DETECTION OF ABRUPT SPECTRAL CHANGES USING SUPPORT VECTOR MACHINES AN APPLICATION TO AUDIO SIGNAL SEGMENTATION

Manuel Davy

IRCCyN - UMR CNRS 6597
1, rue de la Noë – BP92101
44321 Nantes Cedex 3 - FRANCE
Manuel.Davy@irccyn.ec-nantes.fr

Simon Godsill

Cambridge University
Engineering Department
Trumpington street, Cambridge CB1 2PZ - UK
sjg@eng.cam.ac.uk

ABSTRACT

In this paper, we introduce an hybrid time-frequency/support vector machine algorithm for the detection of abrupt spectral changes. A stationarity index is derived from support vector novelty detection theory by using sub-images extracted from the time-frequency plane as feature vectors. Simulations show the efficiency of this new algorithm for audio signal segmentation, compared to another nonparametric detector.

1. INTRODUCTION

Abrupt changes detection is a common but difficult signal processing task. Typical applications are vibration monitoring [1, 2] or music transcription [3]. Strong theoretical results hold in the case of known statistical models of the data [1], however, some real-life systems are difficult to model accurately and model-based statistical techniques are useless in such cases. One needs then resort to nonparametric data analysis techniques, together with nonparametric abrupt change detection algorithms.

The new method presented in this paper addresses nonparametric situations and focuses more specifically on spectral abrupt change detection, with an application to audio processing. Since detecting abrupt spectral change involves both *time* and *frequency*, a natural nonparametric framework is given by Cohen's class Time-Frequency Representations (TFRs) [4]. These representations provide an understandable description of the signal energy content, especially for vibration signals which are composed of spectral components. In this representation space, abrupt changes are characterized by the abrupt beginning/ending of one or several spectral components at a given time, as demonstrated in [3] where an empirical TFR-based abrupt change detector is derived. Another major advantage of Cohen's class is that the TFR itself can be adapted to a specific situation by choosing a convenient TFR kernel as [5] show in the context of signal classification.

TFRs provide an intelligible representation space, yet they do not provide a detection algorithm. The statistical properties of an arbitrary TFR being intractable, nonparametric detection algorithms are necessary. Kernel techniques[1] such as Support Vector Machines (SVMs) are among the most powerful nonparametric tools for classification, regression and novelty detection [6]. Support vector novelty detection consists of deciding whether a vector (e.g. the current observation) is abnormal or new, compared to a set of so-called training vectors (e.g. a set of past observations). In this paper, we present a hybrid TFR/SVM abrupt change detector. The efficiency of these hybrid methods has already been shown for signal classification [7], and we adopt a similar approach here to novelty detection.

This paper is organised as follows: Section 2 describes a generic TFR-based abrupt change detector. Support vector novelty detection is exposed in Section 3, and Section 4 presents results obtained with audio signals. Finally, some conclusions and reseach directions are given.

2. A GENERIC TFR-BASED DETECTOR

The detection algorithm we propose relies on the computation of a stationarity index $I[k]$ characterizing the level of spectral changes in the signal $y[k]$, where k denotes the discrete time. Then, at each time k, the stationarity index $I[k]$ is compared to a decision threshold η, and an abrupt change is detected whenever $I[k] > \eta$.

Assume that a set $\mathcal{X}_k$ of vectors $\mathbf{x}_{k-i}$, $i = 1, \ldots, m$ is available, with each vector $\mathbf{x}_{k-i}$ characterising the spectral contents of the nonstationary signal at time $k - i$. More precisely, the vectors $\mathbf{x}_{k-i}$ are computed using a sliding window from the signal $y[k]$. Assume moreover that we can compute a function $g_{\mathcal{X}_k}(\mathbf{x}_k)$ which measures the dissimilarity of $\mathbf{x}_k$ (characterizing the spectral contents of $y[t]$ at time k) with the vectors in $\mathcal{X}_k$. Then, the index

$$I[k] \triangleq g_{\mathcal{X}_k}(\mathbf{x}_k) \tag{1}$$

measures the nonstationarity of $y[k]$, and can be used to detect abrupt spectral changes. This strategy was initiated in [3], where the feature vectors were extracted from the Cohen's class time-frequency distribution of $y[k]$. However, in [3] the stationarity index was based on time-frequency distance measures, which have been shown to be overperformed by SVMs [7].

In continuous time t and continuous frequency f, Cohen's class TFRs are defined as the convolution of the Wigner-Ville TFR $\mathcal{W}_x$ of the signal $y(t)$ with a TFR kernel ϕ [4]:

$$\mathcal{C}_y^{\phi}(t,f) = \int \mathcal{W}_y(s,\nu)\phi(s-t,\nu-f)dsd\nu \tag{2}$$

The TFR kernel ϕ completely defines the TFR, and its shape can either be chosen from heuristics (e.g., the spectrogram), or result

This work was sponsored by the European research project MOUMIR, http://www.moumir.org

[1] In this paper, we consider both TFR parameterization functions, or TFR kernels, and kernels for decision algorithms. In order to avoid confusion, they will be referred to as *TFR kernels*, and *SVM kernels*.

from an optimization procedure [5]. In the following, we consider discretized TFRs [8], denoted $C_y^{\phi}[k,n]$, where n is the discrete frequency (there are N frequency bins in $[0,0.5[$).

Cohen's class TFRs characterize the spectral energy content of the signal at a given time, and provide therefore natural feature vectors $\mathbf{x}_k[i]$, $i = 1,\ldots,m$. Namely, once the kernel ϕ is selected, the feature vector is a subimage of width $T \geq 1$:

$$\mathbf{x}_k[i+N(j-1)] \;=\; C_y^{\phi}[k-T+j, 0.5(i-1)/N]$$

where $(i = 1,\ldots,N,\ j = 1,\ldots,T)$ and $\mathbf{x}_k[l]$ denotes the l-th component of the vector $\mathbf{x}_k$. The corresponding detection algorithm depends on the parameters η, T and ϕ which can be easily tuned as shown in Section 4. The decision function $g_{\mathcal{X}_k}(\mathbf{x}_k)$ is now derived from SVM novelty detection theory.

3. SUPPORT VECTOR NOVELTY DETECTION

Support vector novelty detection is a specific class of SVM algorithms [6], which have been succefully applied to, e.g., jet engine pass-off tests [2]. The objective of novelty detection is to decide whether a given vector is similar to a set of training vectors (assumed to be "normal" or "non-novel"), or not.

Suppose a set of training vectors $\mathcal{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_m\}$ is available in a so-called input space, denoted Ω. Each $\mathbf{x}_i$ is supposed to be a "normal" vector. Given a vector $\mathbf{x}$, we want to derive a fuction $f_{\mathcal{X}}(\mathbf{x})$, such that

$$\begin{array}{lcll} f_{\mathcal{X}}(\mathbf{x}) & > & 0 & \text{if } \mathbf{x} \text{ is similar to } \mathcal{X} \text{ (i.e., } \mathbf{x} \text{ is normal)} \\ f_{\mathcal{X}}(\mathbf{x}) & < & 0 & \text{if } \mathbf{x} \text{ is not similar to } \mathcal{X} \text{ (i.e., } \mathbf{x} \text{ is abnormal)} \end{array}$$

This definition of $f_{\mathcal{X}}(\mathbf{x})$ requires "similar" and "not similar" to be precisely defined; this is the aim of support vector machines. Imagine $\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}$ are samples distributed according to some underlying probability distribution P. What we really want to know is whether $\mathbf{x}$ is distributed according to P ($\mathbf{x}$ is normal), or not ($\mathbf{x}$ is abnormal). A simple solution consists of determining a region R of the space Ω that captures most of P probability mass, namely, R is such that $\int_R P(d\mathbf{x}) \approx 1$. The detection function is then such that $f_{\mathcal{X}}(\mathbf{x}) > 0$ if $\mathbf{x} \in R$, and $f_{\mathcal{X}}(\mathbf{x}) < 0$ otherwise. A solution to estimate the region R (or its bounding hypersurface) consists of fitting a SVM kernel $k(\mathbf{x}_i,\mathbf{x})$ on the support training vectors $\mathbf{x}_i$, with a certain weight α_i as depicted in Fig. 1. The weighted addition of the SVM kernels $\sum_{i=1}^m \alpha_i k(\mathbf{x}_i,\mathbf{x})$ defines a hypersurface (see Fig. 1, middle), and the region R is such that:

$$\mathbf{x} \in R \;\text{ if and only if }\; \sum_{i=1}^{m} \alpha_i k(\mathbf{x}_i,\mathbf{x}) - b > 0 \qquad (3)$$

Eq. (3) defines the contour of R by cutting the hypersurface at a given altitude b, Fig. 1, bottom. As Eq. (3) depends on several parameters, the problem to be solved consists of determining optimally the SVM kernel k, the weights α_i and the threshold b.

Many possible SVM kernels have been proposed in the literature, and a good choice is the Gaussian kernel (or RBF kernel) [2, 6], defined by:

$$k(\mathbf{x},\mathbf{y}) \;=\; \exp -\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2} \qquad (4)$$

where $||\cdot||^2$ is the L_2 norm defined on Ω. This SVM kernel depends on one parameter σ, related to the kernel spread. Once the SVM kernel is selected, one needs to determine the parameters α_i, $i = 1,\ldots,m$ and b such that the region R actually captures most of P probability mass. However, additional constraints ensuring that the bounding hypersurface of R is simple have to be considered, in order to limit over-training. By mapping the vectors into a dot product space, support vector machines provide a simple solution through an optimisation problem.

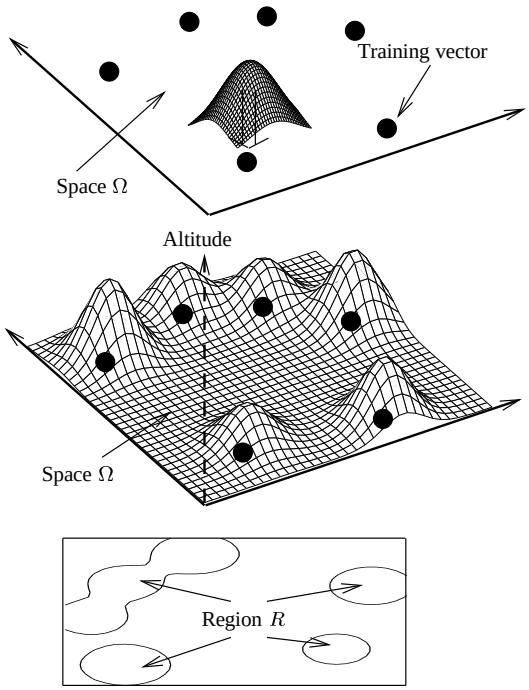

Fig. 1. By fitting weigthed SVM kernel functions on training vectors, represented by black points (top figure), one defines a surface (middle figure). Cutting this surface at a given altitude b defines the contour of the region R (bottom figure).

3.1. Optimisation in a dot product space

SVM kernels such as in Eq. (4) have the following property: there exists a so-called feature space $\mathcal{E}$, endowed with a dot product denoted $<\cdot,\cdot>$ and a mapping $\Phi : \Omega \longmapsto \mathcal{E}$ such that $k(\mathbf{x},\mathbf{y}) = <\Phi(\mathbf{x}),\Phi(\mathbf{y})>$ (this is because RBF kernels satisfy Mercer's conditions [6]). In the following, we denote $x_i = \Phi(\mathbf{x}_i)$, $i = 1,\ldots,m$ and $x = \Phi(\mathbf{x})$. In the feature space, the training vectors are distributed by an underlying distribution P' (related to P by the mapping Φ), and the problem is again to determine a region R' of $\mathcal{E}$ that captures most of this distribution probability mass.

Since $||x_i||^2 = k(\mathbf{x}_i,\mathbf{x}_i) = 1$, for all $i = 1,\ldots,m$, the training vectors x_i in $\mathcal{E}$ are located on an hypersphere centered at the origin of $\mathcal{E}$ with radius 1. In $\mathcal{E}$, the training data are located on a portion of the hypersphere as depicted in Fig. 2, and can be separated from the rest of the space $\mathcal{E}$ by a hyperplane $p(w,b)$ parametered by the vector w and the distance to the origin $b/||w||$. The parameters w and b need now be computed such that the region R' is the part of the sphere that is bounded by $p(w,b)$, and this is done by maximizing the distance $b/||w||$ from the origin, under the constraint that all the training vectors are located on the

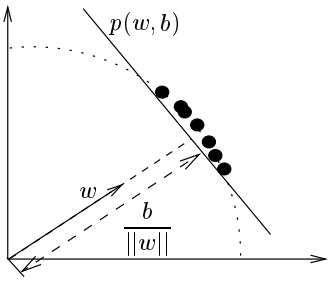

Fig. 2. In the feature space $\mathcal{E}$, the vectors are located on a hypersphere. The hyperplane $p(w,b)$ parametered by w and b separates the training vectors from the rest of the surface of the hypersphere.

opposite side of $p(w,b)$. This corresponds to the following constrained quadratic optimisation problem, where the scale parameter b is given:

$$\text{Minimise } \frac{1}{2}||w||^2 \text{ subject to } < w, x_i > \geq b, \ i = 1, \ldots, m \quad (5)$$

Theoretical results [6] state that the optimal hyperplane $\widehat{p}(w,b)$ exists and is unique. The training vectors located on both the hypersphere and the hyperplane are called *support vectors*. Note that in the feature space, the hypersurface bounding R' is simple, which limits over-training.

3.2. Soft margin novelty detection

In many situations, the training set may contain a small number of abnormal vectors, and the optimal hyperplane should be positioned such that these vectors are located between the origin and $\widehat{p}(w,b)$. In this case, all the training vectors may satisfy

$$< w, x_i > \geq \ b - \xi_i \ \text{ with } \ \xi_i \geq 0 \ \text{ for all } i = 1, \ldots, m \quad (6)$$

where $\{\xi_i\}$ are so-called slack variables, which allow for some abnormal vectors. However, in order to control the number of such vectors, the objective function of Eq. (5) becomes

$$\left[\frac{1}{2}||w||^2\right] \ + \ \left[\frac{1}{\nu m}\sum_{i=1}^{m}\xi_i \ - \ b\right] \quad (7)$$

where ν is a parameter than tunes the number of possible abnormal training vectors, whose properties are summarised in the following theorem [6]:

Theorem:

i. ν is an upper bound for the fraction of outliers;

ii. ν is a lower bound for the fraction of support vectors;

iii. Suppose the training vectors were sampled independently from a distribution P which does not contain discrete components. Suppose, moreover, that the SVM kernel is analytic and non-constant. Then, with probability 1, asymptotically, ν equals both the fraction of support vectors and the fraction of outliers

Standard calculations [2] show that the optimization problem of Eq.'s (6) and (7) is equivalent to minimizing

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j < x_i, x_j > \quad (8)$$

subject to the constraints $\sum_{i=1}^{m}\alpha_i = 1$ and $0 \ \leq \ \alpha_i \ \leq \ 1/\nu m$ for all $i = 1, \ldots, m$. Once the dual variables are computed (using, e.g., the Loqo algorithm in Vanderbei [9]), the decision function becomes:

$$f_{\mathcal{X}}(\mathbf{x}) \ = \ \text{sign}\left[\sum_{i=1}^{m}\alpha_i < x_i, x > -b\right]$$

which gives the connection with the input space, since $< x_i, x >= k(\mathbf{x}_i, \mathbf{x})$. Finally, one has the decision function of Eq. (3) with $b = \sum_{i=1}^{m}\alpha_i k(\mathbf{x}_i, \mathbf{x}_j)$ where j is chosen in $1, \ldots, m$.

3.3. Interpretation – Comments

As can be seen, the weights α_i can be optimally computed by "linearizing" through the use of SVM kernels. Moreover, the mapping Φ *does not have to be calculated explicitely*, since only k appears in the key equations. The simplicity of the bounding hypersurface in Ω is a consequence of the simplicity of the decision hypersurface in $\mathcal{E}$. Note finally that b can be seen as a scale factor, and we propose the decision function $g_{\mathcal{X}_k}(\mathbf{x}_k)$ in Eq. (1) to be:

$$g_{\mathcal{X}_k}(\mathbf{x}_k) \ = \ -\log\left[\frac{\sum_{i=1}^{m}\alpha_i k(\mathbf{x}_{k-i}, \mathbf{x}_k)}{b}\right]$$

The SVM optimisation procedure is repeated for each time sample. In our present implementation, the optimisation is run at each time instant, but a faster strategy consists of simply updating the α_i, $i = 1, \ldots, m$ at each time, since one vector is removed from the start of $\mathcal{X}_k$ and one is added to the end.

4. APPLICATION TO AUDIO SIGNAL SEGMENTATION

The abrupt spectral change detector presented in this paper is well adapted to audio signals. More precisely, musical records as well as human voice are composed of one or several fundamental frequencies and corresponding sets of harmonics, which are interpreted as spectral components in the time-frequency plane. Fig. 3 represent the audio signal of a classical guitar playing alone, and Fig. 4 corresponds to a jazz band (bass guitar, electric guitar and drums). In addition to the new detector, the TF distance measure-based detector presented in [3] is also implemented for comparison.

In these simulations, we selected the spectrogram $C_y^{\text{SP}}[k,n]$ with Hamming window (6.8 ms), which is a correct compromise between time and frequency resolutions. In order to enhance high frequencies, the detector is applied to $(C_y^{\text{SP}}[k,n])^{0.05}$. The tuning parameters have been chosen to rough values: the window width is $T = 0.23$ms, and the SVM parameters are $m = 50$, $\sigma = 0.05$, and $\nu = 0.2$. Fig. 3 displays the index computed from the spectrogram of the guitar signal, as well as the index computed with the algorithm of [3] (in this latter method, the subimage width is $T = 1.18$ms so that both methods use the same amount of TFR information). The SVM-based index shows sharper peaks located at the exact change times (determined by an experienced listener), whereas the index of [3] misses most of the changes. Moreover, all the time changes are found, and located at accurate times. The computational load is about 30 times higher with the SVM technique than for the TFR-based technique, but note that we have not implemented the most efficient SVM strategy. Moreover, the SVM optimisation procedure is implemented in Matlab, whereas the TF distance computation used in [3] is implemented in C.

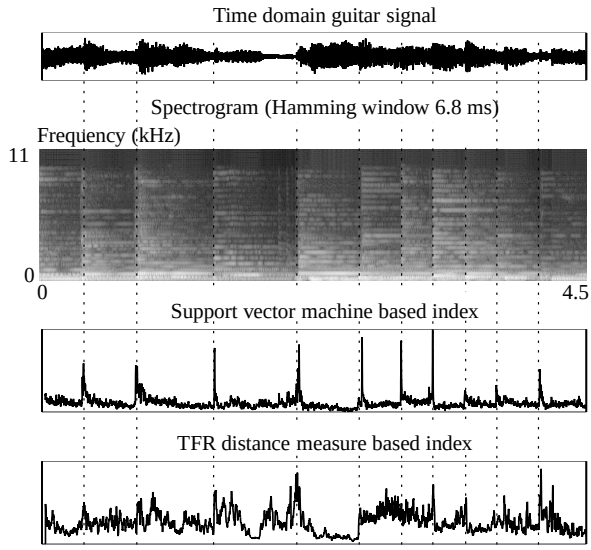

Fig. 3. Detection results for the SVM-based new detector, and the TF distance measure-based detector of [3] using the spectrogram, for a musical signal (guitar alone). The vertical dotted lines indicate the abrupt change times found by an experimented listener.

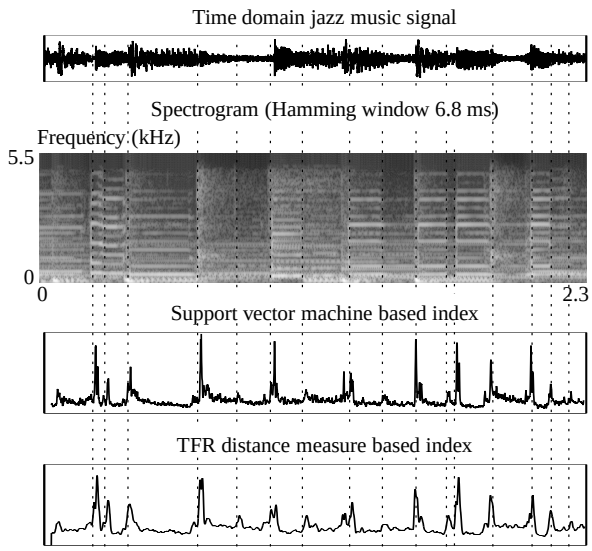

Fig. 4. Detection results for the SVM-based new detector, and the TF distance measure-based detector of [3] using the spectrogram, for a musical signal (jazz with bass guitar, electric guitar and drums). The vertical dotted lines indicate the abrupt changes times found by an experienced listener.

Fig. 4 gives the results obtained with the more complex jazz signal. Again, the abrupt changes have been accurately detected. This case was more complex since abrupt changes are caused by three different instruments. Difficult cases arise when, e.g., the bass guitar plays one note without changes, whereas the electric guitar notes change several times. In such situations, an abrupt change does not involve all the spectral components and it is more difficult to detect. This kind of changes are generally correctly detected by our algorithm. The TF distance measure-based detector also performs statisfactorily, but the time accuracy is lower compared to the SVM-based detector, since it requires wider subimages. A narrower subimage would improve the detector time accuracy to the detriment of its smoothness.

5. CONCLUSION

The new abrupt change detector presented in this paper is derived from time-frequency representations together with support vector machines. This combination of tools is, again, proven efficient for nonparametric decision problems involving nonstationary signals.

Future research directions will consider optimisation of the detector parameters by using a database of expertized (i.e., pre segmented) audio signals, as well as sequential optimization of the support vector parameters.

6. REFERENCES

[1] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes - Theory and Application*, Prentice-Hall, April 1993.

[2] P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis, "Support vector novelty detection applied to jet engine vibration spectra," in *NIPS'2000*, 2000.

[3] H. Laurent and C. Doncarli, "Stationarity index for abrupt changes detection in the time-frequency plane," *IEEE Signal Processing Letters*, vol. 5, no. 2, pp. 43 – 45, February 1998.

[4] P. Flandrin, *Time-Frequency/Time-Scale Analysis*, Academic Press, 1999.

[5] M. Davy, C. Doncarli, and G. Faye Boudreaux-Bartels, "Improved optimization of time-frequency based signal classifiers," *IEEE Signal processing letters*, vol. 8, no. 2, pp. 52–57, February 2001.

[6] A. Smola and B. Schölkopf, *Learning with Kernels*, MIT press, To appear.

[7] A. Gretton, M. Davy, A. Doucet, and P.J.W. Rayner, "Nonstationary signal classification using support vector machines," in *IEEE workshop on Statistical Signal Processing (SSP)*, Singapor, August 2001.

[8] M. Davy and E. Roy, "The ANSI C Time-Frequency Toolbox for use with Matlab," 2000, http://www-sigproc.eng.cam.ac.uk/ md283/.

[9] R. J. Vanderbei, "Loqo: An interior point code for quadratic programming," Tech. Rep. TR SOR-94-15, Department of Civil Engineering and Operations Research, Princeton University, 1995.