Sparse Regression in Time-Frequency Representations of Complex Audio

Monika Dörfler, Gino Velasco
Nuhag, Faculty of Mathematics
University of Vienna, Austria
monika.doerfler@univie.ac.at
gino.velasco@univie.ac.at

Arthur Flexer, Volkmar Klien
Austrian Research Institute for Artificial Intelligence
arthur.flexer@ofai.at
volkmar.klien@ofai.at

ABSTRACT

Time-frequency representations are commonly used tools for the representation of audio and in particular music signals. From a theoretical point of view, these representations are linked to Gabor frames. Frame theory yields a convenient reconstruction method making post-processing unnecessary. Furthermore, using dual or tight frames in the reconstruction, we may resynthesize localized components from so-called sparse representation coefficients. Sparsity of coefficients is directly reinforced by the application of a ℓ^1-penalization term on the coefficients. We introduce an iterative algorithm leading to sparse coefficients and demonstrate the effect of using these coefficients in several examples. In particular, we are interested in the ability of a sparsity promoting approach to the task of separating components with overlapping analysis coefficients in the time-frequency domain. We also apply our approach to the problem of auditory scene description, i.e. source identification in a complex audio mixture.

1. INTRODUCTION

Time-frequency representations such as the spectrogram or short-time Fourier transform seem to be well suited for the representation of music. However, due to the uncertainty principle, a certain smearing of the time-frequency coefficients is unavoidable. This effect will often create overlap between components that would not be expected to overlap by their nature, e.g. two sinusoids with close frequencies. For other components, the overlapping area may be increased, thus complicating the task of separating certain components with approximately disjoint support in the time-frequency domain. For example, one might be interested in suppressing a certain instrument's contribution from a music signal. Such approaches are used in Computational Auditory Scene Analysis by the name of Time-Frequency masks. In this contribution, we describe the nature of time-frequency representations from a mathematical point of view. We then introduce a model and a corresponding algorithm for actively obtaining a sparse signal representation. The model rests on the fact that the time-frequency representations typically used in the audio signal processing community, are highly redundant. Hence, the representation coefficients are not unique and we may impose additional assumptions on the coefficients. Here, we will describe the effect of imposing an ℓ^1-penalization on the coefficients. We will show for several synthetic and real signals, that this leads to sharper representations and better separation properties. We apply the presented methods to the problem of auditory scene description. More precisely, given a mixture of known source sounds, we wish to determine the activity pattern for each source. This will be achieved by correlating representation coefficients of the sources with those of the mixture. In this setting, we compare the canonical time-frequency coefficients to those obtained from sparse regression in the time-frequency domain. The idea to use the sparse coefficients in place of the the canonical ones rests on the assumption, that these coefficients accurately capture the main characteristics of each of the sources. We will show that results obtained from sparse time-frequency representations improve those obtained from canonical representations. In particular, the amount of false positives is reduced, which is an important issue as pointed out in [1]. Thus, sparsity constraints help to avoid artificial correlations between signal components.

2. GABOR FRAMES: ANALYSIS AND SYNTHESIS

Given a discrete sequence of real or complex numbers, $x[n]$, $n \in \mathbb{Z}$, as well as a, usually compactly supported, window function $\varphi[n]$, $n \in \mathbb{Z}$, the short-time Fourier transform (or STFT) of $x[n]$ is given, for $k \in \mathbb{Z}$ and $\omega \in [-0.5, 0.5]$ by

$$\mathcal{V}_\varphi x(k,\omega) = \sum_{n=-\infty}^{\infty} x[n]\varphi[n-k]e^{-2\pi i\omega n}. \qquad (1)$$

Now, in practice, a subsampled version of (1) will usually be applied. Also, since the window φ has finite length l, we deal with a finite number of frequency bins. Hence, the result of the sampled STFT, also called Gabor transform, [2], is a matrix of size $N \times M$, where N is the number of time shifts by a time-constant, or hop-size, a considered. M is the number of frequency bins, hence the length of the FFT, given by l/b, b being the frequency-shift constant.
To gain a more general point of view, it is convenient, to consider the coefficients $\mathcal{V}_\varphi x(ka, mb)$ obtained from subsampling in (1), as inner products between the signal x and

time-frequency-shifted windows. We define the inner product x and $y \in \mathbb{C}^L$ as

$$\langle x, y \rangle = \sum_{n=0}^{L-1} x[n]\overline{y[n]}. \tag{2}$$

φ will from now on be called window function.

Definition 1 (Time-frequency shifts) *Let $x \in \mathbb{C}^L$.*
$T_k x[n] := x[n-k]$ is called translation operator or time shift by k.
$M_l x[n] := e^{\frac{2\pi i l n}{L}} x[n], l \in \mathbb{Z}$ is called modulation operator or frequency shift by l.
The composition of these operators, $M_l T_k$, is a time-frequency shift operator.

The family

$$\varphi_{k,m} := M_{mb} T_{ka} \varphi \tag{3}$$

for a window function $\varphi \in \mathbb{C}^L$, $m = 0, \ldots, M-1$ and $k = 0, \ldots, K-1$, where $Ka = Mb = L$, is called the set of *Gabor analysis functions*.
We next describe, under which conditions a signal is unambiguously defined by a family of Gabor atoms. The theory of *frames* gives the appropriate framework and we first state the defining inequalities for signals of finite energy.

Definition 2 *A set of Gabor analysis functions $\varphi_{k,m}$ in $\mathbb{C}^L$ is called a Gabor frame, if there exist constants $A, B > 0$, so that, for all $f \in \mathbb{C}^L$*

$$A\|f\|^2 \leq \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} |\langle f, \varphi_{k,m} \rangle|^2 \leq B\|f\|^2. \tag{4}$$

If $A = B$, then the functions $\varphi_{k,m}$ form a *tight frame*.
The above inequality can be understood as an "approximate Plancherel formula", characterizing the preservation of energy by the transform and leading to the invertibility of the frame operator S:

$$Sf = \sum_{k,m \in \mathbb{Z} \times \mathbb{Z}} \langle f, \varphi_{k,m} \rangle \varphi_{k,m} \tag{5}$$

Note that the frame operator is usually defined as an operator on $L^2(\mathbb{R}^d)$ and the relation to the finite discrete case is actually of interest in itself, see [3, 4] for more details. Since we only consider the finite disrete case, which is of interest for implementation, we may think of S as a matrix mapping $\mathbb{C}^L$ to $\mathbb{C}^L$.
The invertibility of S is equivalent to the existence of frame bounds $0 < A, B < \infty$ in the frame inequality in (4). The invertibility of S, now, leads to the existence of so-called dual frames, yielding convenient reconstruction formulas. This can easily be seen as follows: The canonical *dual frame* $\tilde{\varphi}_{k,m}$, is given by

$$\tilde{\varphi}_{k,m} = S^{-1} \varphi_{k,m}. \tag{6}$$

For Gabor frames, the elements of the dual frame $S^{-1}\varphi_{k,m}$ are generated from a single function (the dual window $\tilde{\varphi}$), and will hence be denoted by $(\tilde{\varphi}_{k,m})$. This follows from the fact that S and S^{-1} (the frame operator and its inverse) commute with the modulation and translation operators M_{mb} and T_{ka}, for $m = 1, \ldots, M$ and $k = 1, \ldots, K$, see e.g. [5]. Hence,

$$f = S^{-1}Sf = \sum \langle f, \varphi_{k,m} \rangle \tilde{\varphi}_{k,m}. \tag{7}$$

The coefficients used in (7) are called *canonical* in order to distinguish them from (infinitely many) other possible expansion coefficients with respect to the same frame. In the case of a tight frame, $S = AI$, where I denotes the identity operator, and therefore $S^{-1} = \frac{1}{A} I$. Tight frames will be further discussed in the next subsection.
In the finite discrete case of $f \in \mathbb{C}^L$ a collection $\{\varphi_{k,m}\} \subset \mathbb{C}^L$ with $N = KM$ can only be a frame, if $L \leq N$ and if the matrix G, defined as the $N \times L$ matrix having $\overline{\varphi_{k,m}}$ as its $(n + kM) - th$ row, has full rank. In this case, the frame bounds are the maximum and minimum eigenvalues of the frame operator $S = G^* \cdot G$. Here, G^* denotes the adjoint of G. The eigenvalues of this positive matrix yield information about numerical stability. The closer the frame-bounds are, the closer the frame operator will be be to a diagonal matrix. If A and B differ too much, the inversion of the frame operator is numerically unstable.
In applications in audio signal processing, redundancy of 2, 4 or even higher is common. Further, the effective length of the window φ equals or is shorter [1] than the FFT-length. In this special situation, the frame operator takes a surprisingly simple form:
From the definition of the frame operator

$$Sf = \sum_{k,m} \langle f, \varphi_{k,m} \rangle \varphi_{k,m}$$

a straight-forward calculation (see [3] for details) shows that the single entries of S are given by

$$S_{ji} = \begin{cases} M \sum_{k=0}^{K-1} T_{ka}\varphi[j]\overline{T_{ka}\varphi[i]} & \text{if } |j-i| \bmod M = 0 \\ 0 & \text{else} \end{cases} \tag{8}$$

Since $M \geq l$, where l is the window-length, $j = i$ is the only case for which $|j - i| \bmod M = 0$ holds and $\varphi[j]$ and $\varphi[i]$ are both non-zero. Therefore, the frame operator is diagonal and the dual window $\tilde{\varphi}$ is calculated as

$$\tilde{\varphi}[n] = \varphi[n] / (M \sum_{k=0}^{K-1} T_{ka} |\varphi[n]|^2)$$

2.1 Tight frames: synthesising with the analysis window

As mentioned above, for a *tight frame*, the frame operator equals identity up to a constant factor. This is as close as we may get to an orthonormal basis. As a matter of fact, for any given Gabor frame, a corresponding tight frame can be found and, as for dual frames, by a surprisingly simple formula in many situations of practical relevance.
Note that the frame operator S is a positive and symmetric and therefore selfadjoint operator, from which it follows

[1] E.g. in the case of zero padding.

that S^{-1} and $S^{-\frac{1}{2}}$ are selfadjoint as well and commute with time-frequency shifts.
These properties allow the following manipulations of expansion (7):

$$\sum_{k,m}\langle f,\varphi_{k,m}\rangle\tilde{\varphi}_{k,m} = S^{-1}Sf = S^{-\frac{1}{2}}SS^{-\frac{1}{2}}f$$
$$= \sum_{k,m}\langle f, S^{-\frac{1}{2}}\varphi_{k,m}\rangle S^{-\frac{1}{2}}\varphi_{k,m} = \sum_{k,m}\langle f,\varphi^t_{k,m}\rangle\varphi^t_{k,m}$$

Remark 1 *Note that the tight window given by $\varphi^t = S^{-\frac{1}{2}}\cdot\varphi$ is closest to the original window in the following sense: Let φ be a window generating a frame for lattice constants a and b and let φ^t be the tight window given as $\varphi^t = S^{-\frac{1}{2}}\varphi$. Then for any function h generating a tight frame for lattice constants a and b, the following holds [6]:*

$$\|\varphi - \varphi^t\|_2 \leq \|\varphi - h\|_2$$

This result shows that the tight window calculated as $\varphi^t = S^{-\frac{1}{2}}\varphi$ combines the advantage of using the same window for analysis and synthesis with optimal similarity to a given analysis window. At the same time no "correction" by multiplication with a gain function is necessary after processing, which makes processing more efficient and the results less ambiguous in the case of modification of the synthesis coefficients. This property becomes even more relevant, if the canonical Gabor coefficients are modified in some sense before resynthesis, e.g. in the case of time-frequency masking. In this case, the choice of a tight frame for analysis and synthesis minimizes the error arising from sampling in the coefficient domain. In the case of sparse coefficients, which we consider in the next section, tight frames also allow for a reliable interpretation of the obtained coefficients as well as satisfying reconstruction from these coefficients.

In analogy to the dual window and under the same conditions, we may deduce that the tight window φ^t corresponding to a given window φ and the time constant a can be calculated as:

$$\varphi^t[n] = (S^{-\frac{1}{2}}\varphi)[n] = \varphi[n]/\big(M\sqrt{\sum_{k=0}^{K-1} T_{ka}|\varphi[n]|^2}\big)$$

3. ENFORCING SPARSITY BY AN ℓ^1 CONSTRAINT

Being convinced that the signal components of interest have a sparse, at least approximative, representation in the atomic systems we use, we may directly look for relevant coefficients only. The prior information can be introduced by assuming an appropriate distribution of the coefficients. Mathematically, minimization of an ℓ^1-constraint on the coefficients yields explicit solutions and fast algorithms. [2]
In fact, the ℓ^1-constraint corresponds to a prior on the coefficients $\mathbf{c}$ of the form

$$p(\mathbf{c}) = \exp(-\mu\|\mathbf{c}\|_1).$$

[2] Note, that it has been proved that certain situations ℓ^1-minimization in fact yields the optimally sparse solution, see [7].

This prior leads to the following minimization problem. Given a tight Gabor frame with elements $\varphi^t_{k,m}$, we wish to minimize the following expression:

$$\Delta(x) = \|\sum_{k,m} c_{k,m}\varphi^t_{k,m} - \hat{x}\|_2^2 + \mu\|\mathbf{c}\|_{\ell^1} \qquad (9)$$

where $\|\mathbf{c}\|_{\ell^1} = \sum_{k,m}|c_{k,m}|$ is the ℓ^1-norm of the coefficient sequence and $\hat{x} = x + n$ is the observed signal, possibly contaminated by noise n. For orthonormal bases (instead of frames), the problem formulation in (9) leads to a well-known soft-thresholding solution. However, in the over-complete situation of frames, the situation is more intricate and an iterative procedure has to be applied.

3.1 Landweber iterations

While the classical basis pursuit [8] may be solved by linear programming algorithms, iterative thresholding is commonly used to solve the minimization problem posed in (9). Other algorithms exist, see [9, Chapter 12] for a thorough overview, however, the Landweber algorithm proposed in the current situation appeals by its simplicity and easy implementation. In the statistics community the iterative soft thresholding has been known for some time under the name "the lasso", [10]. The choice of μ is usually a delicate task and mirrors assumptions on the noise variance present in the signal model. In fact, increasing μ corresponds to the assumption of a higher noise variance and will thus lead to a sparser solution.
Let G_{φ^t} denote the Gabor analysis matrix corresponding to the tight system at hand and let $G^*_{\varphi^t}$ denote the corresponding synthesis system which is just the adjoint of G_{φ^t} in the case of a tight frame. To find the solution of (9), consider the sequence of iterates

$$\mathbf{c}^n = \mathbb{S}_\mu(\mathbf{c}^{n-1} + G_{\varphi^t}(\hat{x} - G^*_{\varphi^t}\mathbf{c}^{n-1})), \qquad (10)$$

where $\mathbf{c}^n$ are the expansion coefficients obtained in the n^{th} step, $\mathbf{c}^0$ is arbitrary and the thresholding operator $\mathbb{S}_\mu$ is given by

$$\mathbb{S}_\mu(c_{k,m}) = \begin{cases} c_{k,m} + \frac{\mu}{2}, & c_{k,m} \leq -\frac{\mu}{2} \\ 0 & |c_{k,m}| < \frac{\mu}{2} \\ c_{k,m} - \frac{\mu}{2}, & c_{k,m} \geq \frac{\mu}{2} \end{cases} \qquad (11)$$

It should be noted that $G_{\varphi^t}\hat{x}$ are the coefficients of the original data, which have to be calculated just once. The iterations thus consist of a gradient step (10), in which the coefficients are updated, and the soft thresholding step given in (11). Usually a stopping criterion is built into the algorithm using a fixed tolerance for $\|\mathbf{c}^n - \mathbf{c}^{n-1}\|$, unless a maximum number of iterations is reached before the stopping criterion is met.
According to [11], the corresponding iterative algorithm converges to the solution of (9).

3.2 Two examples

We first consider a synthetic signal comprised of two sinusoids with frequencies 1300Hz and 1400Hz, given a sampling rate of 8192. We use a Gaussian window of 400

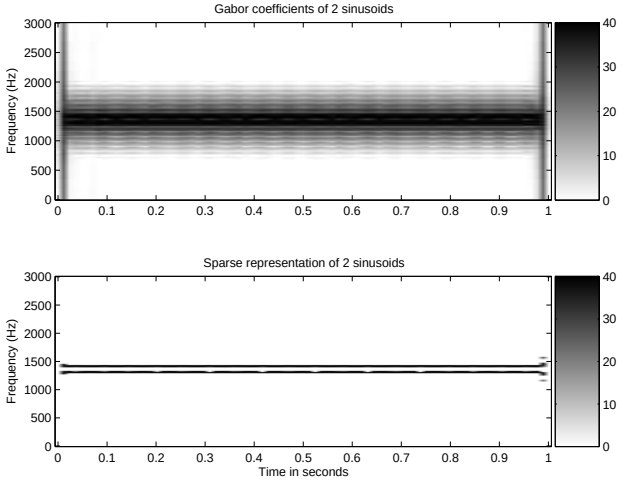

Figure 1. Gabor coefficients and sparse representation of two close sinusoids

samples length, the time-shift parameter $a = 100$ and calculate the canonical Gabor coefficients, shown in the first display of Figure 1. The second display, then, shows the coefficients resulting from ℓ^1-penalization on the expansion coefficients according to (9). It is immediately obvious, that the algorithm visually separates the two signal components. Note that approximate reconstruction from these coefficients is possible, as the correct phase factors are generated by the algorithm. A small error occurs, depending on the choice of μ in (9).

Our second example is a short extract from a music signal consisting of a piano, a double-bass and drums, see Figure 2. Again, we calculate the sparse coefficients, once with a wide Gaussian window, to represent the tonal parts, and once with a narrow Gaussian window to obtain sparser coefficients for transient parts. The results are shown in the 2nd and last display. Reconstructing from sparse coefficients, obtained with the wide window, yields a rather satisfying reproduction of the tonal part (bass, piano), while the residual coefficients mainly contain the cymbals' contribution. Note that, to this point, we have not applied any sophisticated statistical model to suppress noise or separate signal components, nor have we used any more sophisticated sparsity constrained as suggested in [12] to better encode dependencies between the single coefficients. We only use the fact that a relevant part of the signal has a sparse representation in the frame used for analysis. This example underlines the possible merits of the approach to the task of separating components in the time-frequency domain. Similar results using other overcomplete dictionaries have been obtained before, see e.g. [13], where sparse representations of signals were considered using the modulated complex lapped transform (MCLT). Important issues concerning the efficient encoding of the positions and values of the significant coefficients that arise in obtaining sparse coefficients such as those mentioned in e.g. [13, 14] will not be discussed in this paper. Let us just remark, however, that the number of significant coefficients in our experiments amounts to 0.3% to 3% of the size of

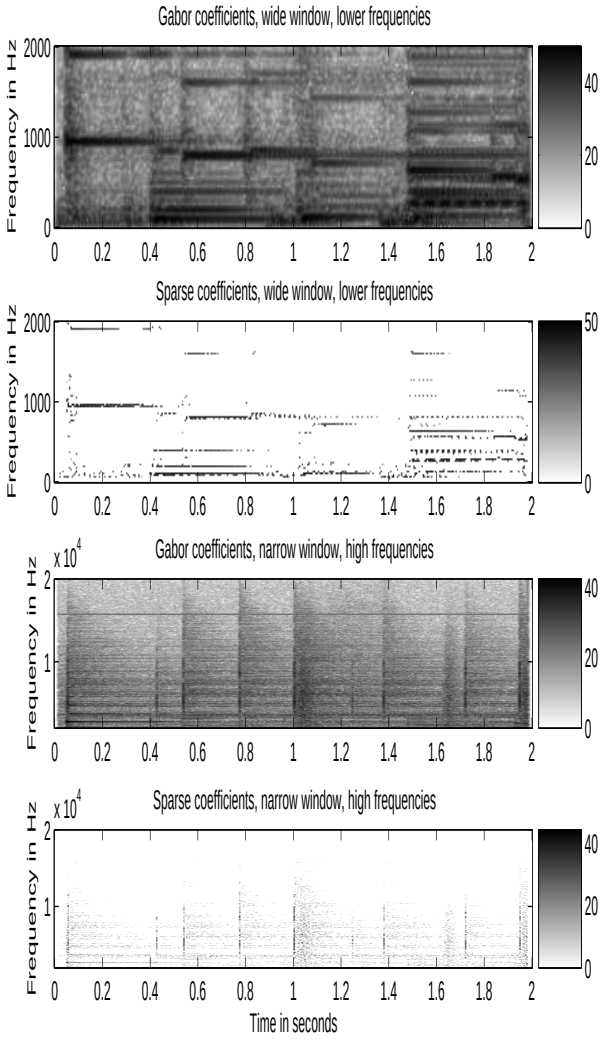

Figure 2. Sparse coefficients for a music signal

the coefficients space.

4. APPLICATION: AUDITORY SCENE DESCRIPTION

Our next application is in the area of auditory scene description, i.e. the classification of audio sources in sound mixtures of several sources. Please note that we are not aiming at source separation but at the easier task of source identification. On the other hand we are going beyond recognition of instrumental sources [15] by including a wider variety of sounds [1]. Also, although closely related to the work done in the very active research domain of music transcription, see, e.g. [16, 17], our approach addresses a slightly different task. Since our primary interest is in electro-acoustic music, for which often well defined sound grains are either pre-defined or can be extracted from a given composition, we wish to automatically determine, whether a particular source sound is present at a specific time in the piece. This is an important step in the process of automatic or semi-automatic annotation of electro-acoustic

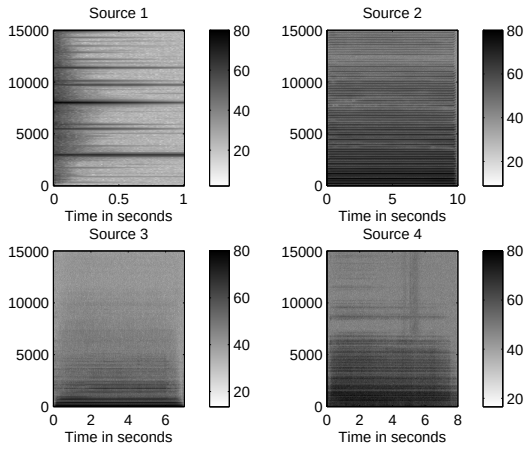

Figure 3. Canonical coefficients of four sources

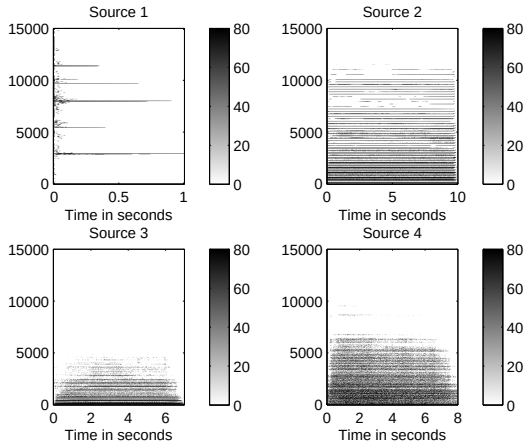

Figure 4. Sparse coefficients of four sources

music.

4.1 Sound Sources

The sources in our example are:

- source 1: single high note played on a Glockenspiel (one second long)
- source 2: long note played on an accordion of medium pitch (ten seconds long)
- source 3: long clarinet tone, low pitch (seven seconds long)
- source 4: noisy clarinet sound, without pitch, air only (eight seconds long)

The canonical and sparse coefficients of the sources are shown in Figure 3 and Figure 4, respectively.

While sources 2 and 3 are harmonic sounds produced by instruments, source 1 is an inharmonic sound produced by a bell and sound 4 is a noisy sound with little harmonic constituents. These sources were chosen to reflect the wide variety in spectral and temporal characteristics displayed by sounds commonly used in electro-acoustic composition.

4.2 Method

We consider the following setting: we are given a sound file generated by N known sources s^j, $j = 1, \ldots, N$, which can be active for any given time t. The signal then is a sum of shifted copies of the sources s^j at time t_k, i.e. $s^j(t-t_k)$. For the purpose of our experiment however, we approximate f at uniform overlapping time intervals I_l as a linear combination of the sources:

$$f_{I_l} \approx \sum_{j=1}^{N} a_{j,I_l} s^j, \tag{12}$$

where a_{j,I_l} is a function storing the activation pattern, i.e. a binary function with values in $\{0, 1\}$. We wish to recover a_{j,I_l} for $j = 1, \ldots, N$, over the intervals I_l.
To do so, we observe the following. Since we expect approximate orthogonality of the various sources in the transform domain, we may assume that the correlation between the coefficients of the mixture and each of the sources reflects the presence of the sources. We therefore proceed as follows: time-frequency coefficients of both the source specimen and the mixture are being computed; overlapping time slices of the time-frequency coefficients are correlated with time-slices of the same length from all four source specimen.

In the sequel we are going to use the absolute values of both the canonical Gabor coefficients and the sparse coefficients $c_{k,m}$ obtained as solution of (9). For brevity, we set $\hat{s}_{k,m} = |\mathcal{V}_\varphi s(k,m)| = |\langle s, \varphi_{k,m}\rangle|$ and $\hat{c}_{k,m} = |c_{k,m}|$. In order to judge the approximate orthogonality of the source specimen in the coefficient domain, we define the inner product of coefficient matrices $\hat{s}^1, \hat{s}^2$ as

$$\langle \hat{s}^1, \hat{s}^2 \rangle_M = \sum_k \sum_m \hat{s}^1_{k,m} \cdot \hat{s}^2_{k,m}$$

and consider the following matrices:

$$CM_{i,j} = \langle \hat{s}^i, \hat{s}^j \rangle_M, \text{ and } CM^{spars}_{i,j} = \langle \hat{c}^i, \hat{c}^j \rangle_M.$$

We normalize the coefficients corresponding to the various sources, so that we can say that deviation from orthogonality between the sources is reflected in deviation from diagonality of the matrices CM and CM^{spars}, respectively. On the other hand, if the condition number of the obtained matrices is good, the correlation between sources can be corrected by applying the inverse of the respective matrix to the obtained correlations between mixture and sources. For clarity, we describe the de-correlation step for time-frequency coefficients $\hat{f}$ of any signal f without specifying whether the coefficients are canonical or sparse for the moment. For the mixture signal f we observe:

$$\langle \hat{f}_{I_l}, \hat{s}^k \rangle \approx \sum_{j=1}^{N} a_{j,I_l} \langle \hat{s}^j, \hat{s}^k \rangle,$$

so that in order to retrieve the coefficients a_{j,I_l}, we have to solve a system of equations involving the matrices CM or CM^{spars}. The detailed procedure is described in Section 4.3 below.

- Note that it is vital in our method to consider time-frequency coefficients rather than just single spectral representations. This takes the time-structure of the signals into account. This procedure makes it unnecessary to examine the correlation coefficients in every time-instant.
- Since the time-structure of the sources is essential to the performance of our method, transient signal components, in particular clicks, are not well-described. However, methods for extraction of transient signal components exist, see e.g. [18, 19, 20, 21], and therefore, their classification may be considered separately.
- For the simulations discussed below, we chose time-slices of one second length and an overlap of 0.5 seconds. As we will see, for sources with significant time-structure, this approach yields rather satisfying results.

4.3 Experiment and results

In the experiments, we consider a one minute signal mix consisting of the four sources mentioned above, at most three of which are active at any time. We calculate the Gabor coefficients of the whole length of the mix (one minute) but just one second for each of the sources using the following parameters: a Hanning window of length 1024 samples (corresponding to $23ms$ at a sampling rate of $44100Hz$) with a hop size of 512, from which a tight window is obtained. This yields a Gabor coefficient matrix of size 1024×5169 for the signal mix and 1024×88 for each of the sources. We then consider time-slices $\hat{f}_{I_l}$ of the Gabor coefficient matrix of the same size as the coefficient matrix of the sources, hence corresponding to a duration of 1 second. We consider an overlap factor of 2 between subsequent time-slices, resulting in 116 time positions in our setting. We then compute the correlation-matrix C, of size 6×116, whose entries are given by

$$C_{j,l} = \langle \hat{f}_{I_l}, \hat{s}^j \rangle_M .$$

Since the model for each time-slice is approximated to be a linear combination of the sources, solving for the coefficients a_{j,I_l} amounts to computing for $\mathrm{inv}(CM) * C$. For the four sources in our experiments, we obtain condition numbers 2.9936 and 1.6425 for CM and CM^{spars}, respectively, which reflects, in this case, their deviation from the identity. This can be interpreted by saying that the sparse coefficients of the sources have less overlap than the canonical coefficients, as expected.
The resulting activation matrix is then normalized for each source, such that the maximum value assumed is 1 for each source. Hence the same threshold is simultaneously applied to all sources, entries above the threshold are set to 1 while the rest are set to 0. The same procedure is applied to the sparse representation of the signal and the sources. The sparse coefficients have been obtained by applying Landweber iterations with $\mu = 0.04$.
Figure 5 shows the true map of time positions where each source is present (target map, black indicating presence of a source), as well as the matrices obtained from the above mentioned procedure, using the Gabor coefficients and the sparse representation with threshold values 0.145 and 0.2 (with the lower threshold being optimal for the sparse representation and the higher for Gabor coefficients).

In comparing the resulting matrices with the true maps, we analyze the receiver operating characteristic (ROC) curve and compute the following:

- accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- specificity $= \frac{TN}{TN+FP}$
- sensitivity $= \frac{TP}{TP+FN}$,

where TP, TN, FP, FN signify true positives, true negatives, false positives, and false negatives, respectively. We see in Figure 6 the corresponding graphs plotted over varying threshold values.

An ideal method for identifying the sources in the mixture would have both specificity and sensitivity equal to one. In more realistic settings it is necessary to find a threshold where a good compromise between high specificity and sensitivity exits. As can be seen from the rightmost plot in Figure 6, both the canonical and sparse representations yield almost equal results in terms of sensitivity. With increasing threshold, less and less sources are detected correctly. But as can be seen from the middle plot in Figure 6, the optimum value for specificity is reached earlier for the sparse representation. This means that one can choose a threshold where specificity is optimal (i.e. no false positives) while still having very high sensitivity (i.e. high amount of true positives). This also results in the optimum value in terms of accuracy being reached earlier for sparse representations (leftmost plot in Figure 6). These slightly improved results are due to the lower cross-correlation between signal components in the sparse representation.

5. CONCLUSIONS AND PERSPECTIVES

We suggested to apply a sparsity-promoting norm on the coefficients in a Gabor expansion. We also recalled how to calculate dual and tight Gabor frames for the situation most commonly encountered in audio signal processing. It seems highly recommendable to prefer tight frames if modification of the canonical coefficients is envisaged. In an application to classification of sound sources, experimental results indicate that sparse coefficients help to avoid false positives as compared to the results obtained from using canonical Gabor coefficients. Since sensitivity is comparable for both sets of coefficients, choosing a sparse representations leads to slightly better over-all classification results. The influence of various parameters, in particular the effects of the threshold in the Landweber iterations is yet to be investigated. Future work will also

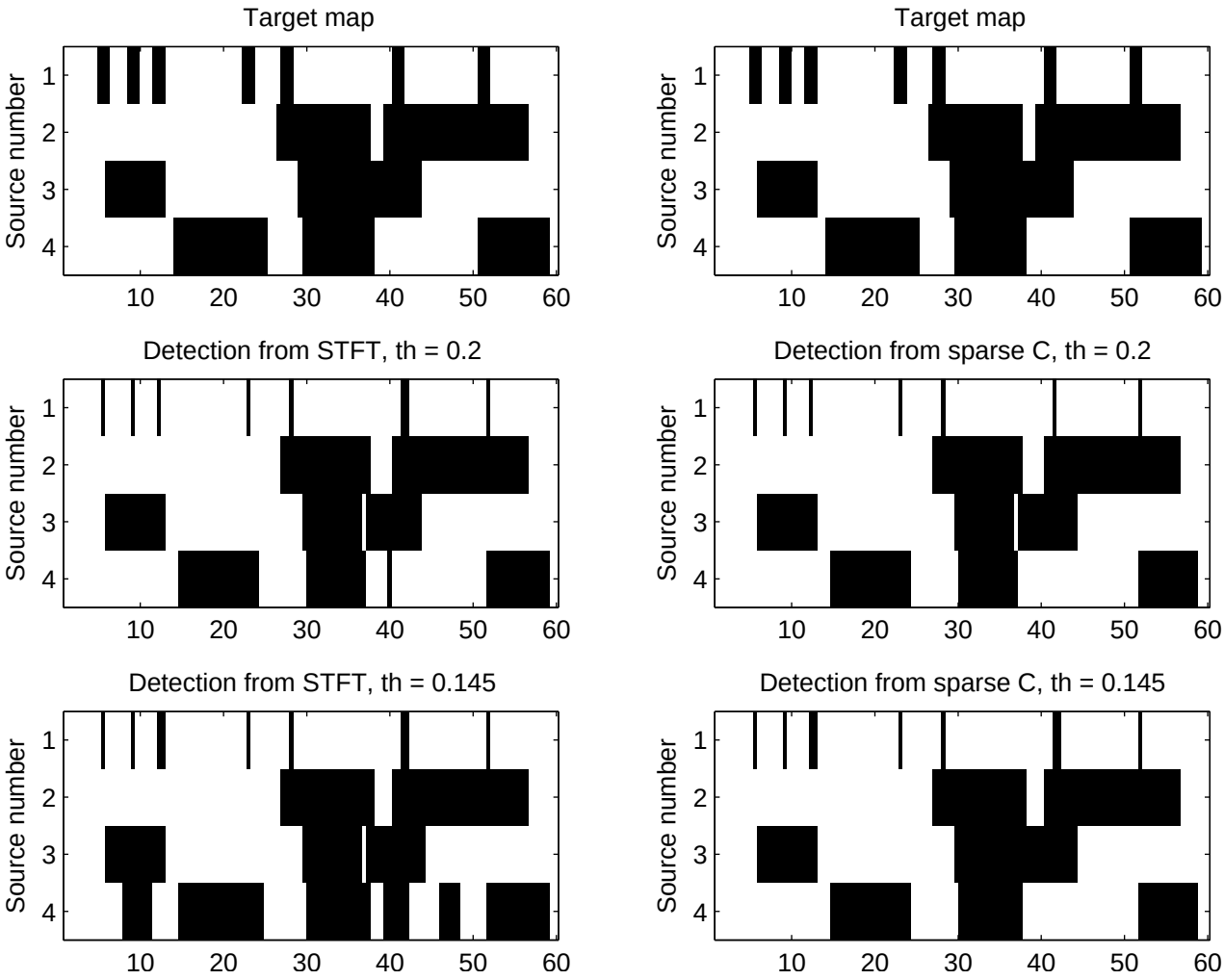

Figure 5. Detection from canonical and sparse coefficients

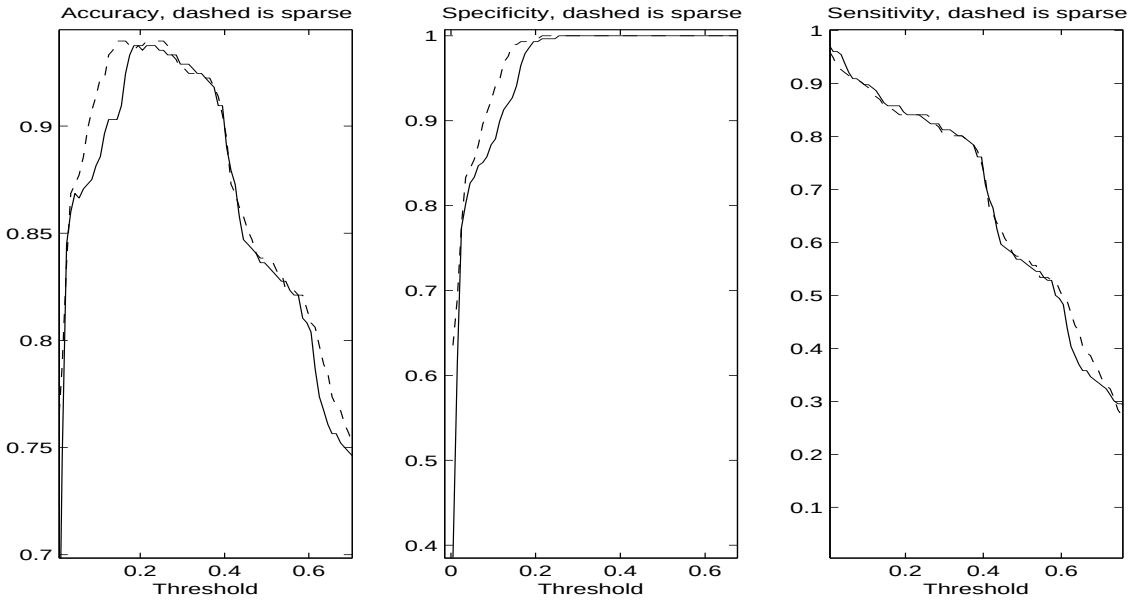

Figure 6. Evaluation of detection from canonical and sparse coefficients

include the application of more sophisticated coefficient norms as suggested in [12] as well as the usage of frames other than Gabor frames, e.g. wavelets, in order to include transient components. Furthermore, systematic evaluation on a more extensive data base will allow us to judge the influence of the various parameters involved.

6. ACKNOWLEDGMENTS

This research is supported by the Austrian Science Fund (Projects T384-N13 and P21247) and the Vienna Science and Technology Fund (WWTF, project Audiominer).
We thank the anonymous reviewers for their valuable comments and suggestions.

7. REFERENCES

[1] D. Hoiem, Y. Ke, and R. Sukthankar, "Solar: Sound object localization and retrieval in complex audio environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 429 – 432, March 2005.

[2] H. G. Feichtinger and T. Strohmer, *Gabor Analysis and Algorithms. Theory and Applications.* Birkhäuser, 1998.

[3] M. Dörfler, "Time-frequency Analysis for Music Signals. A Mathematical Approach," *Journal of New Music Research*, vol. 30, no. 1, pp. 3–12, 2001.

[4] N. Kaiblinger, "Approximation of the Fourier transform and the dual Gabor window," *J. Fourier Anal. Appl.*, vol. 11, no. 1, pp. 25–42, 2005.

[5] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis.," *IEEE Trans. Inform. Theory*, vol. 36, no. 5, pp. 961–1005, 1990.

[6] A. J. E. M. Janssen and T. Strohmer, "Characterization and computation of canonical tight windows for Gabor frames.," *J. Fourier Anal. Appl.*, vol. 8, no. 1, pp. 1–28, 2002.

[7] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l^1 solution is also the sparsest solution," *Commun. Pure Appl. Anal.*, vol. 59, no. 6, pp. 797–829, 2006.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[9] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 2009.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[11] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[12] M. Kowalski and B. Torrésani, "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients," *Signal, Image and Video Processing, doi:10.1007/s11760-008-0076-1*, 2009.

[13] M. Davies and L. Daudet, "Sparse audio representations using the MCLT," *Signal Process.*, vol. 86, pp. 457–470, March 2006.

[14] E. Ravelli, G. Richard, and L. Daudet, "Union of MDCT bases for audio coding," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 8, pp. 1361–1372, 2008.

[15] M. Goto, "Music scene description," in *Signal Processing Methods for Music Transcription* (A. Klapuri and M. Davy, eds.), pp. 327–359, New York: Springer, 2006.

[16] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. Springer, 2006.

[17] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180, IEEE, 2003.

[18] F. Jaillet and B. Torresani, "Timefrequency jigsaw puzzle: adaptive multiwindow and multilayered Gabor expansions," *Int. J. Wavelets Multiresolut. Inf. Process.*, vol. 2, pp. 293–316, 2007.

[19] S. Molla and B. Torrésani, "A hybrid scheme for encoding audio signal using hidden Markov models of waveforms.," *Appl. Comput. Harmon. Anal.*, vol. 18, no. 2, pp. 137–166, 2005.

[20] J. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Processing Letters*, vol. 11(6), pp. 553–556, June 2004.

[21] B. Torrésani and S. Molla, "Transient Detection and Encoding Using Wavelet Coefficient Trees,," in *proceedings of the GRETSI'01 conference* (F. Flandrin, ed.), 2001.