Tracing Sound Objects in Audio Textures

Monika Dörfler
University of Vienna
Faculty of Mathematics
NuHAG
Email: monika.doerfler@univie.ac.at

Ewa Matusiak
University of Vienna
Faculty of Mathematics
NuHAG
Email: ewa.matusiak@univie.ac.at

***Abstract*—This contribution presents first results on two proposed methods to trace *sound objects* within *texture sounds*. We first discuss what we mean by these two notions and explain how the properties of a sound that is known to be textural are exploited in order to detect changes which suggest the presence of a distinct sound event. We introduce two approaches, one is based on Gabor multipliers mapping consecutive time-segments of the signal to each other, the other one on dictionary learning. We present the results of simulations based on real data.**

I. Introduction

Sound signals play a central role in human life and the manner sound is perceived is highly sophisticated, complex and context-dependent. In some applications, one may be interested in distinguishing between what may be called a ”sound object” and more textural sound components constituting an acoustical background. The notion of sound object (”objet sonore”) was introduced by Pierre Schaeffer [10] as a generalization of the concept of a musical note, in particular their definition implies a time-limitation of sound objects.

Human listeners tend to perceive sound in a structured manner, with the ability to focus and de-focus. Whether a particular event is experienced as a relevant sound structure as opposed to background, textural sound, seems to depend both on cultural and educational background, cp. [5], that may be shared by a group of listeners. From a certain point of view, the perception of sound components as background (textural) sound or object (compactly structured) sound, depends on the ”zoom” the listener wishes to adopt or unconsciously assumes. In this contribution, we attempt to mimic these observations in a technical way, by ”defining” a sound to be textural if it does not change certain characteristics which are first to be determined from a certain amount of data. In that sense, we need the a priori knowledge that a particular part of a signal represents textural sound segments. Any signal components representing a significant change are then considered to be new objects in the sense of not belonging to the previous texture sound or background. By definition, a characterizing feature of texture sounds, in particular as opposed to the signal components we would like to call sound objects, is some kind of stationarity over an extended period of time; while micro-changes are always present, the listener integrates them as part of the texture, at least after some time has passed. Therefore, any two sufficiently long slices of a pure texture sound can, and should, be assumed to be correlated. This observation leads us to the following approach: given a signal which is known to present a texture sound, we learn its inherent characteristics. Using the information gained from the learning step, we can then look for significantly different, hence salient, signal components, which we then define to represent a sound object.

For both the learning and the observation period, we divide the signal into overlapping time-slices. Then, during observation, we look for substantial changes from one part of the signal to another, which would indicate the presence of a sound object. We are going to quantify, what we mean by substantial changes, by means of two technical tools: sparsity in an appropriate dictionary and similarity of Gabor transforms. Based on these two tools, we introduce two methods to scan texture signals for the presence of what may be conceived as sound objects. While the proposed framework may also be useful for the task of detecting audio events, this application is not the primary motivation for our study. The latter, challenging task addressed in the framework of CASA [1] requires a much wider and more elaborate evaluation stage and is beyond the scope of the current contribution. Here, we are primarily motivated by a different challenge, which parallels the cognitive process sketched above: we mimic a situation in which a user/listener makes real-time decisions about the property of an event occurring in the signal to be or not to be a sound object which deserves attention. We divide the signal into overlapping slices. In the first approach we propose, we make use of Gabor transforms; more precisely, the variations of the Gabor coefficients between different slices of the signal are tracked by investigating corresponding Gabor multipliers.

The second proposed method is by means of the exploitation of sparsity constraints via dictionary learning. Given a part of a texture sound which is known to be free of sound objects, we learn a dictionary such that each slice admits a sparse approximate representation in that dictionary. We then scan the signal piece by piece by checking its reconstruction error with respect to the corresponding dictionary, in order to detect in which intervals of time an object may occur.

The two methods and the involved tools are presented in the next section. Then, some promising results of preliminary simulations are presented in Section III and we conclude with a short discussion and perspectives.

[1] Computational Auditory Scene Analysis, cp. [13].

II. Technical tools for sound object tracing

The proposed methods aim at deciding about the presence of distinct sound objects within a signal whose first section is known to be textural. Both methods give a decision about the presence or absence of an object within a given slice of the signal. This can be seen as a first step in exact object localization in terms of precise onset and offset times, and later extraction. We will see in Section III that the proposed methods are designed particularly for longer signals and should be applicable to online-applications.

Before describing the two methods in detail we recall some definitions from Gabor analysis and fix notation. We will be working with square integrable functions $L^2(\mathbb{R})$, with norm $\|\cdot\|_2$ induced by an inner product $\langle f,g\rangle = \int_{\mathbb{R}} f(t)\overline{g(t)}\,dt$, $f,g \in L^2(R)$. For $f \in L^2(\mathbb{R})$ and $\omega,\tau \in \mathbb{R}$, the operators $M_\omega f(t) = e^{2\pi i\omega t}f(t)$ and $T_\tau f(t) = f(t-\tau)$ are called frequency and time shift operators, respectively. A collection $\mathcal{G}(g,a,b) = \{g_{k,l} := M_{bl}T_{ak}g\}_{k,l\in\mathbb{Z}}$ is called a Gabor frame for $L^2(\mathbb{R})$ if the operator $S_{g,g}$

$$S_{g,g}f = \sum_{k,l\in\mathbb{Z}} \langle f, g_{k,l}\rangle g_{k,l} \qquad \text{for all} \quad f \in L^2(\mathbb{R}) \tag{1}$$

is bounded and invertible on $L^2(\mathbb{R})$.[2].

For every frame $\mathcal{G}(g,a,b)$ there exists a function γ, called dual window, such that $\mathcal{G}(\gamma,a,b)$ is again a frame, called dual Gabor frame, and $f = S_{g,\gamma}f = S_{\gamma,g}f$ for all $f \in L^2(\mathbb{R})$.

Let $f \in L^2(\mathbb{R})$ be a background, texture signal. We divide it into overlapping slices f_i, $i \in \mathbb{Z}$, in the following way: $f_i(t) = f(t)$ for $t \in [\alpha_i, \beta_i]$ with $\alpha_{i-1} < \alpha_i \le \beta_{i-1}$ and $\alpha_{i+1} \le \beta_i < \beta_{i+1}$.

A. Gabor Multipliers

We first describe the method based on Gabor multipliers. This method not only allows to detect a change but also gives more information on the time-frequency location of a potential object. Let $\mathcal{G}(g,a,b)$ be a Gabor frame and $\mathcal{G}(\gamma,a,b)$ its dual frame. Let $\mathbf{m} = \{m_{k,l}\}_{k,l\in\mathbb{Z}}$ be a bounded complex-valued sequence. Then the Gabor multiplier associated to (g,γ,a,b) with symbol, or mask, $\mathbf{m}$ is given by

$$G_{\mathbf{m}}f = \sum_{k,l\in\mathbb{Z}} m_{k,l}\langle f, g_{k,l}\rangle \gamma_{k,l}\,. \tag{2}$$

The operator $G_{\mathbf{m}}$ is well defined and bounded on $L^2(\mathbb{R})$ [4].

In [8] the authors addressed the problem of transforming one signal into another by means of linear operators. They focus on Gabor multipliers as the transforming operators. More precisely, for two signals f_1 and f_2, given dual frames $\mathcal{G}(g,a,b)$ and $\mathcal{G}(\gamma,a,b)$, the objective is to find a symbol $\mathbf{m}$ such that the Gabor multiplier $G_{\mathbf{m}}$ takes f_1 into f_2 subject to certain constraints on the mask $\mathbf{m}$. The constraints on the mask can be sparsity in time-frequency plane or total energy. An optimal mask, subject to given constraints is a solution to the following minimization problem

$$\min_{\mathbf{m}} \|f_1 - G_{\mathbf{m}}f_2\|_2^2 \quad \text{subject to} \quad d(\mathbf{m}) < \epsilon\,, \tag{3}$$

where d we can chosen to be, for example $d(\mathbf{m}) = \lambda\||\mathbf{m}| - 1\|_1$, to promote sparsity, or $d(\mathbf{m}) = \lambda\|\mathbf{m} - 1\|_2^2$ to control total energy, where λ is a sparsity prior tuning the influence of the second term in (3).

For texture sounds, the slices f_i, as defined in the previous section, are similar, hence also their Gabor transforms. The grade of similarity is learned from the first part of the signal, which is known to be textural. Then, a symbol $\mathbf{m}_i$ of a Gabor multiplier transforming f_i to f_{i+1}, $f_{i+1} = G_{\mathbf{m}_i}f_i$, is close to one, or in other words $d(\mathbf{m}_i)$ is close to zero. During the learning phase, the parameter λ should be tuned to yield small deviations from the constant mask $\mathbf{m} = 1$.

Now, the problem of detecting a sound object versus a stationary background is based on studying masks $\mathbf{m}_i$. If $\mathbf{m}_i$ is significantly different from 1, or $d(\mathbf{m}_i) > \epsilon$ for some chosen $\epsilon > 0$, then the slices f_i and f_{i+1} differ significantly which leads us to assuming the presence of an object in slice f_{i+1}.

a

B. Dictionary Learning with Sparsity Prior

Given a dictionary $\mathbf{D} \in \mathbb{C}^{K\times L}$, $K < L$ and a signal $f \in \mathbb{C}^K$, we say that f admits an $S-$sparse approximation over $\mathbf{D}$ if one can find an approximation of f by S atoms from $\mathbf{D}$. In other words, we are looking for coefficients $x \in \mathbb{C}^L$, such that

$$f \approx \mathbf{D}x \quad \text{while} \quad \|x\|_0 \le S. \tag{4}$$

Here, $\|\cdot\|_0$ is a pseudo-norm counting the non-zero entries in x. Finding the best solution to (4) is an NP-hard problem; however by relaxing the counting pseudo-norm to an ℓ_1 norm, it becomes a convex optimization problem that can be tackled with many existing efficient algorithms, such as basis pursuit (BP) [2], orthogonal matching pursuit (OMP) [12] or FOCUSS [9]. A dictionary yielding sparse approximate representation for a class of signals can be learned from a sufficient number of data samples. Let F be a set of N signals $f_i \in \mathbb{C}^K$, collected into a matrix of size $K \times N$, for which one would like to find a dictionary such that each signal in the group admits an $S-$sparse approximate representation. A dictionary with the desired properties can be built by finding a solution to the following minimization problem

$$\min_{\mathbf{X},\mathbf{D}} \sum_{i=0}^{N-1} \|f_i - \mathbf{D}x_i\|_2^2 \quad \text{subject to, for every } i \quad \|x_i\|_0 \le S\,, \tag{5}$$

where $\mathbf{X} \in \mathbb{C}^{L\times N}$ is the matrix of coefficients $x_i \in \mathbb{C}^L$. Among many algorithms addressing the problem of dictionary learning are K-SVD [1], maximum likelihood methods [7] or the MOD method [3].

For a given texture sound f, we observe the first couple of seconds of the signal and learn a dictionary which gives a sparse approximate representation thereof. We build the

[2] Note that the coefficients $\langle f, g_{k,l}\rangle$ in $S_{g,g}$ are samples of a short-time Fourier transform of f at sampling points (ak, bl).

training data set F by considering slices of first L samples of f, each of length K with $M \geq 0$ samples of overlap, i.e. $f_i(k) = f(i(K-M)+k)$ where $k = 0, \ldots, K-1$ and $i = 0, \ldots, N-1$. Then, assuming ongoing textural characteristics of f, the slices f_i for $i \geq N$ also admit sparse approximate representation in the same dictionary while no significant changes occur. In detail, let $\epsilon > 0$ be given. If it is possible to find a vector x_i of coefficients such that $\|f_i - Dx_i\|_2 \leq \epsilon$ while $\|x_i\|_0 = S$ is satisfied, then we conclude no presence of a sound object. However, if the above relation is violated, we can assume additional components in f_i that are not correlated with elements of $\mathbf{D}$. We scan the signal f slice by slice and verify its representation in $\mathbf{D}$.

III. Simulations

We present numerical results based on two classes of texture sounds f: (heavy) rain and washing machine noise. In order to give a proof of concept, we apply the suggested methods to finding synthetic signals s which unambiguously qualify as sound objects within the background signals; we use damped sums of six different harmonics of 0.5 seconds length. The SNR[3] of the objects present in the texture sound is between $-5dB$ and $-7.5dB$. Note that the sound-files corresponding to the examples as well as supplementary examples, codes and extensions are available at the website homepage.univie.ac.at/monika.doerfler/SoundObj.html.

A. Gabor Multiplier

For the Gabor multiplier approach, we choose slices of approximately half a second (20480 samples) length with 75% overlap. We use a standard tight Gabor frame with a Hann window of length 1024 and 75% overlap. The spectrogram of the test signal is depicted in the upper plot of Fig. 1. The three harmonic and compactly supported synthetic signals are clearly visible. The lower plot shows the deviation $\||\mathbf{m}|-1\|_1$ for the mask corresponding to the transition between two time-slices. Based on the first, purely textural part of the signal, λ is tuned in order to allow only negligible deviation of the absolute value of $\mathbf{m}$ from 1. During our experiments, it turned out that the success depends heavily on an appropriate choice of λ, which was chosen to be 1.2 in the first example.

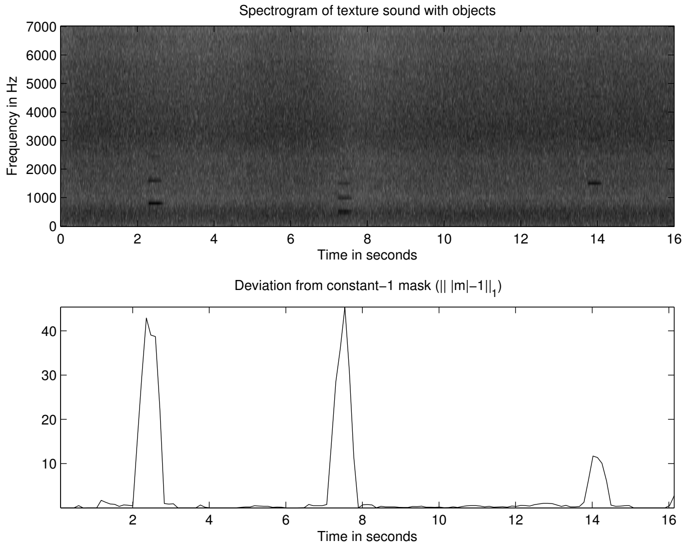

Fig. 1. Detection of sound objects in background noise (Rain) using Gabor multipliers approach.

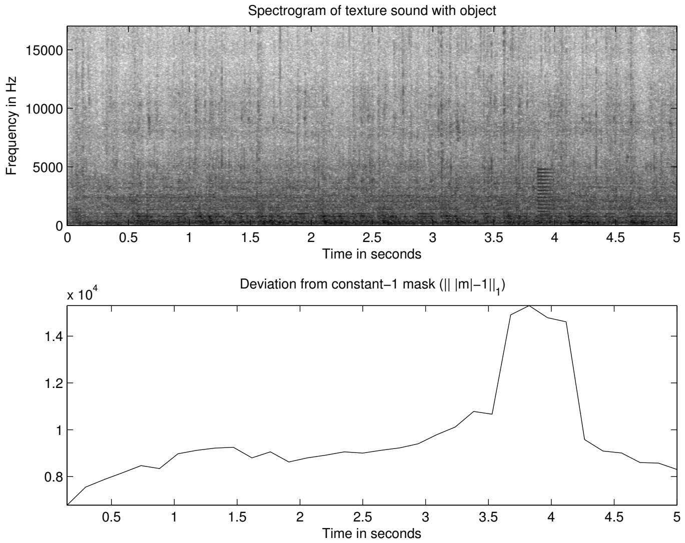

Fig. 2. Detection of a sound object in background noise (Washing machine) using Gabor multipliers approach.

The second example, the distinct noise produced by a washing machine, is a more complex texture sound. Here, the situation is more difficult, since the "stationarity" of the texture is present on a larger scale, as visible in its spectrogram, shown in Figure 2, upper display. For this example, we had to allow for a much smaller $\lambda = 0.01$, i.e. for significant deviations from a constant mask, in oder to obtain meaningful results. Therefore, as opposed to the previous example, we obtain much higher values of the deviation $\||\mathbf{m}|-1\|_1$ also for the textural part. In Figure 3, we show two masks occurring in the investigation of this example; it is clearly visible, that this particular signal contains a lot of energy in low frequency bands, with a certain periodicity (also audible in the signal). It is quite obvious that, without taking these changes of energy into account, no meaningful transition can be expected. On the other hand, inspection of the part of the mask that is related to the sound object has a clear local persistence in time which the texture part lacks, but which is typical for harmonic signals. It is planned to exploit this kind of a priori knowledge - or assumption - about the objects one is interested in, in order to improve the method's success and reliability. In particular, the models introduced as structured or social sparsity, cp. [6], [11], show promising results in first experiments and will be further exploited.

[3] We define the signal to noise ratio (SNR) by $SNR_{dB} = 10\log_{10}(\|s\|_2^2/\|f\|_2^2)$, given in dB, by where f is the background signal, which can be seen as "noise" in which s, the sound object is to be traced.

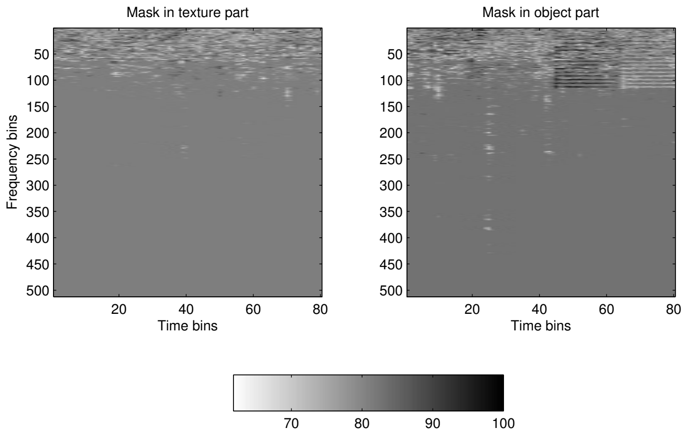

Fig. 3. Change of consecutive masks: between texture slices and between object slices; values are in dB.

B. Sparse Dictionary Representation

We applied the second, dictionary-based method to the signals presented in the previous section; we chose time-slices of 256 samples length and 50% overlap. It turned out during the experiments, that, in the evaluation step, smaller overlap is possible and does not deteriorate the results, since the time-resolution given by the slice-length of about 6ms is fine enough. The resulting evaluation criteria, namely approximation error for a maximum number of atoms and level of sparsity for a chosen error tolerance, are shown in Figure 4. Obviously, both criteria show significant deviation from the texture level during the duration of the sound objects. It should be noted that the amplitude of the time-signals don't visibly increase during the sound objects, also cf. homepage.univie.ac.at/monika.doerfler/SoundObj.html to listen to the audio files.

IV. Discussion and Perspectives

We presented two methods for sound object tracing in background, texture signals. Both methods exploit the assumed quasi-stationary character of texture signals and decide that a 'foreign' sound object should be present, if that stationarity is lost. The suggested methods and numerical experiments need to be extended to a much larger samples of both texture sounds and sound objects in order to draw reliable conclusions about the situations in which the proposed models give satisfactory results; furthermore, there are several open questions as to how long the slices of the signal should be, in both the sparsity method and Gabor multipliers, and what kind of Gabor frames to choose for the latter approach. These questions will be investigated in detail in ongoing work on the topic and results will be presented on the companion website.

Acknowledgments

This research was supported by the WWTF project AudioMiner (MA09-024). The authors would like to thank Anaïk Olivero for sharing with us a source code for computing Gabor masks and Richard Kronland-Martinet and his team in LMA, CNRS Marseille, for giving us permission to use their software SPAD to synthesize texture sounds.

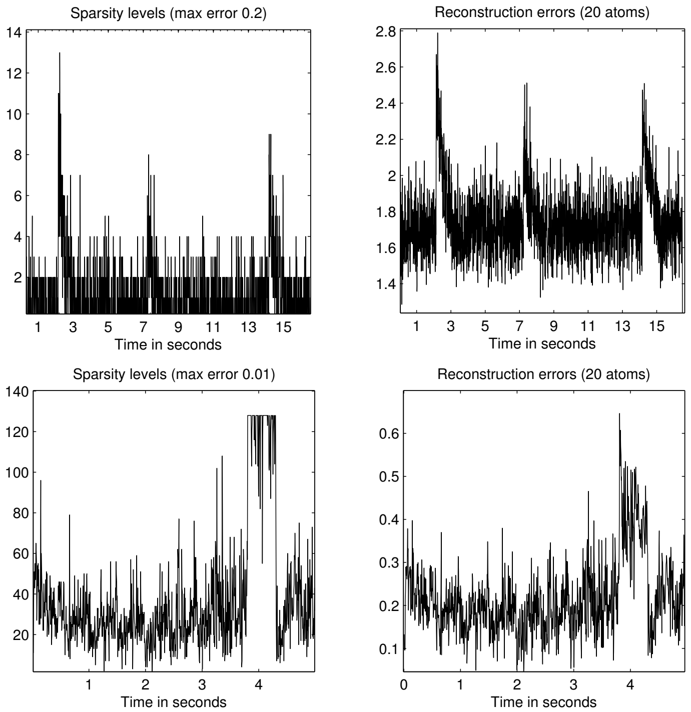

Fig. 4. Tracing sound objects by sparse dictionary representation. Top row: rain signal; bottom row: washing machine signal.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54(11):4311–4322, 2006.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[3] K. Engan, S. O. Aase, and J. H. Hakon Husoy. Method of optimal directions for frame design. *IEEE Int. Conf. Acoust.,Speech, Signal Process.*, 5:2443–2446, 1999.

[4] H. G. Feichtinger and K. Nowak. A first survey of Gabor multipliers. In H. G. Feichtinger and T. Strohmer, editors, *Advances in Gabor Analysis*, Appl. Numer. Harmon. Anal., pages 99–128. Birkhäuser, 2003.

[5] V. Klien, T. Grill, and A. Flexer. On Automated Annotation of Acousmatic Music. *Journal of New Music Research*, 41(2):153–173, 2012.

[6] M. Kowalski, K. Siedenburg, and M. Dörfler. Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators. *IEEE Trans. Signal Process.*, 61(10):2498 – 2511, 2013.

[7] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Comp.*, 12:337365, 2000.

[8] A. Olivero. *Les multiplicateurs temps-fréquence. Applications à l'analyse et à la synthèse de signaux sonores et musicaux.* PhD thesis, 2012.

[9] B. Rao and K. Kreutz Delgado. An affine scaling methodology for best basis selection. *IEEE Trans. Signal Process.*, 47:187–200, 1999.

[10] P. Schaeffer. *On Automated Annotation of Acousmatic Music.* Editions du Seuil, Paris, France, 2002.

[11] K. Siedenburg and M. Dörfler. Persistent Time-Frequency Shrinkage for Audio Denoising. *J. Audio Eng. Soc.*, 61(1/2), 2013.

[12] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[13] D. Wang and G. Brown, editors. *Computational auditory scene analysis: Principles, Algorithms, and Applications.* IEEE Press/Wiley-Interscience, 2006.