Improved Optimization of Time-Frequency-Based Signal Classifiers

Manuel Davy, Christian Doncarli, and G. Faye Boudreaux-Bartels

***Abstract*—Time-frequency representations (TFRs) are efficient tools for nonstationary signal classification. However, the choice of the TFR and of the distance measure employed is critical when no prior information other than a learning set of limited size is available. In this letter, we propose to jointly optimize the TFR and distance measure by minimizing the (estimated) probability of classification error. The resulting optimized classification method is applied to multicomponent chirp signals and real speech records (speaker recognition). Extensive simulations show the substantial improvement of classification performance obtained with our optimization method.**

***Index Terms*—Classification, contrast criterion, distance measure, kernel optimization, time-frequency representations.**

I. Introduction

THE classification of signals is a common problem in signal processing. Important applications include speaker recognition, diagnosis of mechanical systems, and medical diagnosis. In some cases, a statistical model (such as Gaussian distribution) is known and an optimal classification procedure can be implemented [1], [2]. Often, however, no statistical model is available. Here, application of the optimal (Bayesian) classifier would require estimation of the relevant probability density functions, which in turn requires availability of a large learning set of signal realizations. In practice, the learning set is often of small size and, hence, suboptimal procedures must be used. It is then necessary to find a model-free representation space in which the differences between the classes are emphasized and the similarities are deemphasized.

Such a *discriminant* representation space depends on the signals to be classified. A variety of approaches have been previously proposed such as a wavelet packet-based algorithm [3] (selection of the best basis) and methods based on time-frequency representations (TFRs) [4]–[7] (selection of the best TFR within Cohen's class [8]). However, these techniques still suffer from certain restrictions concerning the optimization employed. In this letter, we propose a TFR-based approach that uses an improved optimization procedure. The improvement is due to the fact that 1) the distance measure is optimized (in addition to the TFR) and 2) the optimality criterion used is closely related to the probability of classification error.

Manuscript received July 28, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. G. B. Giannakis.

M. Davy and C. Doncarli are with the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN), Nantes Cedex 03, France (e-mail: Manuel.Davy@irccyn.ec-nantes.fr).

G. F. Boudreaux-Bartels is with the Department of Electrical and Computer Engineering, University of Rhode Island, Kingston, RI 02881 USA.

Publisher Item Identifier S 1070-9908(01)00286-3.

TFR-based classification methods are interesting for several reasons.

- For Gaussian signals, the optimal classifier admits an equivalent formulation using TFRs [9]–[11].
- TFRs are discriminant representations if the time-frequency (TF) structures of signals from different classes are different, a situation that is often encountered in practical applications.
- Adaptation of the TFR to a given learning set allows us to improve classification performance.

In this letter, therefore, we will consider the following situation.

- No statistical model is assumed. In particular, the signals can be non-Gaussian.
- Signals from different classes feature different TF structures, i.e., TFRs are discriminant representations.
- The only explicit prior information available consists of a learning set of labeled samples (supervised classification).

With these assumptions, finding the optimal TF classifier (optimal in the sense of minimal probability of classification error) is intractable. Therefore, we will use a TF classifier with a given, plausible structure and optimize its key components: TFR and distance measure. As a major difference from previous work, our optimization minimizes the (estimated) probability of classification error.

This letter is organized as follows. In Section II, the TF classification method used is explained. Sections III and IV discuss the novel optimization method, with Section III developing the optimality criterion and Section IV explaining the optimization procedure. Finally, simulation results are presented in Section V.

II. TF Classification Method

In this section, we discuss the structure of the TF classification method used. Specifically, we consider the decision rule employed and the two key components to be optimized: TFR and distance measure. In what follows, let $x(t)$ denote a signal to be classified and ω_i, $i = 1, \cdots, N$ the different classes. We assume availability of a learning set $\mathcal{X}$ of labeled signals, i.e., $\mathcal{X}$ is separated into N sets $\mathcal{X}_i = \{x_1^i, \cdots, x_{n_i}^i\}$, $i = 1, \cdots, N$, representing the respective classes ω_i. Each learning set consists of n_i signals $x_k^i(t)$, $k = 1, \cdots, n_i$.

A. Decision Rule

The proposed TF classifier uses the following decision rule: x is assigned to $\omega_{\hat{i}}$ with

$$\hat{i} = \underset{i=1,\cdots,N}{\arg\min}\ d\left(C_x^\phi, \overline{C_i^\phi}\right). \tag{1}$$

Here, $d(\cdot, \cdot)$ is a dissimilarity or distance measure (see Section II-B), $\mathcal{C}_x^\phi(t, f)$ is a real-valued TFR from Cohen's class [8], [12] (see Section II-C), and $\overline{\mathcal{C}_i^\phi}(t, f)$ is a representative TFR characterizing the class ω_i. Here, we define $\overline{\mathcal{C}_i^\phi}(t, f)$ to be the average of the TFRs of the learning signals $x_k^i \in \mathcal{X}_i$

$$\overline{\mathcal{C}_i^\phi}(t, f) \triangleq \frac{1}{n_i} \sum_{k=1}^{n_i} \mathcal{C}_{x_k^i}^\phi(t, f). \tag{2}$$

Several TF classifiers previously proposed use the general decision rule (1) [4]–[6] or a similar rule [9]. Furthermore, with specific choices of d, $\mathcal{C}_x^\phi$, and $\overline{\mathcal{C}_i^\phi}$, this decision rule is optimal for Gaussian signals in the low energy coherence case [9] (see Section II-B).

Given the decision rule defined by (1) and (2), we propose to jointly optimize the dissimilarity/distance measure d and the TFR $\mathcal{C}^\phi$. The specific types of d and $\mathcal{C}^\phi$ that will be used for this optimization are explained in what follows.

B. Dissimilarity/Distance Measures

The optimization of d consists in optimally choosing d from among a finite set of specific dissimilarity/distance measures, some of which are reviewed in this section.

- The Euclidean distance (used, e.g., in [13]) is a special case (with $q = 2$) of the family of L_q distances defined by

$$d_{L_q}\left(\mathcal{C}_x^\phi, \overline{\mathcal{C}_i^\phi}\right) \triangleq \left[\iint \left|\mathcal{C}_{x_1}^\phi(t, f) - \overline{\mathcal{C}_i^\phi}(t, f)\right|^q dt\, df\right]^{1/q}. \tag{3}$$

- The correlation distance [4]

$$d_{cor1}\left(\mathcal{C}_x^\phi, \overline{\mathcal{C}_i^\phi}\right) \triangleq 1 - \frac{2\iint \mathcal{C}_x^\phi(t, f)\overline{\mathcal{C}_i^\phi}(t, f)\, dt\, df}{\|\mathcal{C}_x^\phi\|^2 + \left\|\overline{\mathcal{C}_i^\phi}\right\|^2} \tag{4}$$

  where, e.g., $\|\mathcal{C}_x^\phi\|^2 \triangleq \iint [\mathcal{C}_x^\phi(t, f)]^2\, dt\, df$, is a normalized version of the squared L_2 distance since $d_{cor1}(\mathcal{C}_x^\phi, \overline{\mathcal{C}_i^\phi}) = d_{L_2}^2(\mathcal{C}_x^\phi, \overline{\mathcal{C}_i^\phi})/(\|\mathcal{C}_x^\phi\|^2 + \|\overline{\mathcal{C}_i^\phi}\|^2)$. Alternatively, a nonnormalized negative TF correlation can be used

$$d_{cor2}\left(\mathcal{C}_x^\phi, \overline{\mathcal{C}_i^\phi}\right) \triangleq -\iint \mathcal{C}_x^\phi(t, f)\overline{\mathcal{C}_i^\phi}(t, f)\, dt\, df. \tag{5}$$

  The classifier using d_{cor2} is optimal for Gaussian signals and low SNR (low energy coherence case) if $\mathcal{C}_x^\phi$ is chosen as the Wigner–Ville (WV) distribution of x, and $\overline{\mathcal{C}_i^\phi}$ is chosen as the WV spectrum [12] of the signals from class ω_i [9].
- A broad family of dissimilarity measures is given by the *f-divergences*

$$d_{f-\text{div}}\left(\mathcal{C}_x^\phi, \overline{\mathcal{C}_i^\phi}\right) \triangleq g\left[\iint f\left(\frac{N\mathcal{C}_x^\phi(t, f)}{N\overline{\mathcal{C}_i^\phi}(t, f)}\right) N\overline{\mathcal{C}_i^\phi}(t, f)\, dt\, df\right] \tag{6}$$

  where f and g are functions satisfying certain properties [14] and $N\mathcal{C}_x^\phi$, and $N\overline{\mathcal{C}_i^\phi}$ are positive TFRs whose integral equals one. Here, we use

$$N\mathcal{C}_x^\phi(t, f) \triangleq \frac{|\mathcal{C}_x^\phi(t, f)|}{\iint |\mathcal{C}_x^\phi(s, \nu)|\, ds\, d\nu}$$

$$N\overline{\mathcal{C}_i^\phi}(t, f) \triangleq \frac{\left|\overline{\mathcal{C}_i^\phi}(t, f)\right|}{\iint \left|\overline{\mathcal{C}_i^\phi}(s, \nu)\right|\, ds\, d\nu}.$$

  Examples of $d_{f-\text{div}}$ are the Kolmogorov distance

$$d_{\text{Kol}}\left(\mathcal{C}_{x_1}^\phi, \overline{\mathcal{C}_i^\phi}\right) \triangleq \iint \left|N\mathcal{C}_{x_1}^\phi(t, f) - N\overline{\mathcal{C}_i^\phi}(t, f)\right| dt\, df \tag{7}$$

  and the Bhattacharyya distance [14].

$$d_{\text{Bha}}\left(\mathcal{C}_{x_1}^\phi, \overline{\mathcal{C}_i^\phi}\right) \triangleq -\log \iint \sqrt{N\mathcal{C}_{x_1}^\phi(t, f) N\overline{\mathcal{C}_i^\phi}(t, f)}\, dt\, df. \tag{8}$$

- Finally, the L_q distances can be applied to normalized TFRs:

$$d_{NL_q}\left(\mathcal{C}_x^\phi, \overline{\mathcal{C}_i^\phi}\right) \triangleq \left[\iint \left|N\mathcal{C}_{x_1}^\phi(t, f) - N\overline{\mathcal{C}_i^\phi}(t, f)\right|^q dt\, df\right]^{1/q}. \tag{9}$$

C. TFRs

The optimization of $\mathcal{C}^\phi$ uses a TFR of Cohen's class, of which certain parameters are optimized. Any specific TFR from Cohen's class is uniquely characterized by its ambiguity domain kernel $\phi(\xi, \tau)$ [8], [12]. Here we consider three different kernel types (hence, TFR types) that are parameterized by a small number of parameters, thus facilitating TFR optimization (A small number of parameters is furthermore important, since otherwise, overlearning may occur [2]). All three kernel types can be written as

$$\phi(\xi, \tau) = e^{-\varphi(\xi, \tau)}$$

with the function $\varphi(\xi, \tau)$ given as follows.

- For the radially Gaussian kernel (RGK) [15]

$$\varphi(\xi, \tau) = \frac{\rho^2}{2\sigma(\theta)^2}$$

  where $\rho = \sqrt{\xi^2 + \tau^2}$ and $\theta = \tan^{-1}(\xi/\tau)$ are the polar coordinates corresponding to (τ, ξ). In order for $\mathcal{C}^\phi$ to be realvalued, the "contour function" $\sigma(\theta)$ has to be π-periodic. Thus, we propose using the truncated Fourier series

$$\sigma(\theta) = a_0 + \sum_{p=1}^{p_{\max}} [a_p \cos(2p\theta) + b_p \sin(2p\theta)]$$

  in which a_0 is chosen such that $\sigma(\theta) > \tilde{a}_0 > 0$ (otherwise, $\sigma(\theta)^2$ would present singularities for the values of θ where the sign of $\sigma(\theta)$ changes). The kernel parameters are then $\tilde{a}_0$ as well as a_p and b_p with $p = 1, \cdots, p_{\max}$. In contrast to [15], no volume constraint is used here.

- For the generalized marginals Choi–Williams kernel GMCWK) [16],

$$\varphi(\xi, \tau) = \frac{1}{\omega} \prod_{k=1}^{B} (\xi \cos \theta_k + \tau \sin \theta_k)^2$$

where

B	number of branches;
θ_k with $k = 1, \cdots, B$	branch angles;
ω	branch widths.

- For the multiform tiltable exponential kernal (MTEK) [17]

$$\varphi(\xi, \tau) = \pi \left\{ \left(\left[\frac{\tau}{\tau_0}\right]^2 \left[\frac{\xi}{\xi_0}\right]^{2\alpha} + \left[\frac{\tau}{\tau_0}\right]^{2\alpha} \left[\frac{\xi}{\xi_0}\right]^2 + 2rA \right)^2 \right\}^{\lambda}$$

where A stands for $([\tau\xi/\tau_0\xi_0]^{\beta})^{\gamma}$. The kernel parameters are α, β, γ, τ_0, ξ_0, r, λ.

We note that all of the three kernels described above include the WV distribution as a special case.

III. Optimization

Similar to the optimal Bayes classifier [2], our optimality criterion is the minimal (estimated) probability of classification error $\widehat{P_e}(d, \phi)$ (our notation emphasizes that $\widehat{P_e}$ depends on d and ϕ).

We shall first develop an explicit expression of $P_e(d, \phi)$. We suppose that the class index i and the signal $x \in \omega_i$ are random. The result of classification, $\hat{i}$ in (1), is then random as well. The classes ω_i are assumed equiprobable, i.e., $P[i] = 1/N$, $i = 1, \cdots, N$. The error probability can then be written as

$$\begin{aligned} P_e(d, \phi) &= P[\hat{i} \neq i] \\ &= \sum_{i=1}^{N} P[\hat{i} \neq i | i] P[i] \\ &= \frac{1}{N} \sum_{i=1}^{N} P[\hat{i} \neq i | i]. \end{aligned} \tag{10}$$

The error event $\hat{i} \neq i$ is the union of the events $\hat{i} = j$ with $j \neq i$ (more precisely, $j \in \{1, \cdots, N\} \setminus \{i\}$). These events being disjoint

$$P[\hat{i} \neq i | i] = \sum_{\substack{j=1 \\ j \neq i}}^{N} P[\hat{i} = j | i].$$

Inserting in (10) yields

$$P_e(d, \phi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} P[\hat{i} = j | i]. \tag{11}$$

For given i and j, let d_{ij} denote the random variable $d(\mathcal{C}_x^{\phi}, \overline{\mathcal{C}}_j^{\phi})$ with random $x \in \omega_i$. Furthermore, let

$$e_{ij} = d_{ij} - d_{ii} = d\left(\mathcal{C}_x^{\phi}, \overline{\mathcal{C}}_j^{\phi}\right) - d\left(\mathcal{C}_x^{\phi}, \overline{\mathcal{C}}_i^{\phi}\right). \qquad j \neq i$$

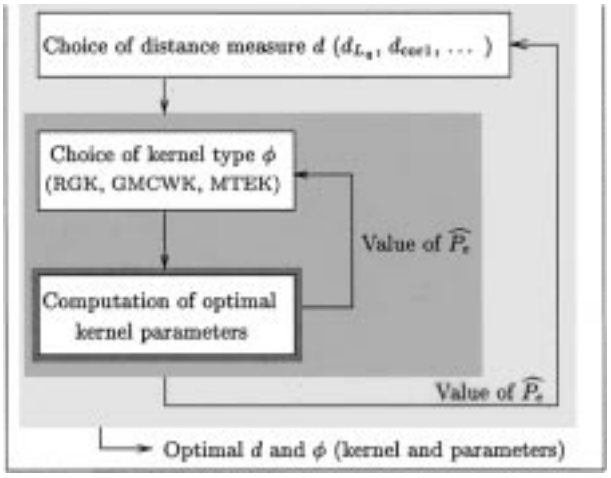

Fig. 1. Optimization procedure. The different dissimilarity/distance measures d and TFR kernel types ϕ are evaluated one by one. For each choice of d and ϕ, the kernel parameters are optimized. The optimal d and ϕ (type and parameters) are those for which $\widehat{P_e}(d, \phi)$ is minimum.

Then $P[\hat{i} = j | i] = P[d_{ij} < d_{ii}] = P[e_{ij} < 0]$ and (11) becomes

$$P_e(d, \phi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} P[e_{ij} < 0]. \tag{12}$$

In a discrete-time/discrete-frequency implementation where the TF integrals in (3)–(9) are replaced by double summations, the distance random variables d_{ij} are the sum of many random components. Thus, it can be conjectured (using a central limit theorem argument) that the distribution of $e_{ij} = d_{ij} - d_{ii}$ can be well approximated by a Gaussian distribution. In Section V-A, this conjecture will be confirmed by experimental observations. Therefore, e_{ij} will hereafter be modeled as Gaussian. It follows that

$$\begin{aligned} P[e_{ij} < 0] &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\,\sigma_{ij}} \exp\left(-\frac{1}{2} \left[\frac{u - m_{ij}}{\sigma_{ij}} \right]^2 \right) du \\ &= Q\left(\frac{m_{ij}}{\sigma_{ij}} \right). \end{aligned} \tag{13}$$

Here, $m_{ij} = E[e_{ij}]$, $\sigma_{ij}^2 = \mathrm{var}[e_{ij}]$, and the Q-function is $Q(c) = \int_c^{+\infty} (1/\sqrt{2\pi}) \exp(-(u^2/2))\, du$.

Based on expression (13), $P_e(d, \phi)$ can be estimated from the learning set. We substitute for m_{ij} and σ_{ij}^2 the following estimates derived from the learning signals x_k^i and x_k^j:

$$\hat{m}_{ij} = \frac{1}{n_i} \sum_{k=1}^{n_i} \left[d\left(\mathcal{C}_{x_k^i}^{\phi}, \overline{\mathcal{C}}_i^{\phi}\right) - d\left(\mathcal{C}_{x_k^i}^{\phi}, \overline{\mathcal{C}}_j^{\phi}\right) \right] \tag{14}$$

$$\hat{\sigma}_{ij}^2 = \frac{1}{n_i} \sum_{k=1}^{n_i} \left[d\left(\mathcal{C}_{x_k^i}^{\phi}, \overline{\mathcal{C}}_i^{\phi}\right) - d\left(\mathcal{C}_{x_k^i}^{\phi}, \overline{\mathcal{C}}_j^{\phi}\right) - \hat{m}_{ij} \right]^2. \tag{15}$$

Hence, the estimated error probability is

$$\widehat{P_e}(d, \phi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} Q\left(\frac{\hat{m}_{ij}}{\hat{\sigma}_{ij}} \right). \tag{16}$$

As explained in next section, $\widehat{P_e}$ is minimized with respect to the dissimilarity/distance measure d, the TFR type (kernel type ϕ), and the parameters contained in ϕ.

IV. Optimization Procedure

As was explained in Section II, only a relatively small number of different dissimilarity/distance measures d and TFR types (kernel types ϕ) are considered. Therefore, the minimization of $\widehat{P_e}(d, \phi)$ in (16) is done by evaluating the different types of d and ϕ one by one while optimizing the parameters of ϕ for each choice of d and ϕ. The overall optimization procedure is thus hierarchical as depicted in Fig. 1. The different steps are as follows.

1) Choose a distance measure d.
2) Choose a kernel type ϕ.
3) Determine the kernel parameters (parameters of ϕ) such that $\widehat{P_e}(d, \phi)$ is minimal for the given d and ϕ.

This procedure is iterated for different choices of d and ϕ. The final result is given by the d and ϕ (type and parameters) for which $\widehat{P_e}(d, \phi)$ is minimal.

The optimization of the kernel parameters (for given dissimilarity/distance measure d and TFR kernel type ϕ) is performed using a suitable numerical algorithm that does not require computation of gradients. We used the Nelder and Mead direct search algorithm [18] (e.g., MATLAB function `fmins`). This algorithm requires computation of $\widehat{P_e}$ for different kernel parameter values. This is done as follows.

1) Compute the TFRs $\mathcal{C}^{\phi}_{x^i_k}$ of $x^i_k \in \mathcal{X}_i$ for $k = 1, \cdots, n_i$ and $i = 1, \cdots, N$.
2) Compute the representative TFRs $\overline{\mathcal{C}}^{\phi}_i$ for $i = 1, \cdots, N$ using (2).
3) Compute the distances $d(\mathcal{C}^{\phi}_{x^i_k}, \overline{\mathcal{C}}^{\phi}_j)$ for $k = 1, \cdots, n_i$, $i = 1, \cdots, N$ and $j = 1, \cdots, N$.
4) For $i = 1, \cdots, N$ and $j = 1, \cdots, N$, compute $\hat{m}_{ij}$ and $\hat{\sigma}_{ij}$ using (14) and (15).
5) Compute $\widehat{P_e}(d, \phi)$ using (16).

V. Results

In this section, we present the results of the proposed TF classifier obtained for two classification problems.

A. Synthetic Multicomponents Signals

We consider synthetic multicomponent chirp signals defined as

$$\omega_1\colon x(t) = \sin 2\pi(f_0 t+\psi_0)+\sin 2\pi(s_1^1 t^2+f_1 t+\psi_1)+\epsilon(t)$$
$$\omega_2\colon x(t) = \sin 2\pi(f_0 t+\psi_0)+\sin 2\pi(s_1^2 t^2+f_1 t+\psi_1)+\epsilon(t)$$

where

ψ_0 and ψ_1	uncorrelated random initial phases (uniformly distributed in [0; 1]);
$f_0 = 0.25$ and $f_1 = 0.4$	fixed normalized frequencies;
ϵ	Gaussian white noise of variance $\sigma_\epsilon^2 = 2$.

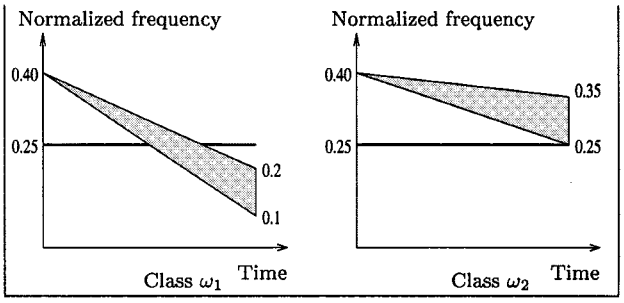

Fig. 2. Idealized TFRs of the signals from ω_1 and ω_2. Each class consists of a tone at normalized frequency 0.25, and a linear chirp signal starting at normalized frequency 0.40 and ending at a random frequency uniformly distributed between 0.1 and 0.2 for class 1 and between 0.25 and 0.35 for class 2.

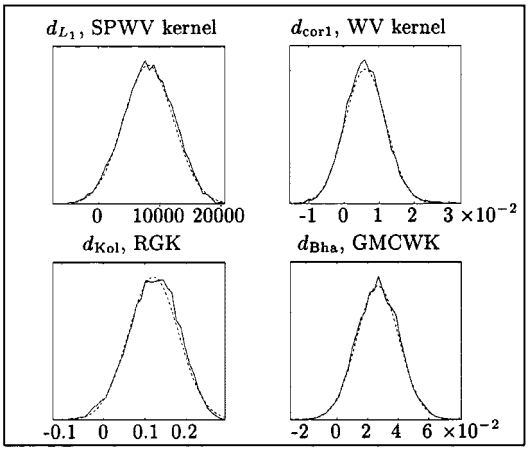

Fig. 3. Histograms of the variable e_{12}. The histograms of e_{12} are estimated using the learning set ($n_1 = n_2 = 50$ signals) for 10 000 signal realizations. Four dissimilarity/distance measures and kernels are tested. Solid line: estimated histogram of e_{12}. Dashed line: Gaussian distribution of same mean and variance as the histogram represented by solid line.

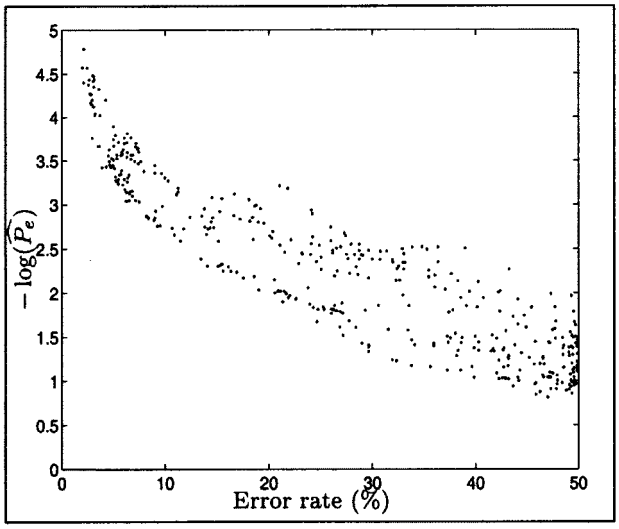

Fig. 4. Estimated probability of classification error $\widehat{P_e}(d, \phi)$ (estimated from the learning set with $n_1 = n_2 = 50$) versus estimated classification error rate (estimated by classifying 2×1000 signals). The Kolmogorov distance is chosen. Each plotted dot corresponds to one of the 500 tested TFR kernels.

The chirp slopes s_1^1 and s_1^2 are random. The distribution of s_1^1 is different from the distribution of s_1^2 (s_1^1 and s_1^2 are the discriminant features). The final chirp frequency is then uniformly

TABLE I
RESULTS OF THE OPTIMIZATION PROCEDURE APPLIED TO THE SYNTHETIC MULTICOMPONENT SIGNALS. FOR EACH TESTED DISSIMILARITY/DISTANCE MEASURE d AND TFR KERNEL TYPE, THE KERNEL PARAMETERS ARE OPTIMIZED USING THE LEARNING SET ($n_1 = n_2 = 50$). THE ERROR RATE (IN %) AND THE VALUE OF $\widehat{L_e} = -\log[\widehat{P_e}(d, \phi)]$ ARE ESTIMATED USING $2 \times 10\,000$ TEST SIGNALS

Distance Measure	MTEK		GMCWK			RGK		
	β=1, γ=1 4 param.	β=2, γ=0.5 4 param.	$B=1$ 2 param.	$B=2$ 3 param.	$B=3$ 4 param.	$p_{\max}=1$ 3 param.	$p_{\max}=2$ 5 param.	$p_{\max}=3$ 7 param.
Correlation d_{Cor1}	9.72% $\widehat{L_e}$ =3.75	9.38% $\widehat{L_e}$ =3.77	2.32% $\widehat{L_e}$ =4.08	7.82% $\widehat{L_e}$ =3.81	8.80% $\widehat{L_e}$ =3.70	16.69% $\widehat{L_e}$ =3.40	16.98% $\widehat{L_e}$ =3.39	16.48% $\widehat{L_e}$ =3.40
Quadratic $d^2_{L_2}$	3.74% $\widehat{L_e}$ =3.77	8.38% $\widehat{L_e}$ =3.57	2.09% $\widehat{L_e}$ =3.77	3.47% $\widehat{L_e}$ =3.73	2.81% $\widehat{L_e}$ =3.71	12.54% $\widehat{L_e}$ =3.34	5.84% $\widehat{L_e}$ =3.69	13.22% $\widehat{L_e}$ =3.32
L_1 d_{L_1}	3.79% $\widehat{L_e}$ =4.15	2.88% $\widehat{L_e}$ =4.26	2.70% $\widehat{L_e}$ =3.93	3.26% $\widehat{L_e}$ =3.94	2.51% $\widehat{L_e}$ =4.15	3.56% $\widehat{L_e}$ =4.14	2.62% $\widehat{L_e}$ =4.29	2.88% $\widehat{L_e}$ =4.29
Kolmogorov d_{Kol}	3.62% $\widehat{L_e}$ =3.61	3.92% $\widehat{L_e}$ =3.66	2.79% $\widehat{L_e}$ =4.03	2.88% $\widehat{L_e}$ =4.01	1.73% $\widehat{L_e}$ =4.43	2.01% $\widehat{L_e}$ =4.29	1.63% $\widehat{L_e}$ =4.48	1.84% $\widehat{L_e}$ =4.40
Bhattacharyya d_{NBha}	3.90% $\widehat{L_e}$ =3.48	4.41% $\widehat{L_e}$ =3.52	1.86% $\widehat{L_e}$ =4.15	3.02% $\widehat{L_e}$ =3.79	2.52% $\widehat{L_e}$ =3.95	1.67% $\widehat{L_e}$ =4.30	1.56% $\widehat{L_e}$ =4.37	1.58% $\widehat{L_e}$ =44.30
NL_2 d_{NL_2}	15.51% $\widehat{L_e}$ =2.80	3.04% $\widehat{L_e}$ =3.81	2.11% $\widehat{L_e}$ =4.21	1.87% $\widehat{L_e}$ =4.36	1.99% $\widehat{L_e}$ =4.16	2.43% $\widehat{L_e}$ =4.07	1.45% $\widehat{L_e}$ =4.31	1.55 % $\widehat{L_e}$ =4.47

distributed in [0.10; 0.20] for $x \in \omega_1$ and in [0.25; 0.35] for $x \in \omega_2$. Fig. 2 displays idealized TFRs of these signals.

Using these signals, we have tested the Gaussian approximation for the statistical distribution of e_{ij} (see Section III). Fig. 3 displays the estimated distribution of e_{12} for several choices of dissimilarity/distance measure and TFR kernel. The Gaussian approximation is seen to be very precise.

Fig. 4 shows the relationship between the estimated probability of classification error $\widehat{P_e}(d, \phi)$ (estimated from the learning set for a given choice of dissimilarity/distance measure d and kernel ϕ), and the error rate obtained by applying the corresponding classifier (estimated using a test set). The distance measure d_{Kol} is selected. The TFR kernel type is either RGK or GMCWK, with randomly selected parameters. For each kernel ϕ (type and parameters), $\widehat{P_e}(d_{\mathrm{Kol}}, \phi)$ is estimated using the learning set (composed of $n_1 = n_2 = 50$ signals), and the classification error rate is estimated by applying the classifier defined by (d_{Kol}, ϕ) to 2×1000 other signals. In Fig. 4, the results corresponding to 500 kernels (type and parameters) are plotted. It is seen that $\widehat{P_e}(d, \phi)$ is strongly related to the actual classifier error rate.

Finally, we applied the optimization procedure to a set of eight kernel types and six dissimilarity/distance measures, which results in 48 kernel parameter optimizations. The results of each kernel parameter optimization are displayed in Table I. The best overall solution uses the NL_2 distance and the RGK with $p_{\max} = 2$.

We have compared the proposed classification method with previous TF methods. The decision rule is as displayed Section II-A. Three different approaches can be distinguished.

- In [4] and [5], the classifier relies on TF correlations, the kernel parameters being optimized using a contrast criterion. The kernel types are smoothed pseudo-WV (SPWV) and spectrogram, respectively.
- In [19], TF distance measures are implemented, the TFR being the Wigner–Ville distribution.
- In [7], the kernel is computed directly in the ambiguity plane by selecting discriminant (τ, ξ) locations. The decision rule involves the Mahalanobis distance d_{Mah} [2].

TABLE II
CLASSIFICATION RESULTS OBTAINED WITH DIFFERENT TF METHODS (SYNTHETIC MULTICOMPONENT SIGNALS). THE LEARNING SET CONSISTS OF 2×50 SIGNALS, THE TEST SET CONSISTS OF $2 \times 10\,000$ OTHER SIGNALS

TFR	d	Error rate	Reference
WV	d_{cor2}	22.46 %	[9]
SPWV	d_{cor1}	3.87 %	[4]
WV	d_{Kol}	20.94 %	[19]
Ambig. plane	d_{Mah}	4.94 %	[7]
Jointly optimized		1.45 %	This letter

TABLE III
SPEAKER RECOGNITION STUDY FOR 3 SPEAKERS OF THE VOWEL "A." THE TFR KERNEL TYPE IS THE MODIFIED RGK

Distance measure	TFR Kernel	Error Rate
Correlation	Wigner-Ville	27.33%
Correlation	Optimized RGK	27.33%
Kolmogorov	Optimized RGK	4.00%
Bhattacharyya	Optimized RGK	3.33%
L_2	Optimized RGK	23.33%

Table II displays the results obtained with these methods. It is seen that the proposed method is the most efficient.

B. Real Speech Signals

The method has been applied to the following real signals: three different speakers pronounce the vowel "a." The aim is to recognize which speaker corresponds to a given utterance. An "a" utterance is numerically recorded, sampled at a rate of 2 kHz. The complete set is made of 65 utterances of one vowel by each of the three speakers, i.e. $3 \times 65 = 195$ signals. Each utterance is 512 points long. The training set features 15 signals in each class, randomly taken from the full set. The remaining signals are used as a test set.

For this example, we will use the φ-divergences, the L_2 distance applied to normalized TFRs, and the Jensen distance as discriminant functions. The kernel type here is the RGK, easy to optimize and versatile. Table III displays the error rates. The φ-divergences are the most efficient in this case. Using the correlation distance, the optimal RGK is identical to the WV kernel

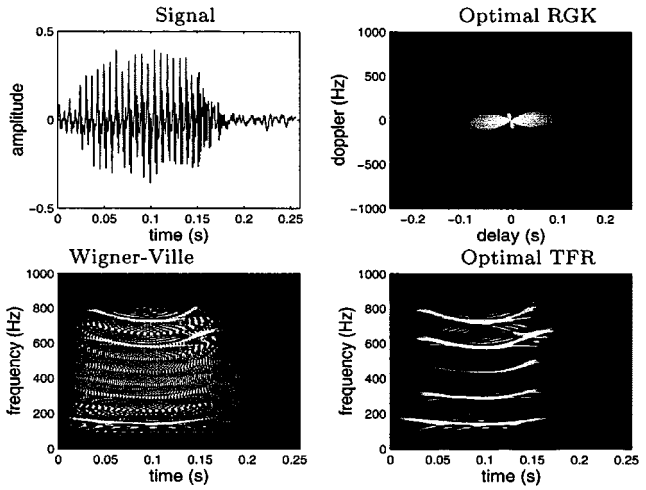

Fig. 5. Speaker recognition. First column: a signal in the first class and its WV representation. Second column: the optimal RGK and the optimal TFR of the signal.

(no volume constraint is used here), which explains the same error rate. Fig. 5 displays the optimal RGK and TFRs of a test signal.

VI. Conclusions

We presented a new TF classification method with a learning set as the only prior information. This method employs an improved optimization procedure, and a new optimality criterion is proposed. Similar to the Bayes classifier, the proposed optimization consists of minimizing the probability of classification error.

The accuracy of the new criterion was demonstrated for two examples. The TF dissimilarity/distance measure and kernel types employed are shown very general, and the method can be applied to a wide variety of classification problems.

Acknowledgment

The authors would like to thank P. Hlawatsch for helpful suggestions and comments on this letter.

References

[1] M. Davy, C. Doncarli, and J. Y. Tourneret, "Supervised classification using MCMC methods," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, Turkey, June 2000.

[2] R. O. Duda and P. E. Hart, *Pattern Classiftcation and Scene Analysis*. New York: Wiley-Interscience, 1973.

[3] N. Saito and R. Coifman, "Local discriminant bases," in *Wavelet Applications in Signal and Image Processing II*, A. F. Laine and M. A. Unser, Eds: Proc. SPIE, 1994, vol. 2303.

[4] C. Heitz, "Optimum time-frequency representations for the classification and detection of signals," *Appl. Signal Process.*, no. 3, pp. 124–143, 1995.

[5] C. Richard and R. Lengellé, "Data driven design and complexity control of time frequency detectors," *Signal Process.*, vol. 77, pp. 37–48, Jan. 1999.

[6] M. Davy and C. Doncarli, "Optimal kernels of time-frequency representations for signal classification," in *Int. Symp. Time-Frequency and Time Scale*, Pittsburgh, PA, 1998, pp. 581–584.

[7] B. Gillespie and L. Atlas, "Optimizing TF kernels for classification," *IEEE Trans. Signal Processing*, to be published.

[8] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Mag.*, pp. 21–67, Apr. 1992.

[9] P. Flandrin, "A time-frequency formulation of optimal detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1337–1384, Sept. 1988.

[10] A. M. Sayeed and D. L. Jones, "Optimal detection using bilinear time-frequency and time-scale representations," *IEEE Trans. Signal Processing*, vol. 43, pp. 2872–2883, Dec. 1995.

[11] G. Matz and F. Hlawatsch, "Time-frequency subspace detectors and application to knock detection," *Int. J. Electron. Commun.*, vol. 53, no. 6, pp. 379–385, 1999.

[12] P. Flandrin, *Time-Frequency/Time-Scale Analysis*. New York: Academic , 1999.

[13] L. Atlas, J. Droppo, and J. McLaughlin, "Optimizing time-frequency distributions for automatic classification," in *SPIE–Int. Soc. Optical Engineering*, 1997.

[14] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Process.*, vol. 18, pp. 349–369, Dec. 1989.

[15] R. G. Baraniuk and D. L. Jones, "A signal-dependant time-frequency representation: Optimal kernel design," *IEEE Trans. Signal Processing*, vol. 41, pp. 1589–1601, Apr. 1993.

[16] X.-G. Xia, Y. Owechko, B. H. Soffer, and R. M. Matic, "On generalized-marginals time-frequency distributions," *IEEE Trans. Signal Processing*, vol. 44, pp. 2882–2886, Nov. 1996.

[17] H. Costa and G. F. Boudreaux-Bartels, "Design of time-frequency representations using a multiform, tiltable exponential kernel," *IEEE Trans. Signal Processing*, vol. 43, pp. 2283–2301, Oct. 1995.

[18] J. M. Ortega and W. C. Rheinboldt, *Iterative Solutions of Nonlinear Equations in Several Variables*. New York: Academic, 1970.

[19] I. Vincent, C. Doncarli, and E. Le Carpentier, "Non-stationary signals classification using time-frequency distributions," in *Int. Symp. Time-Frequency and Time Scale*, Paris, France, 1994, pp. 233–236.