the substitution $z = \sum_{i=1}^{K} x_i$. It is clear that the result is a Gaussian PDF with mean $\mu_i = 1$ for $1 \le i \le K$ and $\mu_i = 0$ for $K+1 \le i \le N$. Note also that it is a PDF, regardless of K (that is to say the sufficiency of z). It is also clear that the PDF $p_x(\mathbf{x})$ generates the PDF $p_z(z)$. In addition, note that if $K = N$, then $p_x(\mathbf{x}) = p_x(\mathbf{x}|H_1)$, as predicted by the theory.

B. Data-Dependent Reference Hypothesis

We now mention a useful property of (7). Let $\mathcal{H}_z$ be a *region of sufficiency* (ROS) of $\mathbf{z}$, which is defined as a set of all hypotheses such that for every pair of hypotheses $H_{0a},\ H_{0b} \in \mathcal{H}_z$, we have

$$\frac{p_x(\mathbf{x}|H_{0a})}{p_x(\mathbf{x}|H_{0b})} = \frac{p_z(\mathbf{z}|H_{0a})}{p_z(\mathbf{z}|H_{0b})}.$$

An ROS may be thought of as a family of PDFs traced out by the parameters of a PDF, where $\mathbf{z}$ is a sufficient statistic for the parameters. The ROS may or may not be unique. For example, the ROS for a sample mean statistic could be a family of Gaussian PDFs with variance 1 traced out by the mean parameter. Another ROS would be produced by a different variance. The "J-function"

$$J(\mathbf{x}, T, H_0) \triangleq \frac{p_x(\mathbf{x}|H_0)}{p_z(T(\mathbf{x})|H_0)} = \frac{p_x(\mathbf{x}|H_0)}{p_z(\mathbf{z}|H_0)}$$

is independent of H_0 as long as H_0 remains within ROS $\mathcal{H}_z$.

Defining the ROS should in no way be interpreted as a sufficiency requirement for $\mathbf{z}$. All statistics $\mathbf{z}$ have an ROS that may or may not include H_1 (it does only in the ideal case). Defining $\mathcal{H}_z$ is used only in determining the allowable range of reference hypotheses when using a data-dependent reference hypothesis.

For example, let $\mathbf{z}$ be the sample variance of $\mathbf{x}$. Let $H_0(\sigma^2)$ be the hypothesis that $\mathbf{x}$ is a set of N independent identically distributed zero-mean Gaussian samples with variance σ^2. Clearly, an ROS for $\mathbf{z}$ is the set of all PDFs traced out by σ^2. We have

$$p(\mathbf{x}|H_0(\sigma^2)) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N} x_i^2\right\}$$

and, since $\mathbf{z}$ is a $\chi^2(N)$ random variable (scaled by $1/N$)

$$p(\mathbf{z}|H_0(\sigma^2)) = \frac{N}{\sigma^2\Gamma\left(\frac{N}{2}\right)} 2^{-N/2} \left(\frac{Nz}{\sigma^2}\right)^{N/2-1} \exp\left(-\frac{zN}{2\sigma^2}\right).$$

It is easily verified that the contribution of σ^2 is canceled in the J-function ratio.

Because $J(\mathbf{x}, T, H_0(\sigma^2))$ is independent of σ^2, it is possible to make σ^2 a function of the data itself, changing it with each input sample. In the example above, since $\mathbf{z}$ is the sample variance, we could let the assumed variance under H_0 depend on $\mathbf{z}$ according to $\sigma^2 = \mathbf{z}$.

However, if $J(\mathbf{x}, T, H_0(\sigma^2))$ is independent of σ^2, one may question what purpose does it serve to vary σ^2. The reason is purely numerical. Note that in general, we do not have an analytic form for the J-function but instead have separate numerator and denominator terms. Often, computing $J(\mathbf{x}, T, H_0(\sigma^2))$ can pose some tricky numerical problems, particularly if $\mathbf{x}$ and $\mathbf{z}$ are in the tails of the respective PDFs. Therefore, our approach is to position H_0 to maximize the numerator PDF (which simultaneously maximizes the denominator). Another reason to do this is to allow PDF approximations to be used in the denominator that are not valid in the tails, such as the central limit theorem (CLT).

In our example, the maximum of the numerator clearly happens at $\sigma^2 = \mathbf{z}$ because $\mathbf{z}$ is the maximum likelihood estimator of σ^2. We will explore the relationship of this method to asymptotic ML theory in a later section. To reflect the possible dependence of H_0 on $\mathbf{z}$, we adopt the notation $H_0(\mathbf{z})$. Thus

$$\hat{p}_x(\mathbf{x}|H_1) = \frac{p_x(\mathbf{x}|H_0(\mathbf{z}))}{p_z(\mathbf{z}|H_0(\mathbf{z}))}\hat{p}_z(\mathbf{z}|H_1), \text{ where } \mathbf{z} = T(\mathbf{x}). \quad (9)$$

The existence of $\mathbf{z}$ on the right side of the conditioning operator $|$ is admittedly a very bad use of notation but is done for simplicity. The meaning of $\mathbf{z}$ can be understood using the following imaginary situation. Imagine that we are handed a data sample $\mathbf{x}$, and we evaluate (7) for a particular hypothesis $H_0 \in \mathcal{H}_z$. Out of curiosity, we try it again for a different hypothesis of $H_0' \in \mathcal{H}_z$. We find that no matter which $H_0 \in \mathcal{H}_z$ we use, the result is the same. We notice, however, that for an H_0 that produces larger values of $p_x(\mathbf{x}|H_0(\mathbf{z}))$ and $p_z(\mathbf{z}|H_0(\mathbf{z}))$, the requirement for numerical accuracy is less stringent. It may require fewer terms in a polynomial expansion or else fewer bits of numerical accuracy. Now, we are handed a new sample of $\mathbf{x}$, but this time, having learned our lesson, we immediately choose the $H_0 \in \mathcal{H}_z$ that maximizes $p_x(\mathbf{x}|H_0(\mathbf{z}))$. If we do this every time, we realize that H_0 is now a function of $\mathbf{z}$. The dependence, however, carries no statistical meaning and only has a numerical interpretation.

In many problems, $\mathcal{H}_z$ is not easily found, and we must be satisfied with *approximate* sufficiency. In this case, there is a weak dependence of $J(\mathbf{x}, T, H_0)$ upon H_0. This dependence is generally unpredictable unless, as we have suggested, $H_0(\mathbf{z})$ is always chosen to maximize the numerator PDF. Then, the behavior of $J(\mathbf{x}, T, H_0)$ is somewhat predictable. Because the numerator is always maximized, the result is a positive bias. This positive bias is most notable when there is a good match to the data, which is a desirable feature.

C. Asymptotic ML Theory as a Special Case of the PDF Projection Theorem

We have stated that when we use a data-dependent reference hypothesis, we prefer to choose the reference hypothesis such that the numerator of the J-function is a maximum. Since we often have parametric forms for the PDFs, this amounts to finding the ML estimates of the parameters. If there are a small number of features, *all* of the features are ML estimators for parameters of the PDF, and there is sufficient data to guarantee that the ML estimators fall in the asymptotic (large data) region, then the data-dependent hypothesis approach is equivalent to an existing approach based on classical asymptotic ML theory. We will derive the well-known asymptotic result using (9).

Two well-known results from asymptotic theory [12] are the following.

1) Subject to certain regularity conditions (large amount of data, a PDF that depends on a finite number of parameters and is differentiable, etc.), the PDF $p_x(\mathbf{x};\boldsymbol{\theta}^*)$ may be approximated by

$$p_x(\mathbf{x};\boldsymbol{\theta}^*) \simeq p_x(\mathbf{x};\hat{\boldsymbol{\theta}}) \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}})'\mathbf{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}})\right\} \quad (10)$$

where $\boldsymbol{\theta}^*$ is an arbitrary value of the parameter, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, and $\mathbf{I}(\boldsymbol{\theta})$ is the *Fisher's information matrix* (FIM) [12]. The components of the FIM for PDF parameters θ_k, θ_l are given by

$$\mathbf{I}_{\theta_k,\theta_l}(\boldsymbol{\theta}) = -\mathbf{E}\left(\frac{\partial^2 \ln p_x(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_k \partial\theta_l}\right).$$

The approximation is valid only for $\boldsymbol{\theta}^*$ in the vicinity of the MLE (and the true value).

2) The MLE $\hat{\boldsymbol{\theta}}$ is approximately Gaussian with mean equal to the true value $\boldsymbol{\theta}$ and covariance equal to $\mathbf{I}^{-1}(\boldsymbol{\theta})$ or

$$p_\theta(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) \simeq (2\pi)^{-P/2}|\mathbf{I}(\hat{\boldsymbol{\theta}})|^{1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\mathbf{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} \tag{11}$$

where P is the dimension of $\boldsymbol{\theta}$. Note that we use $\hat{\boldsymbol{\theta}}$ in evaluating the FIM in place of $\boldsymbol{\theta}$, which is unknown. This is allowed because $\mathbf{I}^{-1}(\boldsymbol{\theta})$ has a weak dependence on $\boldsymbol{\theta}$. The approximation is valid only for $\boldsymbol{\theta}$ in the vicinity of the MLE.

To apply (9), $\hat{\boldsymbol{\theta}}$ takes the place of $\mathbf{z}$, and $H_0(\mathbf{z})$ is the hypothesis that $\hat{\boldsymbol{\theta}}$ is the true value of $\boldsymbol{\theta}$. We substitute (10) for $p_x(\mathbf{x}|H_0(\mathbf{z}))$ and (11) for $p_z(\mathbf{z}|H_0(\mathbf{z}))$. Under the stated conditions, the exponential terms in approximations (10), and (11) become 1. Using these approximations, we arrive at

$$\hat{p}_x(\mathbf{x}|H_1) = \frac{p_x(\mathbf{x};\hat{\boldsymbol{\theta}})}{(2\pi)^{-P/2}|\mathbf{I}(\hat{\boldsymbol{\theta}})|^{1/2}}\hat{p}_\theta(\hat{\boldsymbol{\theta}}|H_1) \tag{12}$$

which agrees with the PDF approximation from asymptotic theory [13], [14].

To compare (9) and (12), we note that for both, there is an implied sufficiency requirement for $\mathbf{z}$ and $\hat{\boldsymbol{\theta}}$, respectively. Specifically, $H_0(\mathbf{z})$ must remain in the ROS of $\mathbf{z}$, whereas $\hat{\boldsymbol{\theta}}$ must be asymptotically sufficient for $\boldsymbol{\theta}$. However, (9) is more general since (12) is valid only when *all* of the features are ML estimators and only holds asymptotically for large data records with the implication that $\hat{\boldsymbol{\theta}}$ tends to Gaussian, whereas (9) has no such implication. This is particularly important in upstream processing, where there has not been significant data reduction, and asymptotic results do not apply. Using (9), we can make simple adjustments to the reference hypothesis to match the data better and avoid the PDF tails (such as controlling variance), where we are certain that we remain in the ROS of $\mathbf{z}$. As an aside, we note that (7) with a fixed reference hypothesis is even more general since there is no implied sufficiency requirement for $\mathbf{z}$.

D. Chain Rule

In many cases, it is difficult to derive the J-function for an entire processing chain. On the other hand, it may be quite easy to do it for one stage of processing at a time. In this case, the chain rule can be used to good advantage. The chain rule is just the recursive application of the PDF projection theorem. For example, consider a processing chain

$$\mathbf{x} \overset{T_1(\mathbf{x})}{\rightarrow} \mathbf{y} \overset{T_2(\mathbf{y})}{\rightarrow} \mathbf{w} \overset{T_3(\mathbf{w})}{\rightarrow} \mathbf{z}. \tag{13}$$

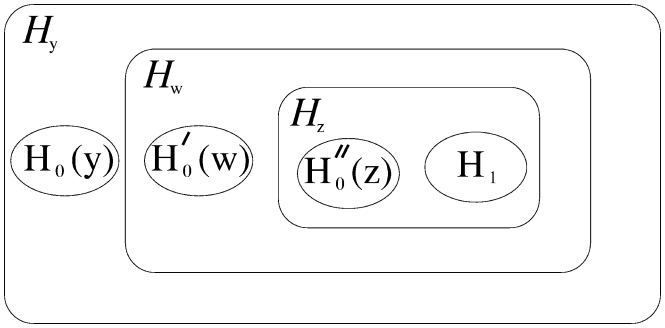

Fig. 1. Required embedding of hypotheses for chain-rule processor corresponding to (13) and (14). The condition $H_1 \in \mathcal{H}_z$ is not necessary for a valid PDF but is desirable for processor optimality.

The recursive use of (7) gives

$$p_x(\mathbf{x}|H_1) = \frac{p_x(\mathbf{x}|H_0(\mathbf{y}))}{p_y(\mathbf{y}|H_0(\mathbf{y}))}\frac{p_y(\mathbf{y}|H_0'(\mathbf{w}))}{p_w(\mathbf{w}|H_0'(\mathbf{w}))} \cdot\frac{p_w(\mathbf{w}|H_0''(\mathbf{z}))}{p_z(\mathbf{z}|H_0''(\mathbf{z}))}p_z(\mathbf{z}|H_1) \tag{14}$$

where $\mathbf{y} = T_1(\mathbf{x})$, $\mathbf{w} = T_2(\mathbf{y})$, $\mathbf{z} = T_3(\mathbf{w})$, and $H_0(\mathbf{y})$, $H_0'(\mathbf{w})$, $H_0''(\mathbf{z})$ are reference hypotheses (possibly data-dependent) suited to each stage in the processing chain. By defining the J-functions of each stage, we may write the above as

$$p_x(\mathbf{x}|H_1) = J(\mathbf{x}, T_1, H_0(\mathbf{y}))J(\mathbf{y}, T_2, H_0'(\mathbf{w})) J(\mathbf{w}, T_3, H_0''(\mathbf{z}))\, p_z(\mathbf{z}|H_1). \tag{15}$$

There is a special embedded relationship between the hypotheses. Let $\mathcal{H}_y$, $\mathcal{H}_w$, and $\mathcal{H}_z$ be the ROSs of $\mathbf{y}$, $\mathbf{w}$, and $\mathbf{z}$, respectively. Then, we have $\mathcal{H}_z \subset \mathcal{H}_w \subset \mathcal{H}_y$. If we use variable reference hypotheses, we also must have $H_0''(\mathbf{z}) \in \mathcal{H}_z$, $H_0'(\mathbf{w}) \in \mathcal{H}_w$, and $H_0(\mathbf{y}) \in \mathcal{H}_y$. This embedding of the hypotheses is illustrated in Fig. 1. The condition $H_1 \in \mathcal{H}_z$ is the ideal situation and is not necessary to produce a valid PDF. The factorization (14), together with the embedding of the hypotheses, we call the chain-rule processor (CRP).

III. Types of J-Functions

We now summarize the various methods we have discussed for computing the J-function.

A. Fixed Reference Hypothesis

For modules using a fixed reference hypothesis, care must be taken in calculation of the J-function because the data is more often than not in the tails of the PDF. For fixed reference hypotheses, the J function is

$$J(\mathbf{x}, T, H_0) = \frac{p_x(\mathbf{x}|H_0)}{p_z(\mathbf{z}|H_0)}. \tag{16}$$

The numerator density is usually of a simple form, so it is known exactly. The denominator density $p_z(\mathbf{z}|H_0)$ must be known exactly or approximated carefully so that it is accurate even in

the far tails of the PDF. The saddlepoint approximation (SPA), which was described in a recent publication [15], provides a solution for cases when the exact PDF cannot be derived but the exact moment-generating function (MGF) is known. The SPA is known to be accurate in the far tails of the PDF [15].

Example 2: As a very simple example of a fixed-reference module, let $\mathbf{x}$ be a time-series, and let $\mathbf{z}$ be the power estimate

$$z = \frac{1}{N}\sum_{n=1}^{N} x_i^2.$$

For H_0 being WGN, $p_x(\mathbf{x}|H_0)$ is quite simple to write, namely

$$\log p_x(\mathbf{x}|H_0) = -\frac{N}{2}\log(2\pi) - \frac{\left(\sum_{n=1}^{N} x_n^2\right)}{2}. \quad (17)$$

Clearly, z is a Chi-square RV with N degrees of freedom scaled by $1/N$. Thus

$$\log p(z|H_0) = \log N - \log\left(\Gamma\left(\frac{N}{2}\right)\right) - \left(\frac{N}{2}\right)\log(2) + \left(\frac{N}{2}-1\right)\log(Nz) - \frac{Nz}{2}. \quad (18)$$

B. Variable Reference Hypothesis Modules

For a variable reference hypotheses, the J function is

$$J(\mathbf{x}, T, H_0(\mathbf{z})) = \frac{p_x(\mathbf{x}|H_0(\mathbf{z}))}{p_z(\mathbf{z}|H_0(\mathbf{z}))}. \quad (19)$$

Modules using a variable reference are usually designed to position the reference hypothesis at the peak of the denominator PDF, which is approximated by the CLT.

Example 3: We can use the Example 2 and redesign the module as a variable reference module. Now, instead of using reference H_0, we use the reference hypothesis $H_0(z)$ that $\mathbf{x}$ has variance $\sigma^2 = z$. Thus

$$\log p_x(\mathbf{x}|H_0(z)) = -\frac{N}{2}\log(2\pi z) - \frac{\left(\sum_{n=1}^{N} x_n^2\right)}{(2z)}. \quad (20)$$

Now, z will still be Chi-square, but we can approximate its PDF by the CLT. Accordingly, z has mean $\sigma^2 = z$ and variance $2\sigma^4/N = 2z^2/N$. Thus

$$\log p(z|H_0(z)) \simeq -\frac{1}{2}\log\left(\frac{4\pi z^2}{N}\right) - \frac{(z-z)^2}{\left(\frac{4z^2}{N}\right)} \simeq -\frac{1}{2}\log\left(\frac{4\pi z^2}{N}\right). \quad (21)$$

Notice the complete cancellation of the last term.

Let us compare the fixed hypothesis method (17) and (18) with the variable hypothesis method (20) and (21) for the power feature. We create input data $\mathbf{x}$ from iid samples of Gaussian noise but with a random scaling. The scale factor was chosen from a uniform distribution in the [0,100] range. The following results were produced.

$\log J$ function		
Fixed ref	Variable ref	error
$-5.658\,667\,5e+03$	$-5.658\,667\,7e+03$	$-0.000\,166\,666$
$-3.559\,442\,1e+03$	$-3.559\,442\,2e+03$	$-0.000\,166\,667$
$-5.228\,754\,2e+03$	$-5.228\,754\,4e+03$	$-0.000\,166\,667$
$-5.186\,465\,0e+03$	$-5.186\,465\,2e+03$	$-0.000\,166\,667$
$-4.969\,499\,2e+03$	$-4.969\,499\,3e+03$	$-0.000\,166\,666$
$-4.184\,531\,1e+03$	$-4.184\,531\,3e+03$	$-0.000\,166\,667$
$-5.693\,948\,5e+03$	$-5.693\,948\,7e+03$	$-0.000\,166\,667$
$-5.656\,036\,5e+03$	$-5.656\,036\,7e+03$	$-0.000\,166\,667$
$-5.691\,540\,8e+03$	$-5.691\,541\,0e+03$	$-0.000\,166\,667$
$-5.267\,565\,5e+03$	$-5.267\,565\,6e+03$	$-0.000\,166\,667$

There is almost no difference between the approaches (a 0.000 16 error in log domain). The error rises as N decreases because the CLT approximation worsens.

C. Maximum Likelihood Modules

A special case of the variable reference hypothesis approach is the ML method, when $\mathbf{z}$ is an MLE (see Section II-C)

$$J(\mathbf{x}, T, H_0) = \frac{p(\mathbf{x}|\hat{\boldsymbol{\theta}})}{(2\pi)^{-P/2}|\mathbf{I}(\hat{\boldsymbol{\theta}})|^{1/2}}.$$

To continue Examples 2 and 3, it is known that the ML estimator for variance is the sample variance which has a Cramér–Rao (CR) bound of $\sigma^2_{\min} = 2\sigma^4/N$. Applying (12), we get exactly the same result as the above variable reference approach. Whenever the feature is also a ML estimate and the asymptotic results apply (the number of estimated parameters is small and the amount of data is large), the two methods are identical. The variable reference hypothesis method is more general because it does not need to rely on the CLT.

D. One-to-One Transformations

One-to-one transformations do not change the information content of the data, but they are important for feature conditioning prior to PDF estimation. Recall from Section II that Theorem 1 is a generalization of the change-of-variables theorem for 1:1 transformations. Thus, for 1:1 transformations, the J-function reduces to the absolute value of the determinant of the Jacobian matrix (4)

$$J(\mathbf{x}, T) = |\mathbf{J}_T(\mathbf{x})|.$$

Our first example is the log transformation that is useful when applied to exponential RVs to obtain a more "Gaussian-like" distribution.

Example 4—Log Transformation: Let $z = \log(x)$. We have $dy/dx = 1/x$; thus, $\log J = \log(1/x) = -\log x = -z$. For vector arguments

$$\log J = -\sum_{i=1}^{N} z_i.$$

A very important one-to-one transformation in signal processing is the conversion from autocorrelation function (ACF) to reflection coefficients (RCs) using the Levinson algorithm [16]. RCs tend to be better features since they are less correlated than ACF estimates.

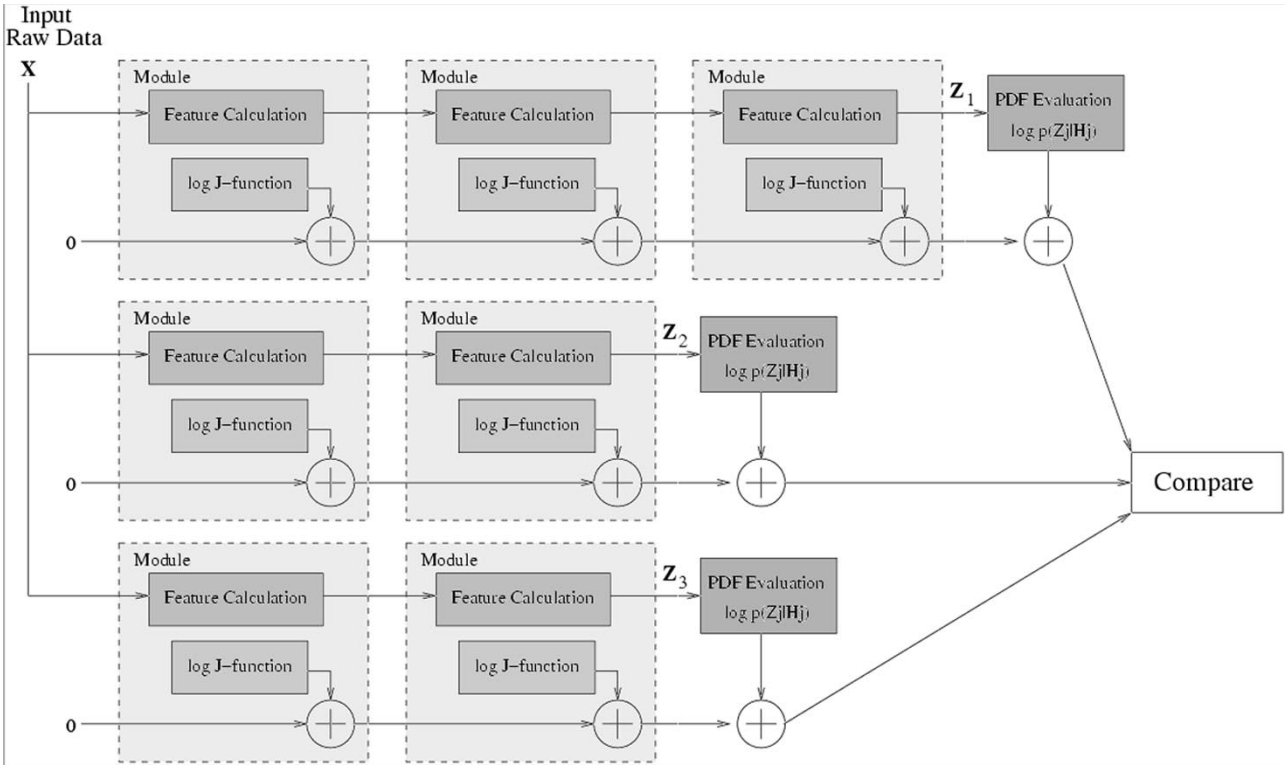

Fig. 2. Block diagram of a class-specific classifier.

Example 5—Conversion From ACF to RCs: Let $\mathbf{z} = T(\mathbf{r})$, where $\mathbf{r} = [r_0, r_1 \dots r_P]'$ and $\mathbf{z} = [r_0, k_1 \dots k_P]'$, where $\{k_1 \dots k_P\}$ are the P RCs. The Jacobian is

$$|\mathbf{J}_T| = r_0^{-P} \prod_{i=1}^{P-1} (1 - k_i^2)^{P-i}.$$

Although the RCs are uncorrelated, they are subject to the limit $|k_i| < 1$, which gives their distribution a discontinuity. To obtain more Gaussian behavior, the log-bilinear transformation is recommended (thanks to S. Kay).

Example 6—Log Bilinear Transformation: Let

$$k_i' = \frac{\log(1 - k_i)}{\log(1 + k_i)},\ 1 \le i \le P.$$

We have

$$|\mathbf{J}_T| = \prod_{i=1}^{P} \frac{(1 - k_i^2)}{2}.$$

Additionally, taking the log of the first feature (r_0) results in a further improvement.

IV. Application of Theorem 1 to Classification

A. Classifier Architecture

Application of the PDF projection theorem to classification is simply a matter of substituting (9) into (1). In other words, we implement the classical Neyman–Pearson classifier but with the class PDFs factored using the PDF projection theorem

$$j^* = \arg\max_j \frac{p_x(\mathbf{x}|H_{0,j}(\mathbf{z}_j))}{p_z(\mathbf{z}_j|H_{0,j}(\mathbf{z}_j))} \hat{p}_z(\mathbf{z}_j|H_j) \text{ at } \mathbf{z}_j = T_j(\mathbf{x}) \tag{22}$$

where we have allowed for class-dependent, data-dependent, reference hypotheses.

The chain-rule processor (14) is ideally suited to classifier modularization. Fig. 2 is a block diagram of a class-specific classifier. The packaging of the feature calculation together with the J-function calculation is called the class-specific module. Each arm of the classifier is composed of a series of modules called a "chain."

B. Feature Selectivity: Classifying Without Training

The J-function and the feature PDF provide a factorization of the raw data PDF into trained and untrained components. The ability of the J-function to provide a "peak" at the "correct" feature set gives the classifier a measure of classification performance without needing to train. In fact, it is not uncommon that the J-function dominates, eliminating the need to train at all. This we call the *feature selectivity effect*. For a fixed amount of raw data, as the dimension of the feature set decreases, indicating a larger rate of data compression, the effect of the J-function compared with the effect of the feature PDF increases. An example where the J-function dominates is a bank of matched filter for known signals in noise. If we regard the matched filters as feature extractors and the matched filter outputs as scalar features, it may be shown that this method is identical to comparing

only the J-functions. Let $z_j = |\mathbf{w}_j'\mathbf{x}|^2$, where $\mathbf{w}_j$ is a normalized signal template such that $\mathbf{w}_j'\mathbf{w}_j = 1$. Then, under the white (independent) Gaussian noise (WGN) assumption, z_j is distributed $\chi^2(1)$. It is straightforward to show that the J-function is a monotonically increasing function of z_j. Signal waveforms can be reliably classified using only the J-function and ignoring the PDF of z_j under each hypothesis. The curse of dimensionality can be avoided if the dimension of $\mathbf{z}_j$ is small for each j. This possibility exists, even in complex problems, because $\mathbf{z}_j$ is required only to have information sufficient to separate class H_j from a specially chosen reference hypothesis $H_{0,j}$.

C. J-Function Verification

One thing to keep in mind is that it is of utmost importance that the J-function is accurate because this will insure that the resulting projected PDF is, in fact, a valid PDF. For example, if the J-function is accidentally scaled by a large positive constant, the classifier will produce false classifications in favor the the class with the erroneous J-function. In contrast, it is not a serious problem, however, if one of the likelihood functions $\hat{p}(\mathbf{x}|H_j)$ is not a perfect match to the data for class H_j because it will be discovered by trial and error. A better PDF estimate can be found simply by comparing the likelihood values for the given class. Therefore, in the following examples, we are *not* very strictabout the sufficiency of the features for the corresponding target class, although their approximate sufficiency is intuitively apparent. The ultimate justification for using a particular feature set can be the maximization of the likelihood values calculated on the raw data space using

$$\hat{p}(\mathbf{x}|H_j) = J(\mathbf{x}, T_j, H_0)p(\mathbf{z}_j|H_j). \tag{23}$$

We can compare competing feature sets based on the likelihood values and can gradually increase the likelihood on the target class by experimenting with different features and PDF models.

To verify the "J" function, we have developed an end-to-end test that we call the "Acid Test" because of its foolproof nature. To use the method, it is first necessary to define a fixed hypothesis, which is denoted by H_s, for which we can compute the PDF $p(\mathbf{x}|H_s)$ readily and for which we can synthesize raw data. Note that H_s is *not* a reference hypothesis. The synthetic data is converted into features, and the PDF $\hat{p}(\mathbf{z}|H_s)$ is estimated from the synthetic features (using a Gaussian Mixture PDF, HMM, or any appropriate statistical model). Next, the theoretical PDF $p(\mathbf{x}|H_s)$ is compared with the projected PDF

$$\hat{p}(\mathbf{x}|H_s) = J(\mathbf{x}, T, H_0)\ \hat{p}(\mathbf{z}|H_s)$$

for each sample of synthetic data. The log-PDF values are plotted on each axis, and the results should fall on the $X = Y$ line. For each example, we will provide acid test results. Since the acid test checks the equality of two entirely different paths, it should find any systematic error in PDF estimation or in the J-function calculation.

V. Example: Linear Functions of Exponential, Chi-Square, or Log-Exponential RVs

A widely used combination of transformations in signal processing is to first apply an orthogonal linear transformation, perform a squaring operation (or magnitude-squared for complex RVs), and then perform a linear transformation. These transformations include widely used features such as MEL cepstrum [17], polynomial fits to power series and power spectra, autocorrelation functions and, through one-to-one transformations, autoregressive (AR) and reflection coefficients (RC).

The general form is the following. Let $\mathbf{x}$ be an N-by-1 real or complex vector. Let $\mathbf{u} = \mathbf{U}^H\mathbf{x}$ be some real or complex orthogonal linear transformation such that $\mathbf{U}^H\mathbf{U} = v\mathbf{I}$. Note that $\mathbf{U}$ does not need to be square if $\mathbf{x}$ is real and $\mathbf{u}$ is complex since we omit any redundant elements of $\mathbf{u}$. Let n be the length of $\mathbf{u}$. For the case of DFT of a real vector, $n = N/2 + 1$. Next, let $\mathbf{y}$ be the vector whose elements are the magnitude squared (if complex) or squared (if real) values of the elements of $\mathbf{u}$

$$y_i = |u_i|^2,\ 0 \le i \le n-1.$$

Finally, let

$$\mathbf{z} = \mathbf{A}'\mathbf{y} \tag{24}$$

where $\mathbf{A}$ is a real n-by-M matrix.

A. Two Approaches to Computing the J-Function

For the features in (24), there is no closed-form solution to the J-function, except in some simple cases [15]. There are, however, two very good approximations discussed in the next sections. The second method (central limit theorem) will be used in the subsequent example.

1) Saddlepoint Approximation Method: The saddlepoint approximation (SPA) was discussed in a previous publication [15]; therefore, we will only give an overview here. In the referenced paper, the case of autocorrelation coefficients computed from a real vector $\mathbf{x}$ was discussed. The reference hypothesis used for this approach, which is denoted by H_0, is white (independent) Gaussian noise of zero mean and variance 1. The numerator PDF of the J-function

$$J(\mathbf{x}, T, H_0) = \frac{p_x(\mathbf{x}|H_0)}{p_z(\mathbf{z}|H_0)} \tag{25}$$

is known exactly, and the denominator PDF is approximated by the SPA. In extreme cases, this approach can potentially suffer from the "tail PDF problem."

To appreciate the tail PDF problem, one can imagine that for a given sample $\mathbf{x}$, as we scale $\mathbf{x}$ by a large positive number K so that when calculating $J(K\mathbf{x}, T, H_0)$, we will quickly reach a point where the J-function will be a ratio of two numbers that are essentially zero and cannot be reliably computed. In practice, we find that if all calculations are made in the log-domain, the $\log$-J function is well-behaved for very large input values. There are limits, however, and we find that the SPA, which is a

recursive search for the saddlepoint itself, will eventually have convergence problems. To alleviate this problem, we use a variable reference hypothesis (see Section II-B). Let v be a rough estimate of the variance of $\mathbf{x}$. Let H_v be the hypothesis that the input variance equals v. Assuming H_v and H_0 are in the ROS of $\mathbf{z}$, (25) is theoretically independent of H_v, and thus

$$\frac{p_x(\mathbf{x}|H_0)}{p_z(\mathbf{z}|H_0)} = \frac{p_x(\mathbf{x}|H_v)}{p_z(\mathbf{z}|H_v)}.$$

However

$$p_z(\mathbf{z}|H_v) = v^{-M} p_z(v^{-1}\mathbf{z}|H_0)$$

where M is the dimension of $\mathbf{z}$. Therefore

$$J(\mathbf{x}, T, H_0) = v^M \frac{p_x(\mathbf{x}|H_v)}{p_z(\mathbf{z}/v|H_0)} \tag{26}$$

which provides a convenient way to normalize $\mathbf{z}$ prior to calculating the SPA.

2) CLT Method: The second method that gives us a workable solution is the CLT. We use the chain rule to separately analyze the two stages: a) orthogonal transformation and squaring and b) linear transformation. We will design a two-module chain for a subset of the autocorrelation function (ACF) estimates. The processing chain necessary to compute the ACF coefficients can be broken down into two stages:

1) Compute $\mathbf{y}$, which are the magnitude-squared FFT bins.
2) Compute $\mathbf{z}$, which is a subset of the elements of IFFT $(\mathbf{y})$, which is the real part of the inverse FFT of $\mathbf{y}$.

B. Structure of the Examples

As explained previously, a class-specific classifier can be organized into "modules". Each module consist of a feature transformation and a J-function calculation. The J-function requires the definition of a reference hypothesis and the calculation of the numerator (input) and denominator (output) PDF. Accordingly, we organize this example and those that follow into modules. For each module, we explain the following.

1) **Features and ROS**. We describe the feature transformation $\mathbf{z} = T(\mathbf{x})$ and the ROS for the features (see Section II-B). Ideally, the ROS, which is denoted by $\mathcal{H}_z$, includes the "target class" H_1 for which this feature set is designed.
2) **Reference Hypothesis**. We define the reference hypothesis H_0 used in the J-function. Often, this hypothesis is a data-dependent reference, which is written $H_0(\mathbf{z})$.
3) **Input PDF**. We define this as the numerator of the J function.
4) **Output PDF**. This is the denominator of the J-function.
5) **Test Results**. When appropriate, we present results of the "acid test" (Section IV-C).

C. Stage 1: DFT Magnitude-Squared

Stage 1 of the two-stage CLT approach is discussed here.

1) Features and Region of Sufficiency: Let $\mathbf{y}$ be the length $N/2+1$ vector of magnitude-squared bins of the DFT of $\mathbf{x}$.

$$\mathbf{y} = [y_0, y_1 \cdots y_{N/2}]'$$

where

$$y_k = \left| \sum_{i=1}^{N} x_i e^{-j2\pi ki/N} \right|^2, \quad 0 \le k \le \frac{N}{2}.$$

The ROS of $\mathbf{y}$ is quite broad, encompassing all Gaussian processes with a power spectrum.

2) Reference Hypothesis: For our reference hypothesis for this stage, we use H_0, which is the standard normal density (WGN hypothesis with unit variance).

3) Input PDF: We have

$$p_x(\mathbf{x}|H_0) = (2\pi)^{-N/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} x_i^2 \right\}. \tag{27}$$

4) Output PDF: Note that under H_0, $\mathbf{y}$ is a set of independent RVs. It is easily shown that $y_0, y_{N/2}$ obey the $\chi^2(1)$ density with mean N and variance $2N^2$. In addition, $y_1 \cdots y_{N/2-1}$ obey the $\chi^2(2)$ or exponential density with mean N and variance N^2. Thus

$$p_y(\mathbf{y}|H_0) = \prod_{i=0}^{N/2} p_y(y_i|H_0) \tag{28}$$

where

$$p_y(y_i|H_0) = \frac{1}{v\sqrt{2\pi}} \left(\frac{y_i}{v}\right)^{-1/2} \exp\left\{ -\frac{y_i}{2v} \right\} \quad i = 0, \frac{N}{2} \tag{29}$$

and

$$p_y(y_i|H_0) = \frac{1}{v} \exp\left\{ -\frac{y_i}{v} \right\}, \quad 1 \le i \le \frac{N}{2} - 1 \tag{30}$$

and v is the mean of the elements of $\mathbf{y}$ ($v = N$).

D. Stage 2: Linear Transformation

Stage 2 of the two-stage CLT approach is discussed here. In stage 2, we apply a linear transformation to $\mathbf{y}$. We use ACF as an example, but the basic method applies to any linear transformation.

1) Features and Region of Sufficiency: We let $\mathbf{z}$ be the first $P+1$ circular ACF samples

$$z_i = \frac{1}{N} \sum_{n=1}^{N} x_n x_{[n+i]} \quad 0 \le i \le P \tag{31}$$

where $[n+i]$ is taken modulo-N. We use the circular ACF estimates in this example for simplicity because they may be written in terms of $\mathbf{y}$, but the J-function may be found for any variety

of ACF estimate. The features (31) may be written in terms of $\mathbf{y}$ as follows:

$$z_i = \frac{1}{N^2}\sum_{k=0}^{N/2} y_k \cos\{\frac{j2\pi ki}{N}\}, \quad 0 \le i \le P. \tag{32}$$

This has a compact matrix notation

$$\mathbf{z} = \mathbf{A}'\mathbf{y}$$

where $\mathbf{A}$ is the $(N/2+1)$-by-$(P+1)$ matrix defined by

$$A_{ij} = \frac{2}{N^2}\cos\left(\frac{2\pi ij}{N}\right), \quad 0 \le j \le P,\ 1 \le i \le \frac{N}{2}-1 \tag{33}$$

$$A_{ij} = \frac{1}{N^2}\cos\left(\frac{2\pi ij}{N}\right), \quad 0 \le j \le P,\ i = 0, \frac{N}{2}. \tag{34}$$

Since $\mathbf{z}$ is the ACF estimates of order P, the approximate ROS is all AR processes of order P and less.

2) Reference Hypothesis: Because we intend to use the CLT to approximate the J-function denominator, we need to use a variable reference hypothesis $H_0(\mathbf{z})$ such that the mean of $\mathbf{z}$ under $H_0(\mathbf{z})$ is equal to or close to $\mathbf{z}$ itself. There are two possible methods. For arbitrary matrices $\mathbf{A}$, this can be done by projecting the input vector upon the column space of $\mathbf{A}$. Let $H_0(\mathbf{z})$ be the hypothesis that $\mathbf{y}$ has mean

$$\mathbf{y}^z = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{z}. \tag{35}$$

Notice that $\mathbf{A}'\,\mathbf{y}^z = \mathbf{z}$, that is, under $H_0(\mathbf{z})$, the mean of $\mathbf{z}$ is $\mathbf{z}$ itself.

One possible problem that can occur is if $\mathbf{y}^z$ in (35) happens to be negative, which is quite possible, but not allowed. A suitable solution is to use a constrained optimization, that is, choose $\mathbf{y}^z$ such that it is positive, and $\mathbf{A}'\,\mathbf{y}^z$ is as close as possible to $\mathbf{y}$.

A more satisfying way to guarantee a positive $\mathbf{y}^z$ in the case of ACF is the following. We let $H_0(\mathbf{z})$ be the hypothesis that $\mathbf{y}$ obeys the AR spectrum corresponding to $\mathbf{z}$. Thus, we must use the Levinson algorithm to solve for the Pth-order AR coefficients σ_0^2, $\mathbf{a}$. If $A(k)$ is the DFT of $\mathbf{a}$ (padded to length N), then

$$P^z(k) = \frac{\sigma_0^2}{|A(k)|^2}$$

is the AR spectrum corresponding to $\mathbf{a}$. We let

$$\mathbf{y}^z \triangleq [P^z(0)\ldots P^z\left(\frac{N}{2}\right)]'. \tag{36}$$

For large N, we have $\mathbf{A}'\,\mathbf{y}^z \to \mathbf{z}$.

3) Input PDF: We need to evaluate $p_y(\mathbf{y}|H_0(\mathbf{z}))$ and $p_z(\mathbf{z}|H_0(\mathbf{z}))$. We assume $\{y_i\}$ are a set of independent exponential and χ^2 RVs with means corresponding to $\{y_i^z\}$, which are the elements of $\mathbf{y}^z$. Specifically

$$p_y(y_i|H_0(\mathbf{z})) = \frac{1}{y_i^z\sqrt{2\pi}}\left(\frac{y_i}{y_i^z}\right)^{-1/2} \times \exp\left\{-\frac{y_i}{2y_i^z}\right\} \quad i = 0, \frac{N}{2} \tag{37}$$

and

$$p_y(y_i|H_0(\mathbf{z})) = \frac{1}{y_i^z}\exp\left\{-\frac{y_i}{y_i^z}\right\},\ 1 \le i \le \frac{N}{2}-1. \tag{38}$$

4) Output PDF: Because $H_0(\mathbf{z})$ is "close" to $\mathbf{z}$, we approximate $p_z(\mathbf{z}|H_0(\mathbf{z}))$ by the central limit theorem (CLT). Under $H_0(\mathbf{z})$, the elements of $\mathbf{y}$ are independent with mean $\mathbf{y}^z$ and diagonal covariance $\mathbf{\Sigma}_y^z$, which are defined by

$$\mathbf{\Sigma}_y^z(i,i) \triangleq \mathcal{E}((y_i - y_i^z)^2|H_0(\mathbf{z})) = \begin{cases} 2(y_i^z)^2, & i = 0, \frac{N}{2} \\ (y_i^z)^2, & 1 \le i \le \frac{N}{2}-1. \end{cases}$$

We can then easily compute the mean and covariance of $\mathbf{z}$:

$$\mathbf{z}^z = \mathcal{E}(\mathbf{z}|H_0(\mathbf{z})) = \mathbf{A}'\mathbf{y}^z$$

and

$$\begin{aligned}\mathbf{\Sigma}_z^z &= \mathbf{A}'\mathbf{\Sigma}_y^z\mathbf{A}. \\ \log p_z(\mathbf{z}|H_0(\mathbf{z})) &= -\frac{(P+1)}{2}\log(2\pi) - \frac{1}{2}\log|\det(\mathbf{\Sigma}_z^z)| \\ &\quad - \frac{1}{2}(\mathbf{z}-\bar{\mathbf{z}}_{0z})'(\mathbf{\Sigma}_z^z)^{-1}(\mathbf{z}-\bar{\mathbf{z}}_{0z}) \\ &\simeq -\frac{(P+1)}{2}\log(2\pi) - \frac{1}{2}\log|\det(\mathbf{\Sigma}_z^z)|\end{aligned} \tag{39}$$

where in the last step, we make the approximation $\mathbf{z}^z \simeq \mathbf{z}$. This approximation becomes better as N becomes larger. Note also that the method just described is closely related to the ML approach. In fact, $\mathbf{\Sigma}_z^z$ is related to the Fisher's information of the ACF estimates [16].

E. Test Results

The acid test was run on the ACF features using both the SPA and CLT methods. A model order of 2 was used giving a feature dimension of 3 (lags 0 through 2). Results are shown in Figs. 3 and 4. A raw data size of $N = 32$ was used with a test hypothesis, H_s of iid Gaussian noise of variance 100. There were 400 samples of synthetic data used for training the feature PDF using a Gaussian mixture. The results show that both methods "pass" the test because the estimates of projected PDF (vertical axis) appear to track the theoretical PDF values (horizontal axis). The errors are quite small, considering that these are PDF estimates of a 32-dimensional PDF. A comparison was made of the difference of the log J-function values output by the two methods, and it was found that the difference was less than 1.0 for all samples.

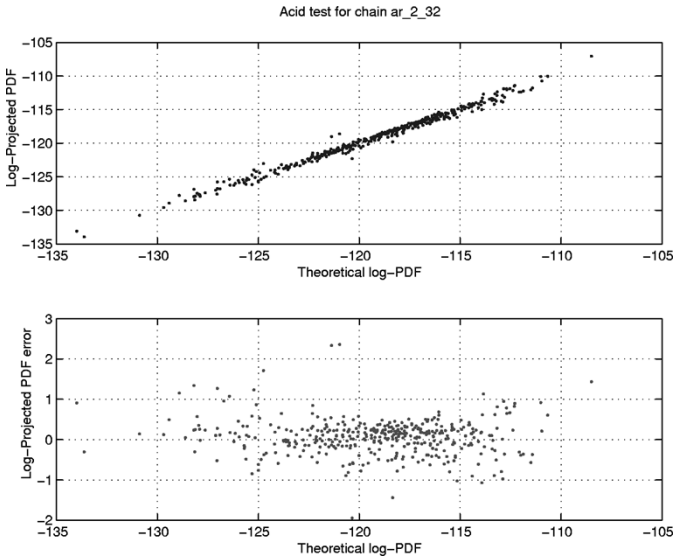

Fig. 3. Acid test results for autocorrelation function using SPA method. Top frame shows the estimate of the feature log-PDF projected to the raw data plotted against the theoretical log-PDF. The bottom frame shows the difference plotted against the theoretical log-PDF. A Gaussian mixture was used to estimate the feature PDF.

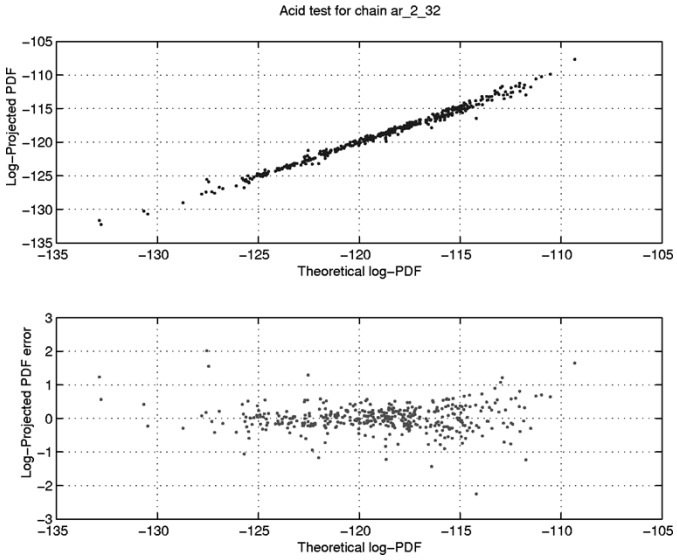

Fig. 4. Acid test results for autocorrelation function using CLT method. Top frame shows the estimate of the feature log-PDF projected to the raw data plotted against the theoretical log-PDF. The bottom frame shows the difference plotted against the theoretical log-PDF. A Gaussian mixture was used to estimate the feature PDF.

VI. Example: Cepstrum and MEL Cepstrum

An important set of features in speech analysis is cepstrum [18] and MEL cepstrum [17]. For the cepstrum, the SPA for the denominator PDF of the J-function for fixed WGN reference hypothesis is described in [15], so we will not need to discuss it further. The MEL cepstrum, however, is a member of the set of transformations in Section V. The MEL cepstrum is computed as follows. Let $\mathbf{y}$ be a DFT magnitude-squared vector of length $N/2-1$, where N is the DFT size. The MEL filter bank is a matrix $\mathbf{A}$ of spectral template vectors

$$\mathbf{A}=[\mathbf{c}_1\mathbf{c}_2\ldots\mathbf{c}_M]$$

where $M=12$ is a commonly used value. The MEL cepstrum equals

$$\mathbf{z}=\mathrm{DCT}(\log(\mathbf{A}'\mathbf{y}))$$

where the log function operates on each element of its argument, and DCT is the discrete cosine transform. Note that DCT and log are both 1:1 transformations, whose J-functions are the determinant of the respective Jacobian matrices. Our primary concern, then, is to analyze the intermediate feature set

$$\mathbf{w}=\mathbf{A}'\mathbf{y}$$

which we have previously described in Section V.

An important warning is that the usual MEL filter bank does not include filters centered at the 0-th and $N/2$ (Nyquist) DFT bins. These two filters need to be included in any class-specific classifier; otherwise, $\mathbf{z}$ will not be sufficient for simple scaling operations. The proper way to eliminate the features is not to exclude them from the MEL filterbank, but rather to assign a noninformative (such as uniform) PDF to them at the output.

VII. Feature Selection

One question we have not yet covered is how does one determine an appropriate feature set for a data class? Choosing features is rarely done through statistical or mathematical analysis. The choice of features remains an art requiring intuition. This intuition is is helped by the methods of resynthesis and model order/segment size selection discussed below.

A. Sufficiency by Resynthesis

In many problems, the ability of a human to classify an event exceeds the ability of the machine. Human performance is almost always a lofty goal. It is therefore reasonable to choose features that can represent the data with enough fidelity to resynthesize the event to the satisfaction of a human observer. For example, the resynthesis of speech data from features has been used for speech analysis to determine the appropriateness of speech analysis methods [19]. We recommend this method whenever it is appropriate.

B. Determination of Segmentation and Model Order

Once a feature set is chosen, it may be possible to fine tune it. This is particularly true if the feature extraction is governed by a set of parameters such as segment size and model order. Full implementation of (7a) may be computationally prohibitive unless a simplified PDF model is used. The method now presented may be a way to automatically determine these parameters.

In many statistical models, there are two parts to the modeling: measurement PDF and spatio-temporal distribution. For example, in an HMM, the state PDFs are measurement PDFs and the state transition matrix describes the spatio-temporal component of the model. By removing the spatio-temporal part of the model, a simplified model results (just a measurement PDF). It may be possible to optimize the feature model order and segmentation based only on the simplified model. The optimized features, it is conjectured, would achieve the highest likelihood once the spatio-temporal parts of the model were

restored. We have conducted many experiments that support this conjecture.

The particulars of the method are now presented. Let the feature data be written $\mathbf{Z}^k = \{\mathbf{z}_1^k, \mathbf{z}_2^k \ldots \mathbf{z}_{N_k}^k\}$, where k is a particular choice of segment size and/or model order and N_k is the corresponding total number of observation vectors corresponding to choice k. Note that we have collected all the available data from all events into one mass, forgetting the temporal or spatial organization, and forgetting which event the observations are from. We also assume that $\mathbf{z}_n^k$ are low enough in dimension that a parametric PDF estimator (i.e., Gaussian mixture) can be estimated from the data. Let the data be divided into a training set ($\mathbf{X}_{\rm tr}$, $\mathbf{Z}_{\rm tr}^k$) and testing set ($\mathbf{X}_{\rm te}$, $\mathbf{Z}_{\rm te}^k$) for cross-validation. Next, we estimate the PDF

$$\hat{p}_k(\mathbf{z}^k)$$

using $\mathbf{Z}_{\rm tr}^k$ for model choice k. The feature PDF is projected to the input data space where it can be compared across different values of k. We have

$$L(k) = \log J(\mathbf{X}, \mathbf{Z}^k) + \sum_n \log \hat{p}_k(\mathbf{z}_n^k)$$

where $\log J(\mathbf{X}, \mathbf{Z}^k)$ is the aggregate log-J-function for the data set. Next, $L(k)$ is calculated for ($\mathbf{X}_{\rm te}$, $\mathbf{Z}_{\rm te}^k$). For added accuracy, $L(k)$ can also be computed by swapping $\mathbf{Z}_{\rm tr}^k$ and $\mathbf{Z}_{\rm te}^k$ and averaging. The optimal choice of k is that which maximizes $L(k)$.

This approach is robust against overparameterization because as the model order (and dimension of $\mathbf{z}^k$) increases above the optimal value, the ability to estimate the PDF worsens and the average of the cross-validated likelihood will begin to fall.

C. Example

To test the approach, we first created a synthetic signal class approximating a "bang" sound. Independent Gaussian noise is passed through a second-order autoregressive filter. The filter output is modulated by an envelope function with an instantaneous attack and an exponential decay. The attack time is chosen at random. Independent noise is added to the result. An example of a typical synthetic event is shown in Fig. 5. A total of 100 events were created, each with a total length of 4096 samples. The features were extracted by segmenting the events into segments of length N, where N ranged from 32 to 512 in powers of 2. Autocorrelation features of order P were extracted from each segment where P was between 2 and 7. The results are shown in Table I and show a peak at $P = 4$, $N = 128$, which is about a 10-ms segment size. This is in agreement with intuition because the width of the event envelope near the peak is about 10 ms.

VIII. Versatile General-Purpose Class-Specific Time-Series Classifier Using Reflection Coefficients and HMM

It is possible to use the material thus-far discussed to arrive at a fully modular, extremely versatile class-specific classifier. A functional block-diagram of this classifier is provided in Fig. 6.

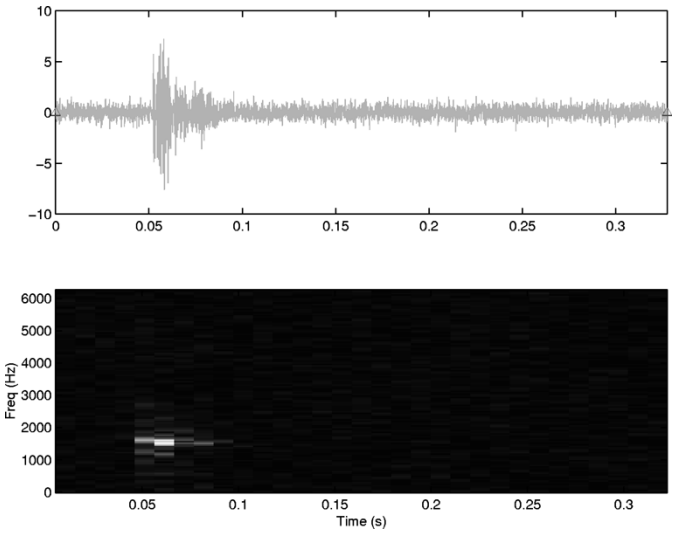

Fig. 5. Example of a typical synthetic event. Time-series (top) and spectrogram (bottom). Sample rate was 12 500 Hz.

TABLE I
Results of Model Order/Segment Size Selection Experiment. Results Log-Likelihoods Relative to Maximum

P	N=32	N=64	N=128	N=256	N=512
2	-7782.6	-5105.3	-5027.9	-8336.7	-16748.1
3	-6590.3	-2033.2	-921.5	-2820.3	-8650.7
4	-7491.6	-1716.7	0.0	-1550.6	-6672.8
5	-9316.0	-2127.8	-119.7	-1499.8	-6546.5
6	-11691.5	-2731.7	-262.8	-1540.3	-6480.7
7	-16159.1	-5437.2	-2169.5	-3164.1	-8080.1

A given time-series is processed by each class-model to arrive at a raw-data log-likelihood for the class. Each block labeled "RC(P)" computes the reflection coefficients of order P from the associated time-series segment. The figure shows two class-models employing different segmentation lengths as well as different model orders. The log-correction terms ($\log J$-functions) of all the segments are added together and the aggregate correction term is added to the HMM log-likelihood (from the forward procedure [20]) to arrive at the final raw data log-likelihood for the class.

Each "RC(P)" block is composed of a series of modules implementing ACF calculation followed by conversion to RCs and ending with feature conditioning by the log-bilinear transformation. This may be implemented four modules corresponding to Sections V-C, V-D, and III-D (Examples 5 and 6). Alternatively, the SPA approach (see Section V-A1) may be used in place of the first two modules and will produce virtually identical features and J-function values. This classifier has the added benefit that the models may be validated by re-synthesis of time-series from features (either computed from actual data or generated at random by the HMM). Using the method of Section VII-B, the segmentation sizes and model orders may be optimized for each

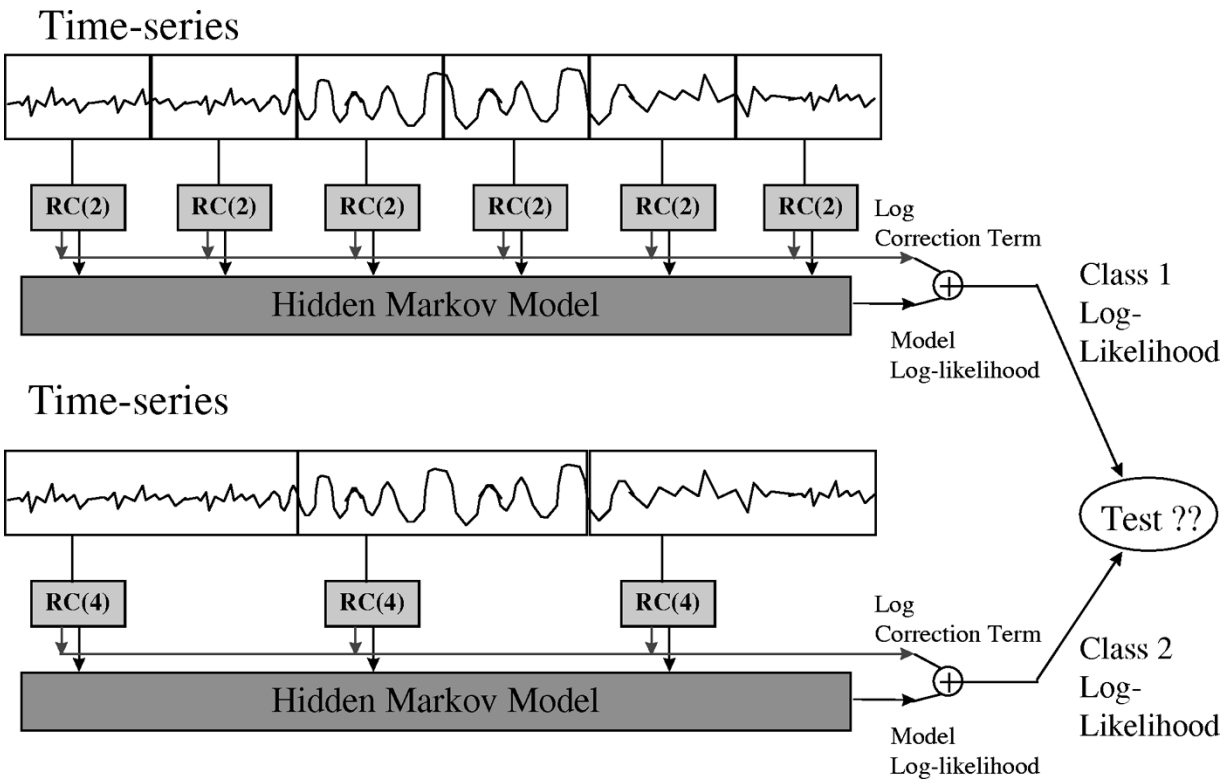

Fig. 6. Block diagram of an HMM and RC-based class-specific classifier. A given time-series is processed by each class-model to arrive at a raw-data log-likelihood for the class. Each block labeled "RC(P)" computes the Pth order reflection coefficients from the corresponding time-series segment and is implemented by a series of modules.

class individually, eliminating the need to "compromise," and, because it is a class-specific classifier, features of any kind may be used. Adding new class processors will not affect the existing class processors or their training.

IX. Conclusions

We have introduced a powerful new theorem that opens up a wide range of new statistical methods for signal processing, parameter estimation, and hypothesis testing. Instead of needing a common feature space for likelihood comparisons, the theorem allows likelihood comparisons to be made on a common raw data space, while the difficult problem of PDF estimation can be accomplished in separate feature spaces. We have discussed the recursive application of the theorem which gives a hierarchical breakdown and allows processing streams to be analyzed in stages. Whereas previous publications on the method have relied on a common fixed reference hypothesis, this paper has presented the use of class-dependent and data-dependent reference hypotheses and has explored the relationship to asymptotic maximum likelihood theory. The use of a data-dependent reference hypothesis allows two new methods of analyzing the feature sets – maximum likelihood (ML) and central limit theorem (CLT). These extensions significantly broaden the applicability of the method. We have illustrated the use of the approach using common feature types including autoregressive and MEL cepstrum features. We have also presented a method of combined feature/model order selection that is enabled by the class-specific approach. Finally, we have provided an example of a versatile class-specific classifier using autoregressive features.

References

[1] P. Baggenstoss, "Class-specific features in classification," *IEEE Trans. Signal Processing*, vol. 47, pp. 3428–3432, Dec. 1999.

[2] S. M. Kay, "Sufficiency, classification and the class-specific feature theorem," *IEEE Trans. Inform Theory*, vol. 46, pp. 1654–1658, July 2000.

[3] Frimpong-Ansah, K. Pearce, D. Holmes, and W. Dixon, "A stochastic/feature based recognizer and its training algorithm," in *Proc. ICASSP*, vol. 1, 1989, pp. 401–404.

[4] S. Kumar, J. Ghosh, and M. Crawford, "A versatile framework for labeling imagery with large number of classes," in *Proc. Int. Joint Conf. Neural Networks*, Washington, DC, 1999, pp. 2829–2833.

[5] ——, "A hierarchical multiclassifier system for hyperspectral data analysis," in *Multiple Classifier Systems*, J. Kittler and F. Roli, Eds. New York: Springer, 2000, pp. 270–279.

[6] H. Watanabe, T. Yamaguchi, and S. Katagiri, "Discriminative metric design for robust pattern recognition," *IEEE Trans. Signal Processing*, vol. 45, pp. 2655–2661, Nov. 1997.

[7] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 711–720, July 1997.

[8] D. Sebald, "Support vector machines and the multiple hypothesis test problem," *IEEE Trans. Signal Processing*, vol. 49, pp. 2865–2872, Nov. 2001.

[9] I.-S. Oh, J.-S. Lee, and C. Y. Suen, "A class-modularity for character recognition," in *Proc. Int. Conf. Document Anal. Recognition*, Seattle, WA, Sept. 2001, pp. 64–68.

[10] E. Sali and S. Ullman, "Combining class-specific fragments for object classification," in *Proc. British Machine Vision Conf.*, 1999, pp. 203–213.

[11] P. M. Baggenstoss, "A modified Baum-Welch algorithm for hidden Markov models with multiple observation spaces," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 411–416, May 2001.

[12] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. London, U.K.: Chapman and Hall, 1974.

[13] R. L. Strawderman, "Higher-order asymptotic approximation: Laplace, saddlepoint, and related methods," *J. Amer. Statist. Assoc.*, vol. 95, pp. 1358–1364, Dec. 2000.

[14] J. Durbin, "Approximations for densities of sufficient estimators," *Biometrika*, vol. 67, no. 2, pp. 311–333, 1980.
[15] S. M. Kay, A. H. Nuttall, and P. M. Baggenstoss, "Multidimensional probability density function approximation for detection, classification and model order selection," *IEEE Trans. Signal Processing*, pp. 2240–2252, Oct. 2001.
[16] S. Kay, "Modern spectral estimation," in *Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1998.
[17] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, pp. 1215–1247, Sept. 1993.
[18] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 221–226, 1968.
[19] C. Bell, H. Fujisaki, J. Heinz, K. Stevens, and A. House, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Amer.*, pp. 1725–1736, Dec. 1961.
[20] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

Paul M. Baggenstoss (M'90) received the Ph.D. degree in electrical engineering (statistical signal processing) from the University of Rhode Island (URI), Kingston, in 1990.

Since then, he has been applying signal processing and hypothesis testing (classification) to sonar problems. From 1979 to 1996, he was with Raytheon Co, Portsmouth, RI. Since 1996, he has been with the Naval Undersea Warfare Center (NUWC), Newport, RI. He is the author of numerous conference and journal papers in the field of signal processing and classification and has taught part-time as an adjunct professor of electrical engineering at the University of Connecticut, Storrs.

Dr. Baggenstoss received the 2002 URI Excellence Award in Science and Technology.