Optimum time-frequency representations for the classification and detection of signals

Christoph Heitz
Fakultät für Physik
D-79104 Freiburg i. Br.

October 1994

Abstract

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.

In this article we develop a new theory of optimum TFRs for classification and detection problems, where two (or more) classes are to be discriminated from one another. Usually this is performed by mapping the signals into some representation space (e.g. a feature space) where a distinction is easily possible. If we regard TFRs as representations in time-frequency space we have to look for TFRs where all signals of one class are similar to each other but dissimilar to all signals of other classes. After introducing a quantitative measure for similarity or, in contrast, for the distance, a measure for the quality of a given TFR for a given classification or detection problem can be defined. Thus, by maximizing this quantity, the optimum TFR can be determined.

Furthermore, for the (realistic) case where the classes are of unknown structure and are only given by a random sample of realizations, an algorithm is presented which finds the optimum TFR automatically. The algorithm's flexibility is demonstrated by application on various types of synthesized signals and on a real problem from the field of acoustic quality tests in industrial production.

1 Introduction

When using time-frequency representations for signal analysis, there is one major problem which is still unsolved for the general case: given a class of signals and a problem to be solved, which is the best kernel to be used?

There have been some answers in the past (see e.g. [3, 17, 13, 12]), but each of them is restricted either to some very special signal classes (e.g. chirps) or to special problems.

Since it is impossible to find a TFR which is optimum for all signals and problems, it has been argued that one should use data adaptive locally optimum TFRs [11, 1, 2]. There, for each point of the time-frequency plane another TFR is used.

In this article we focus on the problem of globally optimum TFRs. We formulate this problem in the very general context of classification and detection. Given any sort of signals and a classification or detection problem to be solved, we want to find a unique optimum TFR for the analysis of the signal. Because of this general framework, the results of this article can be applied to a large variety of situations. Furthermore arbitrarily complex signals can be studied; a restriction on simple function classes is not necessary.

Why does one use TFRs? The answer is that some properties of the signal appear much clearer in a mixed time-frequency representation than in the simple time or frequency representation[1]. In the literature one finds many derivations of kernels which are optimized to extract special predefined properties of the signal, e.g. the chirp rate or the instantaneous frequency (see e.g. [4] for a review). However, these discussions are restricted to those very special cases where the property to be examined is known in advance and can be formulated in a simple way. Furthermore the class of admissible signals is restricted to simple signals like monocomponent ones.

Frequently, when analyzing more complex signals, it is far form obvious, which property is the relevant one. Furthermore the structure of the signals is often unknown. Thus, it can neither be decided in advance which kernel fits to the problem, nor whether at all the TFR yields a better representation of the signal than the simple time or frequency representation. Even if we know that there is a TFR which leads to a good signal representation, the task remains to determine the corresponding kernel function.

Clearly, the question "what is a good signal representation?" depends on the problem and cannot be answered in general. However, frequently the task of the analysis is a *comparison* between several signals. Usually this comparison is made via some sort of distance measure or similarity. This is the case for (e.g.)

- **Classification:** members of one class of signals are similar to each other

[1] The term "frequency representation" is used to denote the Fourier transform of a signal, in order to emphasize that this is nothing else than another representation of the signal.

but not similar to the members of other classes.

- **Detection:** Here we look for signals which are similar to a reference.
- **Detection of outliers:** The task is to detect signals which are *not* similar to a reference.

In this cases a good signal representation is a representation where all signals of one class are similar to each other, but dissimilar to all signals of other classes. In order to formalize what is meant by the intuitive notion of similarity one defines a distance measure where a large distance means little similarity and vice versa.

In classification problems, the classification proceeds in two steps: first a number of features are determined out of the signal (feature extraction). If these features are quantitative ones, each signal is then represented by a set of feature values (feature vector) in the feature space. Then each signal to be classified is assigned to one of the classes by a classification rule (which is sometimes called *classificator*). Frequently, the classification rule is based on the evaluation of some sort of distance between the signal and representatives of the different classes in the feature space. The classificators differ by the definition of their distance measure [8]. It is a fundamental problem that in order to find a good feature set one has to know in advance which properties of the signals are likely to be important for the classification.

In this article, instead, we regard nonparametric representations, especially TFRs. The signal is represented by a two-dimensional function rather than by a feature vector. The similarity of two TFRs is measured via an inner product of TFRs. Our aim is to find the optimum TFR in order to fulfill the above mentioned tasks.

The article is organized as follows: In section 2 the general framework of TFRs of Cohen's class is introduced. In section 3 the concept of similarity and distance when using TFRs is established, and a distance measure is defined. Some of its properties are derived. Section 4 establishes an optimality measure, which allows to judge the ability of a given TFR to solve a classification or detection problem. Thus, it is possible to find the optimum TFR in Cohen's class of TFRs. Section 5 deals with the problem of finding the optimum TFR if the structure of the function classes is not known, but rather each class is given by a random sample of functions. An algorithm to find the optimum TFR is given. In section 6 some examples serve to illustrate the developed theory and demonstrate the flexibility of the optimization algorithm by the application to classification or detection problems of some very different types of signals. In section 7 an application to a problem from the field of acoustic quality tests is shown. Finally, in section 8, two further remarks on the developed theory are made.

2 Representations of signals in the time-frequency domain

A large class of mixed time-frequency representations of a signal $f(t)$ is given by Cohen's class of TFR [4]:

$$C_f(t, \omega; \Phi) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A_f(\xi, \tau)\, \Phi(\xi, \tau)\, e^{-i(\xi t + \tau\omega)}\, d\xi\, d\tau$$

with the *ambiguity function*

$$A_f(\xi, \tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\mu + \frac{\tau}{2})\, f^*(\mu - \frac{\tau}{2})\, e^{i\xi\mu}\, d\mu.$$

The representation is parametrized by the twodimensional kernel $\Phi(\xi, \tau)$.

The equations for the corresponding discrete time case are straightforward and can be found e.g. in [7]. The only difficulty arises with the integration bounds of ξ in the evaluation of the ambiguity function. It is a consequence of the Nyquist theorem for TFRs that, when replacing the integration over ξ by a sum over ξ, the summation has to be performed in the intervall $[-\pi/4, \pi/4]$, unless we use the analytical signal instead of $f(t)$ itself. In this case we have the restriction $|\xi| \leq \pi/2$.

Some important properties of the representations of Cohen's class are [4]:

- **Time shift property:**
  If $\tilde{f}(t) = f(t - t_0)$ (time shift) then

  $$C_{\tilde{f}}(t, \omega; \Phi) = C_f(t - t_0, \omega; \Phi).$$

  This holds for every kernel Φ.

- **Frequency shift property:**
  If $\tilde{f}(t) = f(t)\, e^{i\omega_0 t}$ (frequency shift) then, for all kernels Φ,

  $$C_{\tilde{f}}(t, \omega; \Phi) = C_f(t, \omega - \omega_0; \Phi).$$

- **Kernel restrictions:**
  The kernel $\Phi(\xi, \tau)$ of the TFR is not subjected to any restrictions but may be chosen arbitrarily. We assume, however, that the kernel is well-behaved enough in order to allow the subsequent calculations. Furthermore we allow only kernels that lead to a real TFR. This is the case if

  $$\Phi(\xi, \tau) = \Phi^*(-\xi, -\tau). \qquad (1)$$

- **Non-uniqueness:**
  In general the signal cannot be recovered out of its TFR, although this is possible for some kernels [10]. Usually this fact is regarded as a disadvantage, because information is lost. However, in our context it may be advantageous, for if the dropped information is irrelevant to the problem, its suppression improves the classification or detection.

3 Similarity and Distance Measures

In this section the concept of similarity and distance will be further developed and made more precise.

In order to keep the subsequent equations simple, we introduce the following notation. Let Ω_1 be some linear function space (e.g. $L_2(I)$, where I is an interval, e.g. $I = [0, 1]$). Given a fixed kernel Φ, the set of the TFRs of all functions in Ω_1 is a subset of all twodimensional functions (but not a linear space itself). For other kernels we obtain other sets of twodimensional functions. The smallest linear space which encloses all these sets, is called Ω_2. Thus, all TFRs of functions from Ω_1 (regardless of the used kernel) lie in Ω_2.

$$f(t) \in \Omega_1 \Rightarrow C_f(t, \omega; \Phi) \in \Omega_2.$$

Note however, that in general a function B in Ω_2 has no counterpart in Ω_1, i.e. there need not be a function $f \in \Omega_1$ for which $B(t, \omega) = C_f(t, \omega; \Phi)$.

In Ω_1 and Ω_2 the inner products are defined as usual[2]:

$$< f_1, f_2 > \;=\; \int f_1(t) f_2^*(t)\; dt \tag{2}$$

$$< C_1, C_2 > \;=\; \int\int C_1(t, \omega; \Phi) C_2(t, \omega; \Phi)\; dt\, d\omega. \tag{3}$$

The discrete time formulation is defined analogously.

The norm of a function in Ω_1 or Ω_2 is defined by

$$\|f\|^2 =< f, f > \qquad \text{and} \qquad \|C\|^2 =< C, C > .$$

In the following we will assume that all functions have $\|f\| = 1$ (in Ω_1).

[2]We use the same symbol $< \cdot, \cdot >$ for both the inner product in Ω_1 and in Ω_2. It will be clear from the context which one is meant.

3.1 Similarity and Distance in Ω_1 and Ω_2

A natural measure of similarity is the inner product[3]. Regarding functions in Ω_1 it is given by eq. (2), but may be expressed also in frequency space

$$< f_1, f_2 >= \frac{1}{2\pi} \int \tilde{f}_1(\omega)\tilde{f}_2^*(\omega)\, d\omega, \tag{4}$$

where $\tilde{f}(\omega)$ is the Fourier transform of $f(t)$.

The contrary of similarity is distance. A measure can be introduced by

$$d(f_1, f_2) = 1 - | < f_1, f_2 > |^2$$

for functions with $\|f\| = 1$.

Since we are interested in the inner structure of the signals and not in the phase, we require that $d(f, e^{i\varphi} f) = 0$ for all $\varphi \in [0, 2\pi]$. This implies the use of the modulus $| < f_1, f_2 > |$ of the inner product. Furthermore the use of the modulus results in a real quantity, which is necessary for comparing distances.

These considerations can be translated straightforward into the TFR space Ω_2. Here we have the inner product $< C_{f_1}, C_{f_2} >$, which can be interpreted as some sort of similarity measure of the functions f_1 and f_2 itself:

$$\tilde{c}_\Phi(f_1, f_2) =< C_{f_1}, C_{f_2} > .$$

In order to have the same situation as before the TFRs should be normalized, thus leading to

$$c_\Phi(f_1, f_2) =< C^N_{f_1}, C^N_{f_2} >$$

where $C^N_{f_i} = \frac{C_{f_i}}{\|C_{f_i}\|}$ are the normalized TFRs. To ensure the possibility of normalization, we require that $\Phi(0, 0) = 1$, thus $\|C_f\| \neq 0$ for all $f \neq 0$. If we omitted this restriction, it could happen that, for some kernels, $C_f = 0$ even if $f \neq 0$.

Instead of the normalized functions f_1 and f_2, we now consider the normalized TFRs of these functions, regarding them as special representations of the functions.

Again it is possible to construct a distance measure

$$D_\Phi(f_1, f_2) = 1 - c_\Phi(f_1, f_2) = 1- < C^N_{f_1}, C^N_{f_2} > . \tag{5}$$

Note that it is not necessary to use the modulus of $< C^N_{f_1}, C^N_{f_2} >$ because, firstly, the TFR C_f of a function is by definition independent of phase factors $e^{i\varphi}$: $C_{f(t)} = C_{e^{i\varphi} f(t)}$. Thus, the requirement $D_\Phi(f, e^{i\varphi} f) = 0$, for all φ, is fulfilled anyway. Secondly, since we restrict our discussion on real TFRs, the quantity $< C^N_{f_1}, C^N_{f_2} >$ is already real, in contrast to $< f_1, f_2 >$.

[3]In the case of zero mean functions the inner product of normalized functions is equivalent to the correlation. Therefore the inner product is a natural measure for the similarity of stochastic signals.

Note that both the inner product and the distance depends explicitely on the choice of the kernel Φ. The central idea of this article is that it is exactly this property which may be taken advantage of in classification and detection problems. The freedom in choosing an arbitrary kernel makes it possible to utilize exactly that distance measure which is best fitted both to the problem and the data.

3.2 Properties of the Distance D_Φ

In this section some important properties of the distance measure D_Φ will be derived.

First it should be stressed that the distance measure D_Φ is not a metric in the ordinary mathematical sense since it does not in general possess the property $D_\Phi(f_1, f_2) = 0 \Rightarrow f_1 = f_2$. This follows from the non-uniqueness of the TFR (see section 2).

On the other hand, for each kernel Φ, Schwartz's inequality

$$D_\Phi(f_1, f_2) \leq D_\Phi(f_1, f_3) + D_\Phi(f_3, f_2) \text{ for each} f_1, f_2, f_3$$

is fulfilled, since we can write

$$D_\Phi(f_1, f_2) = \frac{1}{2} \|C^N_{f_1} - C^N_{f_2}\|^2$$

Hence, D_Φ corresponds to the usual Euclidian distance of the normalized TFRs in Ω_2 where it is known that Schwartz's inequality holds.

Second, because of the conservation of the inner product under Fourier transformations, it holds that

$$\int\int C_{f_1}(t,\omega) C_{f_2}(t,\omega)\, dt\, d\omega = \frac{1}{4\pi^2} \int\int A_{f_1}(\xi,\tau)\Phi(\xi,\tau)\; A^*_{f_2}(\xi,\tau)\Phi^*(\xi,\tau)\, d\xi\, d\tau,$$

or, in symbolic notation

$$< C_{f_1}, C_{f_2} >= \frac{1}{4\pi^2} < A_{f_1}\Phi, A_{f_2}\Phi >= \frac{1}{4\pi^2} < A_{f_1}, |\Phi|^2 A_{f_2} >$$

Hence, for c_Φ and D_Φ, only the modulus of Φ is relevant. So there is no loss of generality if we use only real and positive kernels.

The third property regards the connection between the distance d in Ω_1 and D_Φ in Ω_2. Because of Moyal's formula

$$< W_{f_1}, W_{f_2} >= | < f_1, f_2 > |^2$$

where $W_f(t,\omega)$ is the Wigner distribution (kernel $\Phi \equiv 1$), we have the relation $c_{\Phi\equiv 1}(f_1, f_2) = | < f_1, f_2 > |^2$ and thus

$$D_{\Phi\equiv 1}(f_1, f_2) = d(f_1, f_2)$$

For the Wigner distribution it follows that D_Φ is the same as the usual distance d in Ω_1. The introduction of the above kernel dependent distance measure is therefore a generalization of the usual distance measure in Ω_1 which includes the latter. For detection and classification this result is very important, because it means that a clustering of a set of functions in Ω_1 (all distances of two of the functions are small) is represented equally well in the Wigner TFR of the functions. Every algorithm, which uses distance measures constructed with inner products in the function space Ω_1, works equally well when using the Wigner TFR and inner products in Ω_2 instead. Often, however, it will be possible to improve the performance of the algorithm by finding a *better* TFR.

4 Optimum TFRs

In the last section a distance measure has been introduced, which uses the representation of the signal in the time-frequency domain and depends explicitely on the used kernel function. In this section a quantitative optimality criterion will be established in order to find the optimum kernel for detection and classification problems.

First we consider the classification task. We assume that there are two classes of functions. Each class is defined by a specified region R_1 and R_2, respectively, in the function space Ω_1, and a probability distribution P_i which may be uniform in R_i, but may also have another form. Outside the region R_i, the probability distribution P_i vanishes.

In order to separate the classes from each other, we look for the TFR that yields the best distinction between these regions. As before we will restrict our discussion to normalized functions.

A natural measure for the ability to discriminate between the two function classes is

$$m = \frac{d(R_1, R_2)}{d(R_1, R_1) + d(R_2, R_2)}$$

$where d(R_1, R_2)$ is the expectation value of the distance between members of class 1 and class 2, respectively:

$$d(R_1, R_2) = \overline{d(f_1, f_2)} = \int_{R_1} \int_{R_2} P_1(df_1) P_2(df_2)\ d(f_1, f_2)$$

Here $\overline{d(f_1, f_2)}$ denotes means the expectation value of the distance $d(f_1, f_2)$ of f_1 (belonging to R_1) and f_2 (out of R_2), measured with some distance measure. Thus, $d(R_1, R_2)$ is a measure of the distance between the function classes. Analogously, $d(R_1, R_1)$ is the mean distance of two functions in class 1, yielding a measure of the extension of R_1 in the function space.

A large value of m means that the function classes have a large distance in the function space, compared with the extension of the two classes separately.

Thus, the classes can well be discriminated. If $R_1 = R_2$ and $P_1(\cdot) = P_2(\cdot)$, then no discrimination is possible, which corresponds to $m = 0.5$. A m-value of 10 indicates that the mean distance between the classes is 20 times larger than the mean inner class distance $1/2\,[d(R_1, R_1) + d(R_2, R_2)]$.

The quantity m is similar to the Fisher criterion in discriminance analysis [6, 8] for the case of functions on the unit sphere. If, furthermore, the inner class distances $d(R_i, R_i)$ are equal, m resembles the test quantity of the t-test [14]. There, a large test quantity means that the classes are well separated compared to their variances.

Regarding the TFRs of the functions, instead of the two regions R_1 and R_2 we consider now the corresponding regions S_1 and S_2, given by $S_1 = C^N_{R_1}$ and the induced probability distributions. The distance is given by D_Φ according to eq. (5), where we set $C_1 \equiv C^N_{f_1}$ for convenience.

With

$$\begin{aligned}
\overline{D_\Phi(f_1, f_2)} &= 1 - \overline{< C_1, C_2 >} \\
&= 1 - < \overline{C_1}, \overline{C_2} > \\
\overline{D_\Phi(f_1, f_1')} &= 1 - \overline{< C_1, C_1' >} \\
&= 1 - < \overline{C_1}, \overline{C_1} > \\
&= 1 - \|\overline{C_1}\|^2
\end{aligned}$$

the above introduced discrimination measure is then

$$m_\Phi = \frac{1 - < \overline{C_1}, \overline{C_2} >}{(1 - \|\overline{C_1}\|^2) + (1 - \|\overline{C_2}\|^2)}$$

Clearly, this discrimination measure depends on the kernel Φ and can thus be maximized by chosing the adequate kernel. However, since the kernel enters in a complicated way, an analytical optimization is only possible in very simple cases.

So far the results for classification problems involving two classes[4]. In detection problems, in general the appearance of a signal in some sort of background (e.g. white noise) is to be detected. In our context we regard the two function classes R_1 and R_2, where R_1 is given by all possible realizations of the signal to be detected together with the background, whereas R_2 consists of the possible realizations of the background alone.

For a good detection we should maximize

$$\tilde{m}_\Phi = \frac{D_\Phi(S_1, S_2)}{D_\Phi(S_1, S_1)} = \frac{1 - < \overline{C_1}, \overline{C_2} >}{1 - \|\overline{C_1}\|^2} \tag{6}$$

instead of m_Φ, where again $\overline{C_i}$ is the expectation value of the considered function class.

[4]This concept can be generalized for the case of more classes, too. This, however, is beyond the scope of this article.

The difference betweeen m_Φ and $\tilde{m}_\Phi$ is, that in the latter the inner class distance $D_\Phi(S_2, S_2)$ does not appear. This reflects the fact that, for detection purposes, the extension of the second class in Ω_2 is not relevant and may be arbitrary large, in contrast to the classification case. It suffices to have a small distance within class 1 compared with the typical distance between class 1 and class 2.

Given a certain level γ a function f is assumed to be in the considered class if

$$D_\Phi(f, S_1) = 1- < C_f, \overline{C_1} > \; < \gamma.$$

Clearly this scheme can also be used for the detection of outliers.

5 Data dependent optimization of the kernel

In principle, one could maximize m_Φ or $\tilde{m}_\Phi$ analytically, if the structure of the function classes were known. However, this is not an easy task even in simple cases.

On the other hand, the advantage of the proposed new concept is that it can generalized very easily to the case of data dependent optimization of the kernel. This is of great interest, since in relistic problems, there is often no simple model available for the structure of the data, so an analytical derivation of the optimum kernel is not possible. Instead, some realizations of signals are given and one has to estimate the optimum kernel from this sample of realizations.

In our framework it is straightforward how to do this, because the discrimination measures m_Φ and $\tilde{m}_\Phi$ involve only expectation values of TFRs. Given some realizations of functions of each class and a kernel function Φ, the means of the classes can be estimated and thus we get an estimate for the discrimination quality. Maximizing this quantity yields the optimum TFR.

Let us rewrite the full equation for m_Φ, for the given realizations $f_i^1, i = 1, \ldots, N$ of class 1 and $f_j^2, j = 1, \ldots, M$ of class 2:

$$m_\Phi = \frac{1- < \frac{1}{N} \sum\limits_{i=1}^{N} \frac{C_{f_i^1}}{\|C_{f_i^1}\|}, \frac{1}{M} \sum\limits_{j=1}^{M} \frac{C_{f_j^2}}{\|C_{f_j^2}\|} >}{(1 - \frac{1}{N^2} \left\| \sum\limits_{i=1}^{N} \frac{C_{f_i^1}}{\|C_{f_i^1}\|} \right\|^2) + (1 - \frac{1}{M^2} \left\| \sum\limits_{j=1}^{M} \frac{C_{f_j^2}}{\|C_{f_j^2}\|} \right\|^2)}$$

In order to illuminate how the dependance on the kernel Φ, this equation can be expressed in the ambiguity funcion plane (using the conservation of the inner product under Fourier transformations) by

$$m_\Phi = \frac{1- < \frac{1}{N} \sum\limits_{i=1}^{N} \frac{A_{f_i^1}\Phi}{\|A_{f_i^1}\Phi\|}, \frac{1}{M} \sum\limits_{j=1}^{M} \frac{A_{f_j^2}\Phi}{\|A_{f_j^2}\Phi\|} >}{(1 - \frac{1}{N^2} \left\| \sum\limits_{i=1}^{N} \frac{A_{f_i^1}\Phi}{\|A_{f_i^1}\Phi\|} \right\|^2) + (1 - \frac{1}{M^2} \left\| \sum\limits_{j=1}^{M} \frac{A_{f_j^2}\Phi}{\|A_{f_j^2}\Phi\|} \right\|^2)}$$

This quantity is to be maximized by choosing an adequate kernel Φ.

If this has to be done in concrete cases, a general problem arises: the kernel $\Phi(\xi, \tau)$ possesses infinitely many degrees of freedom in continuous formulation or $O(L^2)$ degrees of freedom in discrete notation, where L is the length of the examined time series $f_i(t), i = 1, \ldots, L$. In practice, only a finite number of realizations is available. Thus, the maximization problem is ill–posed. Furthermore it is well known from statistics that a classificator which has too many degrees of freedom (compared with the available number of data) tends to overfit the test set and looses its ability to generalize. Even if the two test sets that have been used to adapt the parameters of the classificator can be separated very well, a new, unknown data point cannot be classified correctly.

Thus, the kernel has to be parametrized by few parameters $\alpha_k, k = 1, \ldots, p$ which determine its shape: $\Phi = \Phi(\xi, \tau; \alpha_k)$. The maximization of m or $\tilde{m}$ can then be expressed as a maximization with respect to the parameters α_k. However, the introduction of kernel parameters reduces the class of admitted kernels, thus rather a suboptimum solution than the real optimum solution of the problem is found by the maximization.

6 Application on simulated data

6.1 TFR versus time representation

This example demonstrates how the use of an adequate TFR leads to a better representation of the signal, if we want the signals of one signal class to be similar to each other.

We consider the case of functions $f(t)$ consisting of two subsequent wave packets where the second packet is of variable phase (see Fig. 1)

$$f^{\varphi}(t) = h(t)e^{i\omega t} + e^{i\varphi}h(t - t_0)e^{i\omega t}. \tag{7}$$

Here t_0 is the time difference between the two packets and $h(t)$ is a real and positive envelope function with $\|h\|^2 = 1/2$. If we further assume that the envelopes have negligible overlap ($< h(t), h(t - t_0) > \approx 0$), then it holds that $\|f^{\varphi}\|^2 = 1$.

We look for a representation that is insensitive for the variations of φ, i.e. the representations of different f^{φ} should be similar to each other. Thus we require $d(f^{\varphi_1}, f^{\varphi_2}) \ll 1$ for all φ_1, φ_2 or, using f^0 as reference signal, $d(f^0, f^{\varphi}) \ll 1$ for all φ in $[0, 2\pi]$.

A simple calculation shows that the inner product of f^0 and f^{φ} is given by

$$< f^0, f^{\varphi} >= \frac{1}{2}(1 - e^{i\varphi})$$

For $\varphi = \pi$, we get $< f^0, f^{\pi} >= 0$, leading to $d(f^0, f^{\pi}) = 1$. For $\varphi = 0$ we have $d(f^0, f^0) = 0$, as expected. Thus, the time representation is *not* insensitive to variations of φ.

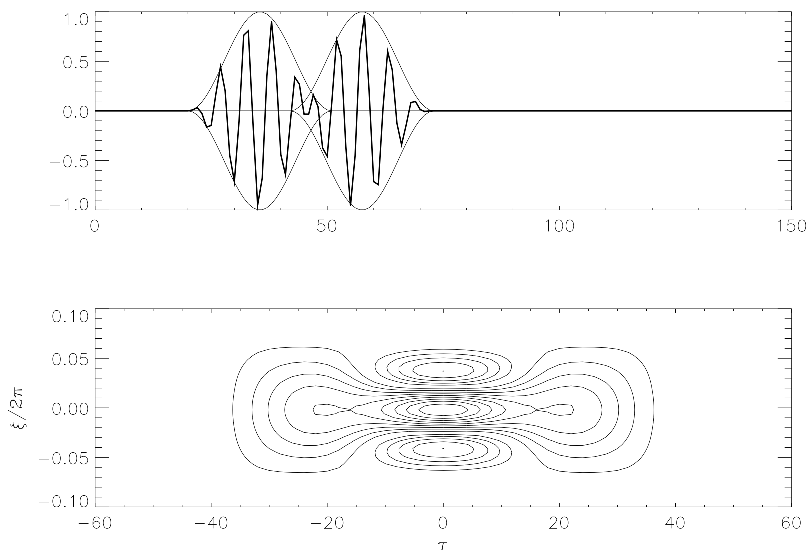

Figure 1: (a) Real part of a signal as in eq. (7) with $t_0 = 21$ and $\omega/2\pi = 0.18$. The envelope $h(t)$ is a hanning window of width 30. (b) The modulus of the corresponding ambiguity function. The crossterms at $(0, \pm t_0)$ can be seen clearly, even if eigenterms and crossterms are not separated very well. The complicated structure of the eigenterm region around the origin results from the superpositon of the two eigenterms. Only the crossterms depend on the phase φ.

The situation is better in the time-frequency domain. The TFR of f^φ consists of the two eigenterms $C_{h(t)e^{i\omega t}}$ and $C_{h(t-t_0)e^{i\omega t}e^{i\varphi}}$ and the two crossterms (see Appendix A). However, the second eigenterm is independent of φ, since the TFR of a function is independent of an overall phase factor. Thus, if we could suppress the crossterms we would have

$$\begin{aligned} < C_{f^0}, C_{f^\varphi} > &= 1 \\ D_\Phi(f^0, f^\varphi) &= 1- < C_{f^0}, C_{f^\varphi} > = 0 \qquad \text{for all } \varphi \end{aligned}$$

as desired.

In the ambiguity function plane the crossterms are located at $(\xi, \tau) = (0, \pm t_0)$, so if we use a kernel $\Phi(\xi, \tau) = g(\tau)$ with $g(\tau) = 0$ for $|\tau| > t_0/2$ the crossterms are indeed suppressed. Of course, if we use the Wigner kernel $\Phi \equiv 1$ then $D_\Phi(f^0, f^\pi) = 1$ as in the time representation, since no crossterm reduction takes place. This behaviour is shown in Fig. 2.

The case of varying phase difference of a signal consisting of two waveforms can be carried over to the case of varying time difference t_0 because if the frequency is not too low compared with the time scale of the envelope $h(t)$ then a small

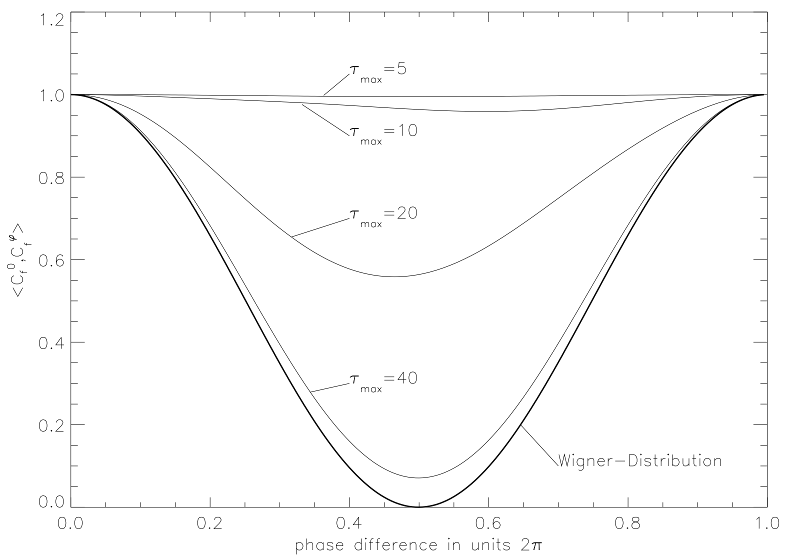

Figure 2: Inner product $< C_{f^0}, C_{f^\varphi} >$ of a signal as in eq. 7 with $t_0 = 21$ and $\omega/2\pi = 0.18$. The envelope $h(t)$ is a hanning window of width 30. C is the Wigner distribution (thick line) with $\Phi \equiv 1$ and a TFR with the kernel $\Phi(\xi, \tau) = g(\tau)$ (thin lines), where $g(\tau)$ is a hanning window of width $2\tau_{\max}$ with various values of $\tau_{\max}$. It can be seen that the larger $\tau_{\max}$ is, the more the crossterms are involved and the inner product drops to 0 for $\varphi = \pi$. As long as $\tau_{\max}$ is less than $t_0/2$ essentially only the inner product of the eigenterms is evaluated which is independent of the phase φ.

variation in t_0 with fixed φ leads to nearly the same waveform as a variation in φ with fixed t_0. Thus, also in the case of a varying t_0 a TFR with the above derived kernel is a better signal representation than the original time represention if we want the representation to be insensitive to these variations.

6.2 Optimization of the Kernel

In simulations we applied the above introduced concept of kernel optimization to discrete time functions $f(t), t = 1, \ldots, L$ with $L = 256$. We performed 100 realizations of each function class to estimate the kernel parameters. Time and frequency variables are given by dimensionless numbers.

For the below described examples we used only two parameters τ_0, ξ_0 as kernel parameters according to

$$\Phi(\xi, \tau) = e^{-\left(\frac{\xi}{\xi_0}\right)^2} e^{-\left(\frac{\tau}{\tau_0}\right)^2} \tag{8}$$

This parametrization yields the kernel for the smoothed Pseudo Wigner distribution.

In order to avoid kernels with $\tau_0 = \xi_0 = 0$, in which case the measure m_Φ is indefined, we have introduced lower bounds τ_l and ξ_l for the parameters which have been set to small positive numbers.

Evaluation of the inner product of the TFR of two functions with this kernel includes two interesting limiting cases which can be expressed in the time and frequency domain, respectively.

For $\tau_0 = 0$ we get $\Phi(\xi, \tau) = \delta(\tau)\, e^{-\left(\frac{\xi}{\xi_0}\right)^2}$. It can be shown (see Appendix B) that for $\xi_0 = \infty$

$$< C_\Phi f_1, C_\Phi f_2 > = < |f_1(t)|^2, |f_2(t)|^2 >$$

For $\xi_0 < \infty$, instead of the inner product of the $|f_i(t)|^2$ we get the inner product of smoothed versions of $|f_i(t)|^2$ where the smoothing is stronger when ξ_0 is smaller. Thus, in this case we compare the (more or less smoothed) squared modulus of the functions in order to quantify their similarity.

Analogously, for $\xi_0 = 0$, we get the result that the inner product of the TFRs equals the inner product of (smoothed) versions of $|\tilde{f}_1(\omega)|^2$ and $|\tilde{f}_2(\omega)|^2$.

We will refer to these results below.

Recalled from section 3 that for $\xi = \tau = \infty$ we get another special case:

$$< C_f, C_g > = | < f, g > |^2$$

As a first example we used the two function classes

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

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. For $a = 0$ this is an example of two nearly non-overlapping wave packets as described in section 6. In figure 3 (a) and (b) two realizations of each signal class are shown.

The two classes differ in their temporal structure (the first class is the time reversed version of the second), but there is no distinctive information in the frequency. Thus, it is clear that the optimum kernel has $\tau_0 = 0$, because in this case the TFR is smoothed as strongly as possible in frequency direction, suppressing all frequency information of the TFR at each time point. Setting τ_0 to a finite value renders no additional information which may be valuable for a distinction of the classes. At the same time it reduces the equality within each class, leading to a reduced m_Φ.

This behaviour can be interpreted in the ambiguity function plane, too. For increasing τ_0 more and more the crossterms at $(\pm\Delta t, 0)$ are involved which reduce the inner product of the ambiguity functions (multiplied with the kernel) of two functions of the same class (see section 6).

The optimum kernel that is found by the algorithm when maximizing the measure m_Φ with $a = 0$ is indeed given by $\tau_0 = \tau_l, \xi = 4.62u$, with $u = 2\pi \cdot 10^{-3}$, confirming the above considerations. The discrimination measure m with these kernel parameters have been found to be $m = 406112$, yielding an excellent discrimination of the two classes.

Since $\tau_0 = 0$ is very small, one of the limiting cases is reached. Hence, the resulting procedure can be easily interpreted: it just compares appropriately smoothed versions of $|f(t)|^2$ and $|g(t)|^2$. The smoothing yields the envelopes of the signals which are similar inside one class, but dissimilar to the envelopes of a function of the other class.

One of the realizations of class 1, which is referred to as f_0, was chosen as representative for this class. For each realization the distances $D_\Phi(f_0, f_i)$ and $D_\Phi(f_0, g_i)$ were calculated, using the estimated optimum kernel. For comparison, also the distance measure d was used for the calculation of the distances. 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. 3 (c) and (d). The classification rule is straightforward: a signal f is supposed to be in class 1 if the distance $D(f_0, f)$ is smaller than a threshold which has to be chosen such that the classification error is minimized.

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. 3, (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. 4, $a = 0.2$ was chosen. This leads to the kernel parameters $\tau_0 = \tau_l, \xi = 16.8u$ and a distinction measure of $m = 37$. The additional noise term leads to a reduction of the discrimination, but also to an enlarged kernel. This is even more obvious in the case $a = 0.5$ (Fig. 5) where we get $\xi_0 = 22.4u$ and $m = 2.6$.

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 in the time domain fails even with $a = 0$.

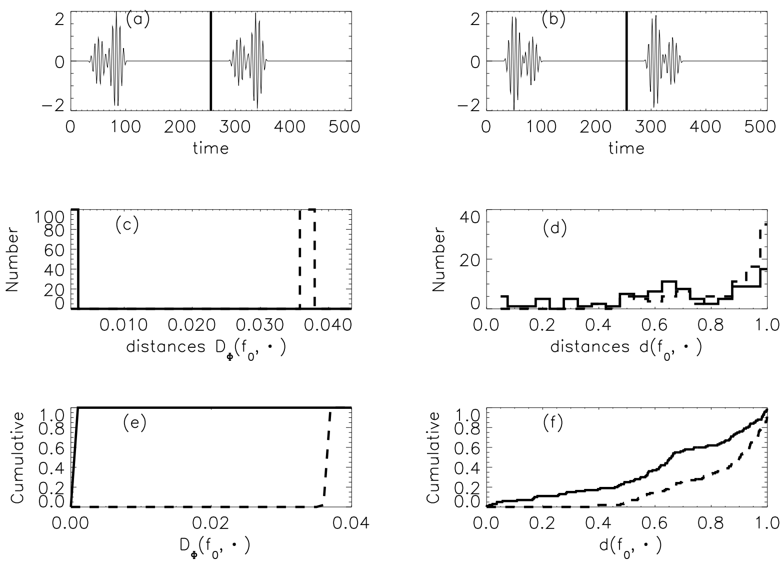

Figure 3: (a) Two realizations of signals of class 1, the two signals of lenght $L = 256$ being represented as two subsequent sections of one single signal. (b) The same for class 2. In (c) and (d) the empirical distance distributions for the two classes are plotted (see text) with D_Φ and the estimated optimum kernel and, on the other side, the distance measure d. (e) and (f) The empirical cumulative distribution of the distances, evaluated with D_Φ and d. The noise parameter has been set to 0.

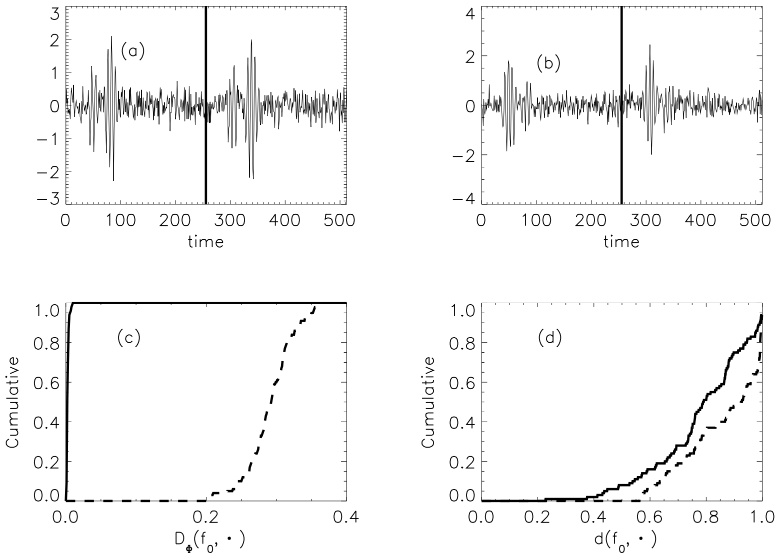

Figure 4: Analogous to Fig. 3, but with $a = 0.2$

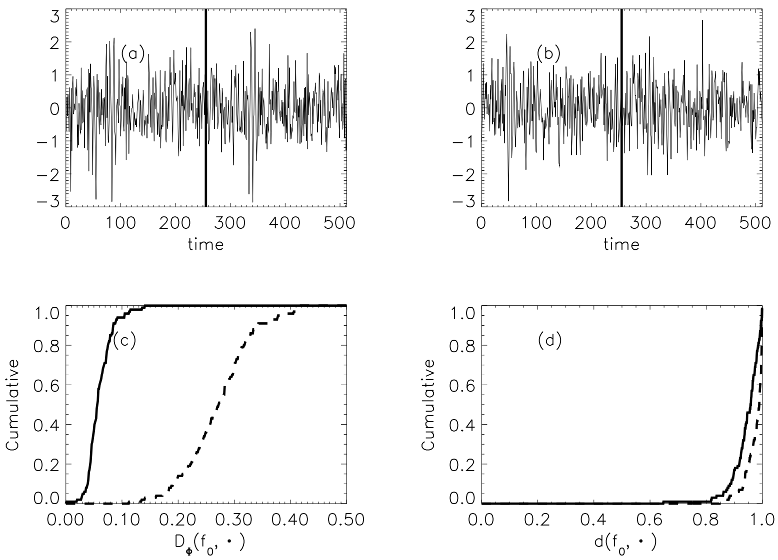

Figure 5: Analogous to Fig. 3, but with $a = 0.5$

In the second example we introduce an additional frequency structure by defining

$$\begin{aligned} f(t) &= h(t)\sin(\omega_1 t + \varphi_1) + 2h(t-\Delta t)\sin(\omega_2 t + \varphi_2) + a\epsilon(t) \\ g(t) &= 2h(t)\sin(\omega_2 t + \varphi_3) + h(t-\Delta t)\sin(\omega_1 t + \varphi_4) + a\epsilon(t) \end{aligned}$$

with $\omega_1/2\pi = 0.15$ and $\omega_2/2\pi = 0.075$. The noise parameter was set to $a = 0.2$. Again the phases are random, $h(t)$ and Δt are chosen as in the first example. For the optimum kernel we obtained $\xi = 21.4u, \tau_0 = 3.44$, leading to a discrimination of $m = 42.2$. Here $\tau_0 > 0$ as expected, since the information about the frequencies can be used additionally to separate between the two classes. Again the separation of the two classes is possible in TFR space, using the distance measure D_Φ with the optimum kernel (see figure 6).

In the third example the two classes are given by two purely stochastic processes, namely AR(2)-processes

$$\begin{aligned} f(t) &= 1.6373631\, f(t-1) - 0.67\, f(t-2) + \epsilon(t) \\ g(t) &= 1.6373197\, g(t-1) - 0.67\, g(t-2) + \epsilon(t) \end{aligned}$$

where $\epsilon(t) \sim N(0,1)$.

The spectrum of each process is given by a peak at frequency $\omega/2\pi = 0.1$ and $\omega/2\pi = 0.12$, respectively. Since the processes are stationary no distinctive information can be found in the temporal structure. Thus, we expect the optimum kernel to have $\xi_0 = 0$. Maximization of m_Φ with respect to the parameters renders indeed $\tau = 1.12, \xi_0 = \xi_l$. With these parameters a discrimination of $m = 12.4$ is achieved (see Fig. 7).

Again, as in the first example, since $\xi_0 \approx 0$ one of the above mentioned limiting cases is realized: for the evaluation of D smoothed versions of $|\tilde{f}(\omega)|^2$ are used. This is exactly what one would do intuitively when comparing stationary stochastic processes where all the second order properties of the process are contained in the spectrum. In order to estimate the spectrum of a stochastic process the periodogram $|\tilde{f}(\omega)|^2$ has to be smoothed, but, of course, it should not be smoothed too much if we want to separate the two classes. Thus, there exists an optimum smoothing which is found by our algorithm. In Fig. 7 the signals and the empirical cumulative distribution of $D(f_0, f_i)$ and $D(f_0, g_i)$ for all realizations $f_i, g_i, (i = 1\ldots, 100)$ are shown with an arbitrary member f_0 as representative of the first class. Here again, in the TFR a separation of the classes is possible whereas in the time domain it fails.

The forth example deals with the detection of a signal of the form

$$f(t) = h(t)\sin(\omega t + \varphi) \tag{9}$$

in white noise. Here $h(t)$ is a hanning window with width 230, the phase φ and the frequency ω are random numbers, uniformely distributed in $[0, 2\pi]$ and

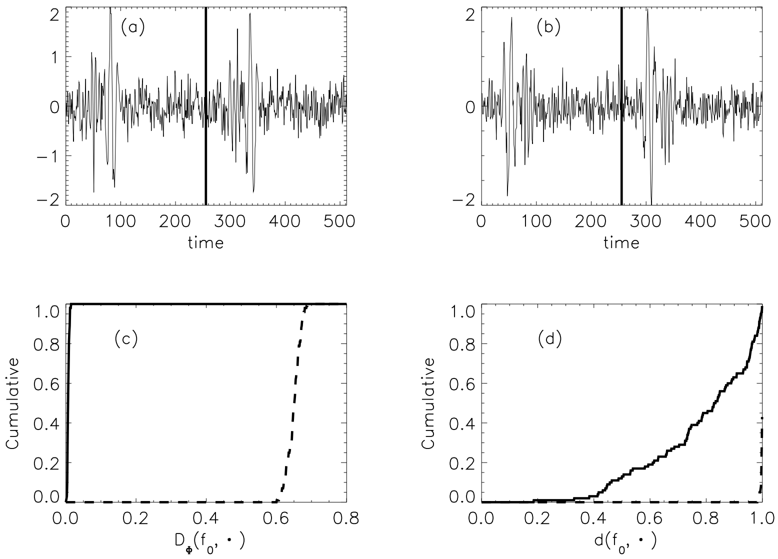

Figure 6: Analogous plots as in Fig. 3, but this time for a signal class which exhibits a temporal and a frequency structure (see text). The noise parameter is $a = 0.2$.

$[0, \pi/2]$, respectively. Thus, the two hypotheses H_0 and H_1 to be tested are

$$\begin{aligned} H_0 &: f(t) = a\epsilon(t) \\ H_1 &: f(t) = h(t)\sin(\omega t + \varphi) + a\epsilon(t) \end{aligned}$$

The noise parameters a has been set to 0.5 in this example.

For $h(t) \equiv 1$ this is a standard problem of detection theory for which a lot of solutions already exist (see e.g. [15]). In our case, for detection we have to maximize $\tilde{m}_\Phi$ as defined in section 4. The algorithm renders $\tilde{m} = 63.4$ with an optimum kernel given by $\tau_0 = \tau_u, \xi_0 = 5.3u$. Once again with this kernel a detection is easily possible by evaluating the distance to an arbitrary member of the class and introducing a threshold (see Fig. 8).

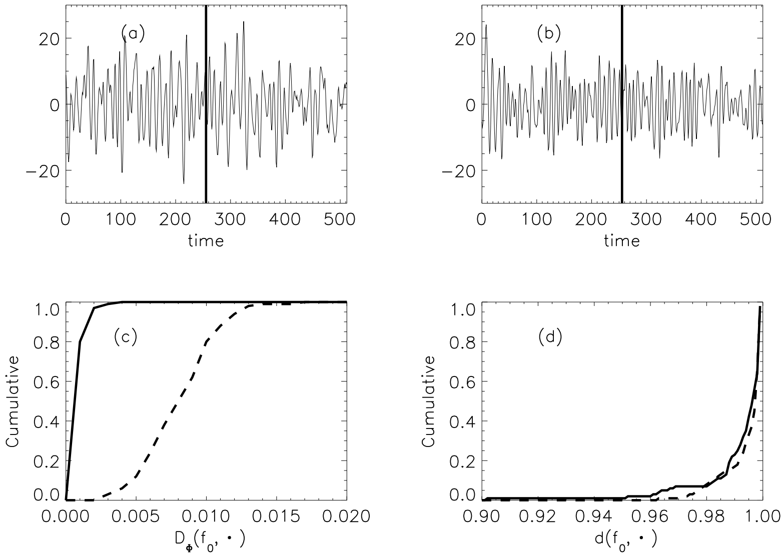

Figure 7: Analogous plots as in Fig. 3, for the two signal classes given by AR(2) processes with different resonance frequencies.

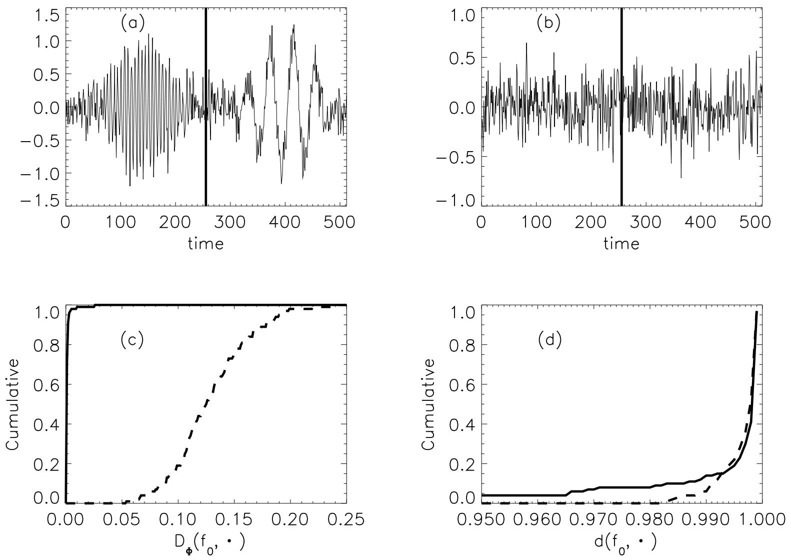

Figure 8: Analogous plots as in Fig. 3, here for the detection of a windowed sinusoid in white noise, the class to be tested against is white noise.

7 Application on real data

The framework of optimum TFRs has been successfully applied on a problem of quality control in industrial production: After the assembly of small eletromotors with attached gear unit an acoustic quality check had to be implemented in order to find out defective motors. From the acoustic noise emitted from regular running a computer should decide whether the motor was faulty or not.

In this section it is shortly described how, for one class of faults, this task could be performed by finding an optimum TFR. For additional information and details we refer to [9].

One of the faults to be detected was "beating". In regular time intervals of about $0.3s$ a short "tick" could be heard in the motor's noise. The origin of these beats can be explained as follows: The rotation of the motor axle is reduced by the gear in a ratio 35:1, i.e. for each rotation of the gear axle the motor axle has to rotate 35 times. Each time the gear axle passes a certain rotation angle, a beat is produced by some fault in the gear unit. Thus, within every 35th rotation of the motor axle, an additional short sound event is produced. Although these beats can be heard rather well, the total power output of the beats is very low. Simple methods like estimating the local power or spectral methods failed in detecting these beats.

However, by optimizing a TFR to this kind of problem, the beating motors could be detected. The idea was the following: Each rotation of the motor axle produces a signal which is given by a time segment of the motor's acoustic noise signal. The single segments should be similar to each other except for beating motors, where each 35th segment (which includes the additional sound event of the gear fault) should differ significantly from the other segments. Again the segments are represented by their TFRs, the distance between them beeing measured by D_Φ. For corrupted segments the distance to the regular segments should be large, where for good motors all segments should have a little distance from each other. As before, the distance is kernel dependent and can thus be optimized in order to render an optimum distinction between the two motor classes.

The following algorithm was implemented: First the signal $f(t)$ was cut into the single periods $p_i, i = 1, \ldots, N$. A number $N \approx 200$ periods were used which corresponds to a signal time of about 2 seconds. For a given kernel the TFR C_i for each segment p_i was calculated. Now, for $k = 0, \ldots, 34$ the value $\tilde{m}_\Phi(k)$ (see Eq. 6) was calculated with class 2 consisting of all segments i with $i \bmod 35 = k$. The maximum of these $\tilde{m}_\Phi$-values was stored as $\hat{m}_\Phi$.

The value of $\hat{m}_\Phi$ is large if there is one set of segments $i = k, k+35, k+70, \ldots$, which differs significantly from the other segments. This should be the case for the beating motors but not for the good ones. Clearly, $\hat{m}_\Phi$ is kernel dependent, and until now nothing has been said about the choice of the kernel. One could imagine that it would be best to use the kernel which maximizes $\hat{m}_\Phi$. Then the

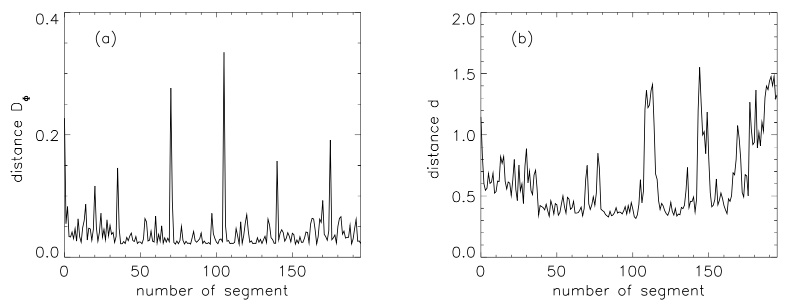

Figure 9: Beating motor: distances of the single segments to the mean of class 1. (a) Distance evaluated with D_Φ for the TFR with maximum $\hat{m}_\Phi$. The corrupted segments which have a distance of 35 segments from one another can be detected easily. (b) Distance evaluated with the the ordinary distance d for the time segments directly, without reference to a TFR.

corrupted segments could be detected optimally thus indicating a beating motor. For the same kernel parametrization is in section 6.2, Fig. 9(a) gives an example of the result of such an optimization which rendered $\hat{m}_\Phi = 7.6$, whereas in Fig. 9(b) an $\hat{m}_\Phi$-value of 1.2 was achieved.

However, this leads to different kernels for different motors. For the detection of the beating motors, in contrast, we want to find a single TFR for the analysis of all signals. In addition, not only $\hat{m}_\Phi$ must be large for the beating motors, but also $\hat{m}_\Phi$ must be small for the good motors. Then, by introducing a level γ, the motors can be classified.

For the optimization of the kernel Φ therefore the empirical cumulative distribution of the $\hat{m}_\Phi$-value for the two classes were calculated. The maximum value of the difference between the two cumulatives is the classification rate (see section 6.2). The optimum kernel was estimated by maximizing this classification rate.

For the found optimum kernel a classification rate of 96.7% was achieved, the corresponding empirical cumulative distribution being shown in Fig. 10(a).

In order to illustrate the advantage of using the optimum TFR instead of the time signal directly, the analogous cumulative for the $\hat{m}$-values evaluated directly on the time segments p_i (instead of C_i) is shown in Fig. 10(b). It can be seen clearly that no class separation is possible in this case.

In this example the optimum kernel has $\tau_0 > \tau_l$ and $\xi_0 > \xi_l$. Thus, none of the simple limiting cases is reached and the full TFR theory is necessary to describe the resulting classification procedure.

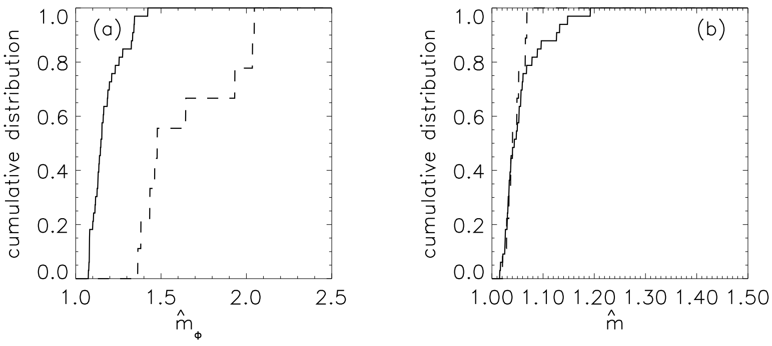

Figure 10: (a) Empirical cumulative distribution of the $\hat{m}_\Phi$-values for the good motors (solid line) and the beating motors (dashed line), where the optimum kernel Φ was used. (b) Analogous as (a), but here $\hat{m}$ being evaluated with the orginal time segments p_i instead of the optimum TFRs C_i.

8 Remarks

In this section two further remarks on the above developed theory shall be made.

1. The concept of optimizing a variable distance measure according to the problem to be solved can be carried over to the ordinary time or frequency representation. By generalizing eq. (2) and introducing a weighting function $g(t)$ one could write

   $$c_g(f_1, f_2) = \int g(t)\, f_1(t) f_2^*(t)\, dt.$$

   Thus, some time regions are weighted stronger than others. In an analogous way, starting with eq. (4), a weighting of some frequencies can be introduced.

   For some applications this may be advantageous. Note however, that in the case of the TFRs all times and all frequencies have the *same* weight. The weighting of the kernel Φ refers to the second order properties of the signal. The results obtained with the optimized kernels in the last section are *not* due to some sort of frequency or time weighting. If, on the other hand, a weighting of frequencies improves the classification, it can be performed *before* the signal enters the time-frequency algorithm, i.e. instead of using the original signal one could use a frequency filtered version of the signal.

   Thus, these time or frequency methods can be implemented additionally if necessary. On the other hand, *after* the transformation into the time-frequency domain the frequently used mixed time-frequency weighting methods (multiplying the TFR with some time-frequency window) can be performed additionally in order to further improve the discrimination.

None of these techniques were applied in the above described examples. Only the second order properties of the signal were used with no weighting of some time, frequency or time-frequency region.

2. The ideas of this article can be generalized for other quadratic signal representations, such as , e.g. , time-scale representations [5, 16]. The use of TFRs implies (see section 2) that the representation properties are independent of time and frequency shifts. That means that if all signals are shifted in time and/or frequency by the same amount the estimated optimum kernel remains optimum.

9 Conclusions

In this article it has been shown that the high degree of flexibility in chosing a TFR of Cohen's class can be used for the improvement of classification and detection algorithms. Starting with the concept of similarity, expressed by means of inner products in the time-frequency domain and a distance measure which is based on this similarity concept, we have introduced a possibility to describe quantitatively the usefullness of a TFR for classification and detection problems. Given two classes to be distinguished one from the other, a discrimination measure m_Φ has been introduced which can be maximized by choosing the appropriate kernel.

It has been demonstrated on some examples that the proposed algorithm can be applied to various types of signal classes, leading to a highly improved representation of the signal in time-frequency space, compared with the ordinary representation in the time or frequency domain.

Of course, each of the above mentioned classification or detection problems could as well be solved by other already known methods. This was mentioned in the first and third example in section 6.2. However, for each considered problem, a different method has to be applied. In addition, also the structure of the signal has to be known. In contrast, with our new algorithm, nothing about the structure of the function classes has to be known *a priori*. Just feed the available realizations into the algorithm and optimize the criterion m_Φ or $\tilde{m}_\Phi$. If any of the above mentioned limiting cases (which can be expressed in time or frequency space without reference to TFR theory) fits for solving the classification or detection problem, the algorithm will find it. Thus, the usually applied methods are included in our algorithm and, additionally, are found automatically.

If, on the other hand, a mixed time-frequency strategy fits better, the optimum kernel will adopt to this strategy.

Thus, the advantage of the proposed algorithm is its generality and the fact that no *a priori* knowledge is needed. Furthermore, a set of realizations of each class is sufficient to estimate the optimum kernel parameters. Thus, the algorithm is best suited for complex data analysis problems where nothing is known about the signal structure.

Simple second order representations like the smoothed periodogram or smoothed $|f(t)|^2$ are included as special cases in the set of all possible TFRs with the proposed parametrization.

Thus, a general algorithm for classification and detection problems has been presented which is based on the second order properties of the signals in their time-frequency representation can be applied to a great variety of signals.

The power of the proposed new algorithm has been furtherly shown by an application on a problem from the area of acoustic quality tests.

10 Acknowledgements

The author likes to thank Prof.Dr. J. Honerkamp for his stimulating interest in this work. Many thanks go to Dr. J. Timmer and Dr. B. Dülfer for fruitful discussions and comments.

A Ambiguity function of two subsequent wave packets

Consider a signal composed of two wave packets with envelope function $h(t)$ and frequency ω and a time distance of t_0:

$$f(t) = h(t)\, e^{i\omega t} \;+\; h(t-t_0)\, e^{i(\omega t+\varphi)} \tag{10}$$

Then the ambiguity function $A_f(\xi,\tau)$ is given by

$$\begin{aligned}
A_f(\xi,\tau) &= \frac{1}{2\pi}\int e^{i\xi\mu}\left[h(\mu+\frac{\tau}{2})\, e^{i\omega(\mu+\tau/2)} \;+\; h(\mu+\frac{\tau}{2}-t_0)\, e^{i(\omega(\mu+\tau/2)+\varphi)}\right] \\
&\qquad \left[h(\mu-\frac{\tau}{2})\, e^{-i\omega(\mu-\tau/2)} \;+\; h(\mu-\frac{\tau}{2}-t_0)\, e^{-i(\omega(\mu-\tau/2)+\varphi)}\right]\, d\mu \\
&= \frac{1}{2\pi}\int e^{i\xi\mu}\, h(\mu+\frac{\tau}{2})\, h(\mu-\frac{\tau}{2})\, e^{i\omega\tau}\, d\mu \\
&+ \frac{1}{2\pi}\int e^{i\xi\mu}\, h(\mu+\frac{\tau}{2})\, h(\mu-\frac{\tau}{2}-t_0)\, e^{i\omega\tau}\, e^{-i\varphi} d\mu \\
&+ \frac{1}{2\pi}\int e^{i\xi\mu}\, h(\mu+\frac{\tau}{2}-t_0)\, h(\mu-\frac{\tau}{2})\, e^{i\omega\tau}\, e^{i\varphi} d\mu \\
&+ \frac{1}{2\pi}\int e^{i\xi\mu}\, h(\mu+\frac{\tau}{2}-t_0)\, h(\mu-\frac{\tau}{2}-t_0)\, e^{i\omega\tau}\, d\mu
\end{aligned}$$

The first and the last term on the right side of this equation, which are called eigenterms, reduce to $e^{i\omega\tau}\, A_h(\xi,\tau)$ and $e^{i\omega\tau}\, e^{i\xi t_0}\, A_h(\xi,\tau)$ by a change of the variables, whereas the other terms, the crossterms, correspond to $e^{i\omega\tau}\, e^{-i\varphi}\, e^{i\xi t_0/2}\, A_h(\xi,\tau+t_0)$ and $e^{i\omega\tau}\, e^{i\varphi}\, e^{-i\xi t_0/2}\, A_h(\xi,\tau-t_0)$. Thus, the eigenterms are located at the origin, whereas the crossterms are located at $(0,\pm t_0)$, a fact that is often mentioned in the literature. The modulus of each term depends only on the ambiguity function of the envelope function $h(t)$.

It can be derived analogously that, for wave packets with an additional frequency difference of ω_0, the crossterms are located at (ω_0, t_0) and $(-\omega_0, -t_0)$.

Furthermore, it is clear that the eigenterms are independent of the phase factors φ in contrast to the crossterms. Thus, calculating the inner products of two different functions f as in eq. (10) which differ in the phase φ we have (if the overlap of eigenterms and crossterms is small) two contributions: one form the inner products of the eigenterms which is independent of the phases, and the other from the inner products of the crossterms which have a phase factor, depending on the difference of the two phases φ.

Until now we have discussed complex wave packets. For real wave packets we use $\cos(\omega t) = \frac{1}{2}(e^{i\omega t} + e^{-i\omega t})$. Thus, we get a decomposition of $f(t)$ into 4 terms, leading to 4 eigenterms and 12 crossterms. In the ambiguity function plane these crossterms are located at $(0,\pm t_0), (\pm 2\omega, 0)$ and $(\pm 2\omega, \pm t_0)$.

B Special cases of the kernel function

The kernel parametrization (see eq. (8)) includes two special cases ($\tau_0 = 0$ and $\xi_0 = 0$) for which the algorithm can easily be interpreted.

For $\tau = 0$ we get

$$
\begin{aligned}
< C_f, C_g > &= \frac{1}{4\pi^2} \int\int |\Phi(\xi,\tau)|^2\, A_f(\xi,\tau) A_g^*(\xi,\tau)\, d\xi\, d\tau \\
&= \frac{1}{4\pi^2} \int\int |\delta(\tau)|^2\, |\Phi(\xi,0)|^2\, A_f(\xi,\tau)\, A_g^*(\xi,\tau)\, d\xi\, d\tau
\end{aligned}
$$

Here we have to integrate $\delta^2(\tau)$-functions. It can be easily shown that

$$\int \delta^2(\tau) f(\tau)\, d\tau = \infty$$

Nevertheless, if we assume a discrete time (which is always the case when calculating with digital computers), replacing the integrals over dt by sums over $t = 1 \ldots N$, we have

$$\sum \delta^2(\tau) f(\tau) = \sum \delta(\tau) f(\tau) = f(0)$$

So, we use the formula

$$\int \delta^2(\tau) f(\tau)\, dt = f(0)$$

keeping in mind that this holds, strictly speeking, only for the discrete time formulation.

Thus, we arrive at

$$\begin{aligned} < C_f, C_g > &= \frac{1}{4\pi^2} \int |\Phi(\xi,0)|^2 \, A_f(\xi,0) \, A_g^*(\xi,0) \, d\xi \\ &= \frac{1}{4\pi^2} < A_f(\xi,0)\Phi(\xi,0), A_g(\xi,0)\Phi(\xi,0) > \end{aligned}$$

Since the inner product is conserved under Fourier transformation $\xi \to t$, it can be evaluated also in the time space. Since $A_f(\xi,0)$ is the Fourier transform of $|f(t)|^2$, the Fourier transform of $A_f(\xi,0)\,\Phi(\xi,0)$ is just a smoothed version of $|f(t)|^2$. Thus, for $\tau_0 = 0$ the inner product of the two TFRs equals the inner product (in the time domain) of smoothed versions of the squared moduli of the signals, the smoothing being stronger if ξ_0 is smaller. In a strictly analogous way it can be shown that the inner product of the TFRs equals the inner product (in the frequency domain) of smoothed versions of $|\tilde{f}(\omega)|^2$ and $|\tilde{g}(\omega)|^2$, if we use a parametrization with $\xi_0 = 0$. Regarding the distance measure D_Φ this implies: setting $\xi_0 = 0$, the distance D_Φ between two functions is small, if the functions have similar spectra $|\tilde{f}(\omega)|^2$. This holds analogously for $\tau_0 = 0$.

References

[1] Richard G. Baraniuk and Douglas L. Jones. A signal-dependent time-frequency representation: Optimal kernel design. *IEEE Trans. Accoustics, Speech and Signal Processing*, 41(4):1589–1602, 1993.

[2] Richard G. Baraniuk and Douglas L. Jones. A signal-dependent time-frequency representation: Fast algorithm for optimal kernel design. *IEEE Trans. Accoustics, Speech and Signal Processing*, 42(1):134–146, 1994.

[3] H. Choi and W. J. Williams. Improved time–frequency representation of multicomponent signals using exponential kernels. *IEEE Trans. Accoustics, Speech and Signal Processing*, 37(6):862–871, 1989.

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

[5] Leon Cohen. The scale representation. *IEEE Trans. Accoustics, Speech and Signal Processing*, 41(7):3275–3292, 1993.

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

[7] Patrick Flandrin. *Temps-Fréquence.* Edition Hèrmes, 1993.

[8] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, Inc., 2 edition, 1990.

[9] C. Heitz. Using optimized time-frequency representations for acoustic quality tests. *in preparation.*

[10] Franz Hlawatsch. Transformation, inversion and conversion of bilinear signal representations. In *Proc. of the IEEE Int. Conf. on Acoust. , Speech and Signal Processing ICASSP-85*, pages 27.5.1–27.5.4, Tampa, FL, 1985. New York.

[11] Douglas L. Jones and Thomas W. Parks. A high resolution data-adaptive time-frequency representation. *IEEE Trans. Accoustics, Speech and Signal Processing*, 38(12):2127–2135, 1990.

[12] Steven Kay and G. Faye Boudreaux-Bartels. On the optimality of the Wigner distribution for detection. In *Proc. of the IEEE Int. Conf. on Acoust. , Speech and Signal Processing ICASSP-85*, pages 27.2.1–27.2.4, Tampa, FL, 1985. New York.

[13] Antonia Papandreou and G. Faye Boudreaux-Bartels. Generalization of the Choi-Williams distribution and the Butterworth distribution for time-frequency analysis. *IEEE Trans. Accoustics, Speech and Signal Processing*, 41(1):463–472, 1993.

[14] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes.* Cambridge University Press, 2 edition, 1992.

[15] M. B. Priestley. *Spectral Analysis and Time Series.* Academic Press, 1981.

[16] Olivier Rioul and Patrick Flandrin. Time-scale energy distributions: A general class extending wavelet transforms. *IEEE Trans. Accoustics, Speech and Signal Processing*, 40(7):1746–1757, 1992.

[17] Yunxin Zhao, Les E. Altes, and Robert J. Marks. The use of cone-shaped kernels for generalized time-frequency representations of nonstationary signals. *IEEE Trans. Accoustics, Speech and Signal Processing*, 38(7):1084–1091, 1990.