Optimized Time-Frequency Representations for the Classification of Signals

CHRISTOPH HEITZ

Centre for Data Analysis and Model Building,
University of Freiburg, Albertstr. 26-28, D-79104 Freiburg

Résumé
Les représentations temps-fréquences (TFR) sont souvent employées dans le domain du traitement du signal. Généralement il n'est pas facile de trouver la représentation qui est adaptée la mieux à la structure des signaux considerés et au problème de l'analyse. Pour le cas de classification avec deux classes de signaux données nous derivons des mésures d'optimalité pour TFRs. Ces mésures aussi bien que l'idée de classification ne posent pas aux *features* mais à la représentation temps-fréquence non-parametrique des signaux. De plus il n'est pas nécessaire de supposer un modèle des signaux. La représentation optimale est celle qui rend la mésure maximale. Pour le cas d'une connaissance imparfaite des classes de signal, òu on ne connait que quelque rèalisations de chaque classe, nous montrons comment la représentation optimale peux être estimée.

Abstract
Time-frequency representations (TFRs) are widely used in signal processing. However, it is always difficult to find the optimum TFR for a given signal class and analysis task. For the problem of classification with two given signal classes, we derive two different optimality measures for TFRs. The measures as well as the underlying classification scheme are not feature based but use distances between the (non-parametric) TFRs of the signals. Furthermore no model assumptions on the signals have to be made. Maximizing the appropriate measure leads to the optimum TFR. For the case of imperfect knowledge of the signal classes it is shown how the optimum TFR can be estimated with a finite sample of training data.

1 Introduction

Time-frequency representations (TFRs) are powerful tools for signal analysis and thus widely used in signal processing. It is well known, however, that there is no single TFR which is "the best" for all problems. It is still an unsolved problem how to determine the optimum TFR for a given signal class and analysis task.

We address the problem of classification of signals of a fixed length. For simplification only two classes are regarded. The question is: How can we find the TFR which is best fitted for the classification task? We restrict ourselves to TFRs of Cohen's class [1].

The usual way is to define from an *a priori* knowledge of the signals structure one or more features which are supposed to give a good representation of the interesting qualities of the signals. Then a TFR is searched which allows a good (reliable, robust, exact, ...) estimation of these features. Afterwards a classification is performed in the low-dimensional feature space.

However, this method is restricted on cases where the structure of the signal classes is known and, furthermore, a knowledge about relevant features is available in advance. Frequently, these conditions are not fulfilled. Especially when analyzing complex signals, often it is far from obvious which property is the relevant one. Furthermore, the structure of the signals is often unknown.

Here we propose a new method for dealing with the classification problem: Quite intuitively, a TFR is called "good" if all signals of one class are similar to each other, but dissimilar to the members of the other class. Once such a "good" representation is found, an unknown signal can be assigned to one of the classes by just comparing the degree of similarity to the different classes and choosing the class that shows the largest similarity to the signal.

2 Optimizing the TFR

In the following this idea is formulated in a mathematical framework: Each signal $f(t)$ is represented in the time-frequency domain by its normalized TFR $C \equiv \frac{C_f}{\|C_f\|}$, where we only allow TFRs out of Cohen's class which are determined by a two-dimenisional complex kernel function $\Phi(\xi, \tau)$. The TFR of an arbitrary function $f(t) \in L_2(I\!R)$ is an element of the function

space $L_2(\mathbb{R} \times \mathbb{R})$, where a norm is defined by

$$\|C_f\|^2 \equiv \int |C_f(t,\omega;\Phi)|^2 \, dt \, d\omega$$

Hence, a representation C of a signal f lies on the unit sphere S of $L_2(\mathbb{R} \times \mathbb{R})$.

The distance between two signals is defined via the inner product of their representations C_i:

$$D_\Phi(f_1, f_2) = 1 - < C_1, C_2 > .$$

and hence depends on the used TFR.

A signal class is described by a stochastic variable $\mathbf{C}$, i.e. a probability distribution P on the unit sphere S. The simplest case is $P = \text{const}$ in a region of S, and $P = 0$ else. However, in general P will have a more complicated form.

Let two classes be described by the two stochastic variables $\mathbf{C}_1$ and $\mathbf{C}_2$. For localized classes it is natural to take $E(\mathbf{C}_i)$ as the representative of class i:

$$E(\mathbf{C}_i) = \int_S C \, P_i(dC)$$

The mean distance of the funtions of class 2 to the representative $E(\mathbf{C}_1)$ of class 1 is

$$D_\Phi^{12} = \int_S D_\Phi(E(\mathbf{C}_1), C') \, P_2(dC')$$

It can be shown easily that $D_\Phi^{12} = D_\Phi^{21}$.

Analogously the mean inner distance of class i is

$$D_\Phi^{ii} = \int_S D_\Phi(E(\mathbf{C}_i), C') \, P_i(dC')$$

A simple measure for the ability of a given TFR (i.e. a given kernel function Φ) to yield a well discriminating representation of the signals is given by

$$m_\Phi = \frac{D_\Phi^{12}}{D_\Phi^{11} + D_\Phi^{22}}$$

Thus, m_Φ measures the mean distance between the classes compared with the mean inner class distances. A large m_Φ means a good separation of the classes. Note that m_Φ is similar to Fisher's discrimination criterion in discriminance analysis [2].

Clearly, it depends on the chosen kernel how much the classes are separated in the time-frequency function space. The optimum TFR is now given by the kernel Φ that maximizes m_Φ.

A signal f is assigned to a class by calculating its representation C_f and searching the smallest distance $D_\Phi(C_f, E(\mathbf{C}_i))$.

The above measure m_Φ is only appropriate for the case of two well localized classes. Often the problem is not to discriminate such two classes but to decide whether a signal belongs to a class (supposed to be localized) or not. However, this case can also be viewed as a classification task, class 2 consisting of all the signals which do not belong to class 1. Here the classification scheme is slightly different: For a function f to be classified, its TFR C is calculated. f is assumed to be in class 1 if $D(C, E(\mathbf{C}_1))$ is smaller than some threshold γ.

The problem now is that class 2 (all the signals which do not belong to class 1) tends to be very large and no localization can be expected. In [3] a measure similar to m_Φ is introduced for dealing with this problem.

Here we propose a different approach for this kind of problem. Recall that our essential objective is to get an optimum classification rate. The measure m_Φ is, for well concentrated classes, closely related to the classification rate. In general, however, it is better to use the classification rate itself as optimality criterion. This can be done formally by replacing the distances D_Φ by binary distances $\hat{D}_{\gamma,\Phi}$ which are, for given $\gamma > 0$, defined as

$$\hat{D}_{\gamma,\Phi} = \begin{cases} 1, & D_\Phi \geq \gamma \\ 0, & D_\Phi < \gamma \end{cases}$$

Thus,

$$c_i(\gamma, \Phi) = \int_S \hat{D}_{\gamma,\Phi}(C, E(\mathbf{C}_1)) P_i(dC)$$

is just the fraction of the functions of class i whose distance to $E(\mathbf{C}_1)$ is larger than γ.

Given a threshold γ and the just proposed classification scheme, $100\,(1 - c_1)\%$ of the functions of class 1 and $100\,c_2\%$ of the functions of class 2 are classsified correctly. If we assume equal *a priori* probabilities for the two classes, the overall classification rate is $0.5\,(1 - c_1 + c_2)$, which is still depending on the threshold γ.

For a fixed kernel the optimum classification rate is given by maximizing about all possible γ:

$$\tilde{m}_\Phi = \max_\gamma (1 - c_1(\gamma, \Phi) + c_2(\gamma, \Phi))$$

Again this depends explicitely on the kernel Φ. The optimum kernel is chosen by maximizing $\tilde{m}_\phi$.

For both of the presented optimization algorithms the optimum kernel can be derived analytically (at least in principle), if the probability distribution of the two classes under consideration is known. If, on the other hand, each class is given by a sample of realizations, Φ_{opt} must be estimated from this sample. This problem is dealt with in the next section.

3 Estimation of the optimum kernel with a finite sample of realizations

Instead of the true probability distribution for the two classes we often have only a random sample of functions for which the class is known. From this information the optimum kernel has to be estimated. This is

done by estimating the measure m_Φ or $\tilde{m}_\Phi$ and maximizing this estimate with respect to Φ.

In order to get these estimates, the quantities D_Φ^{ij} and $c_i(\gamma, \Phi)$ are replaced by estimators.

With a sample of N_1 signals of class 1 and N_2 signals of class 2, an estimator of D_Φ^{12} is given by

$$\hat{D}_\Phi^{12} = 1 - \frac{1}{N_2} \sum_j < \hat{C}_1, C_2^j >$$

with $\hat{C}_1$ is the natural estimator of $E(\mathbf{C}_1)$: $\hat{C}_1 = 1 - \frac{1}{N_1} \sum_k C_1^k$ Analogously an estimator for D_Φ^{ii} can be constructed.

For the quantities $c_i(\gamma, \Phi)$ we replace again $E(\mathbf{C}_1)$ by $\hat{C}_1$, the $c_i(\gamma, \Phi)$ are estimated by calculating the fraction of the sample with $D(C, \hat{C}_1) > \gamma$. Maximizing over all possible γ yields an estimate for $\tilde{m}_\Phi$.

When maximizing m_Φ or $\tilde{m}_\Phi$ with respect to the kernel Φ, a general problem arises: the kernel $\Phi(\xi, \tau)$ is a complex function with infinitely many degrees of freedom. Thus it cannot be determined by a finite data sample [2]. In order to reduce the degrees of freedom of the kernel, it must be parametrized by few parameters, allowing a reliable estimation of the optimum parameter values.

We chose the following parametrization by the two parameters ξ_0 and τ_0:

$$\Phi(\xi, \tau) = e^{-\left(\frac{\xi}{\xi_0}\right)^2} e^{-\left(\frac{\tau}{\tau_0}\right)^2}, \qquad \xi_0, \tau_0 \geq 0$$

which corresponds still to a large class of TFRs (known as Smoothed Pseudo-Wigner-Ville distributions [1], including all spectrograms with Gaussian windows and the Wigner distribution ($\xi_0 = \tau_0 = \infty$).

4 Application on simulated data

In this section we show an application of the first kernel optimization scheme on simulated data.

We regard two nearly non-overlapping wave packets. The two classes are given by

$$\begin{aligned} f(t) &= h(t)\sin(\omega t + \varphi_1) + \\ &\qquad 2h(t - \Delta t)\sin(\omega t + \varphi_2) + a\epsilon(t) \\ g(t) &= 2h(t)\sin(\omega t + \varphi_3) + \\ &\qquad h(t - \Delta t)\sin(\omega t + \varphi_4) + a\epsilon(t) \end{aligned}$$

with $\Delta t = 32$, $h(t)$ a hanning window envelope function with width 30. The phases φ_i are random numbers and uniformely distributed in $[0, 2\pi]$, the frequency $\omega/2\pi$ being set to 0.15. $\epsilon(t)$ is white noise, $\epsilon(t) \sim N(0, 1)$, whose amplitude is controlled by the parameter a. In figure 1 (a) and (b) two realizations of each signal class with $a = 0$ are shown.

We simulated signals of length $L = 256$ with 100 realizations for each class which served as learning set. The (estimated) discrimination measure m_Φ was maximized by means of a standard nonlinear maximization routine, thus leading to the optimum kernel parameters.

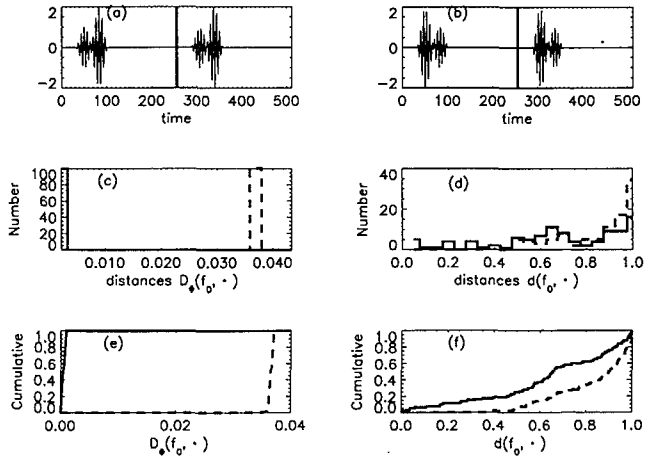

Figure 1: (a) Two realizations of signals of class 1, the two signals of lenght $L = 256$ being represented as two subsequent sections of one single signal. The noise parameter a has been set to 0. (b) The same for class 2. In (c) and (d) a histogram of the empirical distances for the two classes (class 1 solid line, class 2 dashed line) are plotted with D_Φ (using the estimated optimum kernel) and, on the other side, the Wigner distance d. (e) and (f) The empirical cumulative distribution of the distances, evaluated with D_Φ and d.

A second set of 100 realizations for each class served to test the resulting representation: One of the new signals of class 1 was chosen arbitrarily as reference signal f_0. For the estimated optimum kernel Φ, the distances $D_\Phi(f_0, \cdot)$ to each other signal of class 1 and class 2, respectively, were calculated. The distribution of these distances give an impression of how well the classes are separated.

For comparison we also calculated the distances $d(f_0, \cdot)$ where $d = D_{\Phi \equiv 1}$, i.e. the distances if we use the Wigner-Ville distribution.

As explained before, the distances $D_\Phi(f_0, f_i)$ and $D_\Phi(f_0, g_i)$ were calculated, using the estimated optimum kernel. f_0 was an arbitrary realization of class 1 serving as representative for this class. The Wigner-Ville distances $d(f_0, \cdot)$ have been calculated for comparison.

For each class we get an empirical distribution of the distances which is plotted as a histogram with a bin width of $\Delta D = \Delta d = 0.05$ in Fig. 1 (c) and (d).

The ability of the used distance measure (D_Φ or d) for classification can be seen more clearly in the empirical cumulative distribution (see Fig. 1, (e) and (f)). The maximum distance Δ between the two cumulatives is the classification quality. The difference $1 - \Delta$ corresponds to the classification error, if the above mentioned threshold is set to the value D_Φ or d at the point where the largest distance appears.

Several noise parameters a were used. In Fig. 2 and 3 we set $a = 0.2$ and $a = 0.5$, respectively. In each

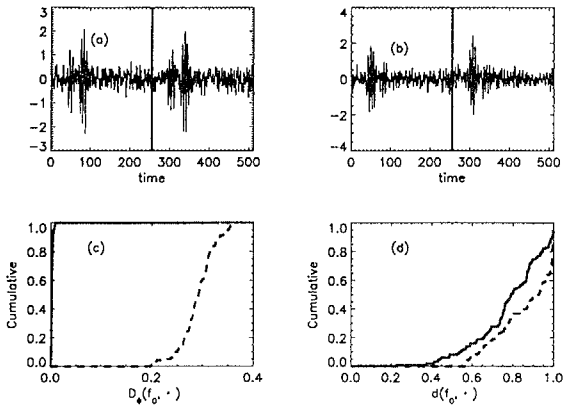

Figure 2: Analogous to Fig. 1, but with $a = 0.2$

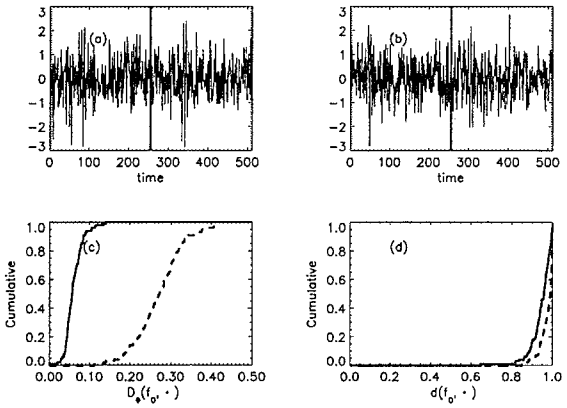

Figure 3: Analogous to Fig. 1, but with $a = 0.5$

case the optimum kernel was estimated. It can be seen that even in the case of highly noisy signals the two classes can be separated very well from each other if the time-frequency distance D_Φ with the optimum kernel is used, but the analogous with the Wigner-Ville distance fails even with $a = 0$.

5 Application on real data

In order to illustrate the second kernel optimization procedure we present a recent application from the field of acoustic quality control: after the baking of roof tiles the tiles have to be tested on fissures and cracks. Defective tiles have to be removed. Until now this task is done by human beings hitting the tile with a hammer and analyzing the resulting sound. A typical signal is shown in Fig. 4.

It was not possible to find discriminating features in the time or frequency domain by standard techniques. However, with optimized TFR and the presented classification scheme good classification rates were obtained.

The two classes consist of the good tiles and the defective ones. For each class 60 signals have been available. From each signal only the first 21 msec (corresponding to 512 data points) were used. The quantity $\tilde{m}_\Phi$ was estimated with 20 signals of each class. $\tilde{m}_\Phi$ for a region of the parameter space is shown in Fig. 5. The kernel parameters with the maximum $\tilde{m}_\Phi$ were found to be $\xi_0 = 0.031\, f_s, \tau_0 = 20\, f_s^{-1}$, where $f_s = 24$ kHz was the sampling rate.

With these kernel parameters an overall classification rate of about 91% has been achieved.

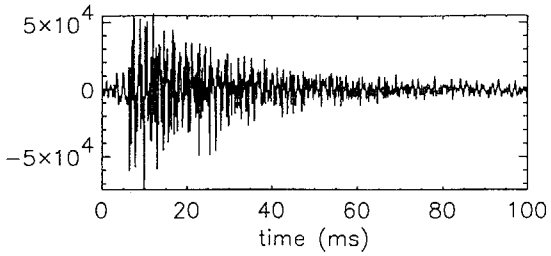

Figure 4: Example of the sound signal of a hit roof tile.

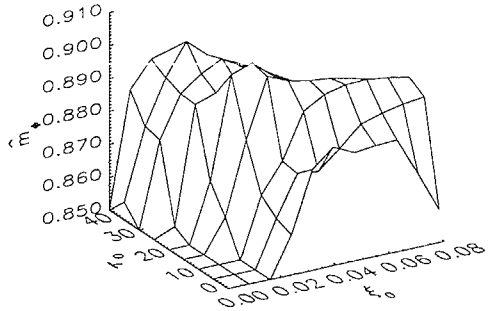

Figure 5: The quantity $\tilde{m}_\Phi$ for various values of the kernel parameters ξ_0 and τ_0.

References

[1] L. Cohen. Time-frequency–distributions – a review. *Proc. IEEE*, 77(7):941–981, July 1989.

[2] Richard O. Duda and Peter E Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, 1973.

[3] C. Heitz. Optimized time-frequency representations for the classification and detection of signals. *Applied Signal Processing.* submitted for publication.