Allocating, Detecting and Mining Sound Structures. An Overview of Technical Tools.

Monika Dörfler*

NuHAG, University of Vienna, Austria
monika.doerfler@univie.ac.at
http://homepage.univie.ac.at/monika.doerfler/

Abstract. In this contribution, we relate the question of discernability of sound structures to the properties of the underlying analysis tools. In particular, we argue, that classical tools that are mainly used in sound processing and lead to features as prominent as the MFCC may be replaced by more accurate methods that are based on rather recent mathematical signal processing tools. In particular, we focus on adaptive representations that lend themselves to efficient computation and, on the other hand, on sparsity-promoting methods which are able to adapt to the structures present in a particular signal class.

Keywords: Sound structures, Time-Frequency, resolution, coefficient priors

1 Introduction

Sound signals play a central role in human life and the manner sound is perceived is highly sophisticated, complex and context-dependent. Since the amount of sound data that are automatically stored, searched and processed, grows dramatically, there is also a growing need for understanding the inherent structures of sound and their implications for human listeners. In particular, in many applications, one may be interested in distinguishing between what may be called a "sound-object" as opposed to more textural sound components that may rather be understood as an acoustical background.

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 structures seems to depend both on cultural and educational background, in the sense of what Smalley calls "experiential basis", cp. [14], that may be shared by a group of listeners. For example, a dense orchestral sound may be perceived as textural sound but also as a sequence and mixture of more compact, plastic, structures, that may be denoted as sound objects. The same observation holds true for environmental sounds and, even more so, for acousmatic music. From a certain point of view, the perception of sound components as background (textural) sound or object

* This research was supported by the WWTF project Audio-Miner (MA09-024)

(compactly structured) sound, therefore depends on the "zoom" the listener wishes to adopt or unconsciously assumes.

Here we propose that a sound component may be called "sound object" if it can be given a certain compact, sparse representation in a dictionary that has a perceptual interpretation, e.g., in a time-frequency dictionary. We will investigate this idea by means of describing some of the most important technical signal processing tools that lie at the base of most modern music information retrieval (MIR) methods. Now, while a lot of high level analysis is pursued and developed in the MIR community, some of the basic techniques are rarely called into question. However, the tools that provide the input for higher level analysis may been seen as a zoom chosen by the spectator and their influence should not be neglected.

Ideally, an analysis tool should be able to render a representation that allows for visual display reflecting a user's acoustical impression. In particular, sound objects should be visible as distinct from a more textural background. In this contribution we will describe several newly designed analysis tools, that render more sophisticated representations of sound signals than classical representations such as the short-time (or sliding window) Fourier transform. In particular, firstly, we show how adaptivity in the transformation parameters can sharpen the visual display while assuring a perfect connection between signal and representation in the sense of invertibility. Secondly, we show how various Bayesian coefficient priors enable us to highlight particular structures by means of informed analysis.

In the sense of reproducible research, all software involved in the production of the simulation examples is available, along with many additional examples and sound files, on the webpages mentioned in Section 2.2 and Section 2.3.

2 Technical tools

We now describe the tools that are at the heart of any sound analysis. While usually the methods are seen as FFT-based time-variant analysis, we take a mathematical point of view and describe the involved dictionaries as *frames*, [3]. It will turn out that this slightly more abstract point of view opens the door to a myriad of useful generalizations of the classical analysis of sound via windowed FFT.

2.1 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 (STFT) of $x[n]$ is given, for $k \in \mathbb{Z}$ and $\omega \in [-0.5, 0.5]$ by $\mathcal{V}_\varphi x(k,\omega) = \sum_{n=-\infty}^{\infty} x[n]\varphi[n-k]e^{-2\pi i\omega n}$. In practice, a subsampled version of the STFT is 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, [7], 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 sub-sampling the STFT, as inner products between the signal x and *time-frequency-shifted* windows. For x and $y \in \mathbb{C}^L$, the inner product is defined as $\langle x, y\rangle = \sum_{n=0}^{L-1} x[n]\overline{y[n]}$. Further, we let $T_k x[n] := x[n-k]$ be called translation operator (time shift) by k and $M_l x[n] := e^{\frac{2\pi i l n}{L}} x[n], l \in \mathbb{Z}$ be called modulation operator (frequency shift) by l.
The family

$$\varphi_{k,m} := M_{mb} T_{ka} g \tag{1}$$

for $m = 0, \ldots, M-1$ and $k = 0, \ldots, K-1$, where $Ka = Mb = L$, is a *Gabor analysis system*. The theory of *frames* gives the appropriate framework for analysis using Gabor systems: A set of Gabor analysis functions $\varphi_{k,m}$ in $L^2(\mathbb{R})$ is called a Gabor frame, if there exist constants $A, B > 0$, so that, for all $f \in \mathbf{L}^2(\mathbb{R})$

$$A\|f\|^2 \leq \sum_{k,m \in \mathbb{Z}\times\mathbb{Z}} |\langle f, \varphi_{k,m}\rangle|^2 \leq B\|f\|^2. \tag{2}$$

This 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:

$$Sf = \sum_{k,m \in \mathbb{Z}\times\mathbb{Z}} \langle f, \varphi_{k,m}\rangle \varphi_{k,m} \tag{3}$$

The invertibility of S leads to the existence of so-called dual frames, yielding convenient reconstruction formulas via the canonical *dual frame* $\tilde{\varphi}_{k,m}$, given by $\tilde{\varphi}_{k,m} = S^{-1}\varphi_{k,m}$. 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 $(\tilde{\varphi}_{k,m})$.

Hence,

$$f = S^{-1}Sf = \sum \langle f, \varphi_{k,m}\rangle \tilde{\varphi}_{k,m} \tag{4}$$

In the finite discrete case of $x \in \mathbb{C}^L$ a collection $\{\varphi_{k,m}\} \in \mathbb{C}^L$ 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. Then, the condition number of the frame operator is given by the fraction of the maximum and minimum eigenvalue, respectively. 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:

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

From the definition of the frame operator a straight-forward calculation (see [6] for details) shows that the single entries of S are given by

$$S_{ji} = \begin{cases} M \sum_{k=0}^{K-1} T_{ka} g[j] \overline{T_{ka} g[i]} & \text{if } |j-i| \bmod M = 0 \\ 0 & \text{else} \end{cases} \tag{5}$$

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 $g(j)$ and $g(i)$ are both have non-zero. Therefore, the frame operator is diagonal and the dual window $\tilde{\varphi}$ is calculated as

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

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. 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.
Since S is positive and symmetric we may write

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

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:

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

Remark 1. One may ask the question, why the consideration of a dual system that guarantees reconstruction has any relevance for analysis. For example, the constant-Q transform has been used with success without being an invertible transform, cf [2]. However, without a reconstruction system at hand, we lack a precise connection between our analysis and the original signal. We may have lost important parts of the signal, we may not correctly interpret the coefficients, if the reconstruction window is very different from the analysis window. In this sense, the tight system generated by $\varphi^t = S^{-\frac{1}{2}} \varphi$ bears an important advantage since it is closest to the original window among all functions h generating a tight frame for lattice constants a and b, cf. [10]. Hence, φ^t 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 analysis 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.

Due to these remarks, tight Gabor frames will be used in our subsequent experiments. We now turn to an important generalization of the time-frequency dictionaries introduced as Gabor frames: we can achieve adaptive frames by allowing for changing windows in either time or frequency. If some mild side-conditions are fulfilled, the analysis *and reconstruction* is similarly straightforward and computationally efficient as in the regular case.

2.2 Adaptivity

While in classical Gabor frames, as introduced in the previous section, we obtain all samples of the STFT by applying the same window φ, shifted along a regular set of sampling points and taking FFT of the same length. Exploiting the concept of frames, we can achieve adaptivity of the resolution in either time or frequency. To do so, we relax the regularity of the classical Gabor frames, which leads to *nonstationary Gabor frames* (NSGT): For $(k, m) \in I_M \times I_M$, we set

(i) $\varphi_{k,m} = \mathbf{M}_{mb_k}\varphi_k$ for adaptivity in time.
(ii) $\varphi_{k,m} = \mathbf{T}_{ka_m}\varphi_m$ for adaptivity in frequency.

A detailed mathematical analysis of NSGTs is beyond the scope of this contribution, but we wish to emphasize, that both analysis *and* synthesis can be done in a similar manner as in the regular case, that is, a diagonal frame operator can be achieved and perfect reconstruction is guaranteed by using either dual or tight windows. For all details, see [15, 1].

Examples and interpretation of adaptive transforms We now illustrate the influence of adaptivity on the visual representation of audio signals. First, an analysis of a short excerpt of G. Ligeti's piano concert is given. This signal has percussive onsets in the piano and Glockenspiel voices and some orchestral background. Figure 1 first shows a regular Gabor (STFT) analysis and secondly, a representation in which the percussive parts are finely resolved by an adaptive NSGT.

Our second example is an excerpt from a duet between violin and piano, by by J.Zorn. We can see three short segments: A vivid sequence of violin and piano notes followed by a calm violin melody with accompanying piano and finally an inharmonic part with chirp component. For this signal, we show an FFT-based Gabor transform (STFT) and an NSGT-based constant-Q transform in Figure 2. In both cases the display of the frequency axis is logarithmic. It is obvious, that the NSGT, with adaptivity in the frequency domain, provides more

accurate resolution of the harmonic components, in particular in low frequency regions. Note that with MFCC, very popular features used in speech and music processing, [12], are obtained from an FFT-based STFT, using a logarithmic spacing of the frequency bins, while the analysis windows are linearly spaced. Given the new opportunities offered by adaptive NSGT it may well be worth reconsidering the underlying basic analysis.

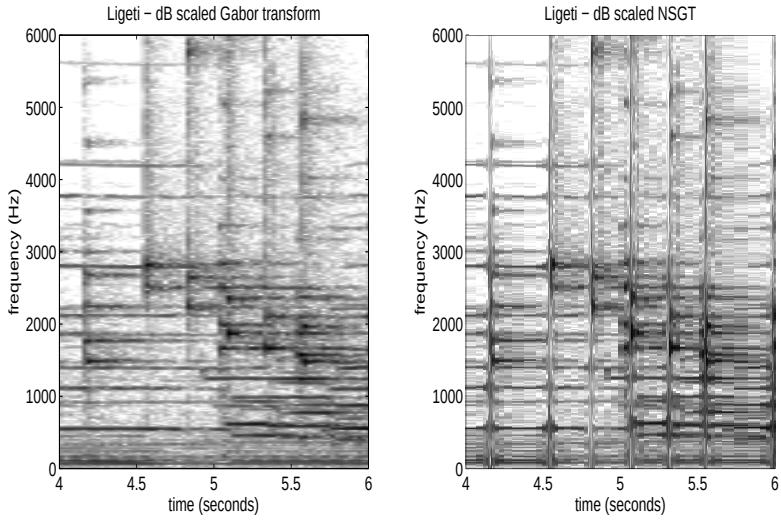

Fig. 1. Ligeti - Regular and nonstationary Gabor representations.

Returning to the quest for salient "sound objects" that stand out from their background, these examples show well, that the analysis tool influences, even by visual inspection, what may be considered as such. In particular, in the Ligeti example, the zooming-in onto the percussive onsets makes these components more distinguishable from their background. On the other hand, the harmonic parts require less coefficients, since they are represented by longer windows. It should be noted that, for further processing, e.g. the extraction of percussive components, this kind of representation is beneficial.
Even more impressively, in the low frequency components of the second example, the single harmonics are not resolved at all in the FFT-based transform, while the NSGT-transform clearly separated them from a soft noise-floor background. Again, apart from pure visual evaluation, frequency separation of single components is necessary for applications such as transposition, cp. [15].
More visual and audio examples for adaptivity in both time and frequency can be found on `http://www.univie.ac.at/nonstatgab/`.

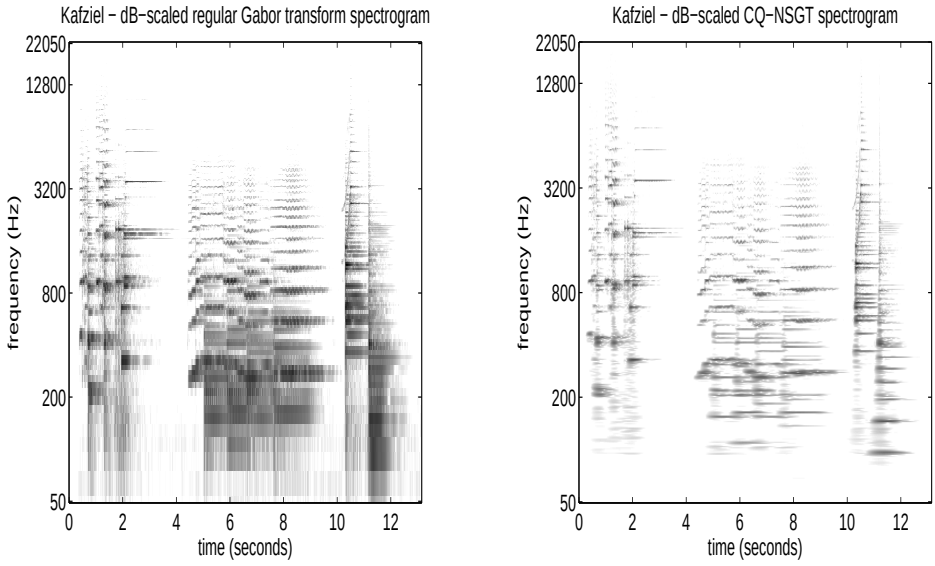

Fig. 2. Time-frequency representations on a logarithmically scaled frequency axis: Regular Gabor spectrogram (left) and constant-Q NSGT spectrogram (right).

2.3 Sparsity

If we are convinced, that the signal components of interest have a sparse, at least approximative, representation in the dictionaries we use, then we may look for relevant coefficients only. A sparse approximation has a small number of nonzero elements, while still giving a satisfying representation and reconstruction of a signal or a certain signal component. One way to enforce sparsity is to choose an expansion of x such that as many coefficients as possible are zero. Mathematically, however, minimization of an ℓ^1-constraint on the coefficients yields explicit solutions and fast algorithms as well as similar solutions.[2] In the present situation, we are going to minimize the following expression for a tight Gabor frame with elements $\varphi^t_{k,m}$:

$$\Delta(x) = \| \sum_{k,m} c_{k,m} \varphi^t_{k,m} - \hat{x} \|_2^2 + \mu \|\mathbf{c}\|_{\ell^1} \tag{6}$$

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 basis (instead of frames), the problem formulation in (6) 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. To

[2] Note, that it has been proved that certain situations ℓ^1-minimization in fact yields the optimally sparse solution, see [5].

find the solution of (6), we then consider the sequence of iterates

$$\mathbf{c}^n = \mathbb{S}_\mu(\mathbf{c}^{n-1} + \mathcal{V}_{\varphi^t}(\hat{x} - \mathcal{V}^*_{\varphi^t}\mathbf{c}^{n-1})), \tag{7}$$

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

$$\mathbb{S}_\mu(z) = \arg(z)(|z| - \mu)^+ \tag{8}$$

According to [4], the corresponding iterative algorithm converges to the solution of (6). The corresponding algorithm is known under the name "The Lasso".

Structured Sparsity The ℓ^1-norm acts independently on each coefficient and therefore does not take the correlations between time-frequency coefficients which are typical for most natural audio signals into account. Coefficient correlation will be particularly pronounced in coefficients contributing to one sound-object. It was therefore suggested in [11] and successfully studied for the denoising of music signals in [13], to introduce neighborhood-systems to emphasize the structural connections in time-frequency representations. This idea is formalized by replacing the thresholding operator (8) by a component-wise weighted thresholding as

$$\mathbb{S}_\omega(c_\gamma) = c_\gamma(1 - \frac{\mu}{(\sum_{\gamma' \in U_\gamma} \nu_\gamma(\gamma')|c_{\gamma'}|^2)^{1/2}})^+, \tag{9}$$

where U_γ is the environment of each coefficient c_γ that is relevant for the weighting process and ν_γ is the sequence of corresponding weights.

This new, neighborhood-smoothed thresholding operator replaces $\mathbb{S}_\mu$ in the iteration (7), leading to what we call *weighted group Lasso* (WGL). The neighborhoods' shapes are parametrized by their size and corresponding weights. For the experiments presented here, we allowed for rectangular domains with either uniform (i.e. rectangular) or triangular (i.e. "tent"-like) weightings. Various other shapes and weights may incorporate prior knowledge about the signal of interest, for example, neighborhoods consisting of several non-connected sets corresponding to expected harmonic structures. The shapes do not necessarily have to be symmetric, since the energy of most audio signals is typically not symmetrically distributed around its peaks either. This observation can be exploited as shown next.

Examples and interpretation of structured sparsity Consider Figure 2.3, where the iterated WGL-shrinkage results with four different neighborhood-shapes, each only extending in time, are compared (based on the analysis Gabor-frame with redundancy 4).

The different neighborhood shapes lead to different sparse views of the particular signal component, giving the extracted component perceptually distinguishable appearances. Whereas the symmetric neighborhoods captures signal energy both before and after the attacks, the asymmetric neighborhoods emphasize

components before (resp. after) the attacks. The orientation of the neighborhood therefore systematically promotes the preservation of different temporal segments of the signal.

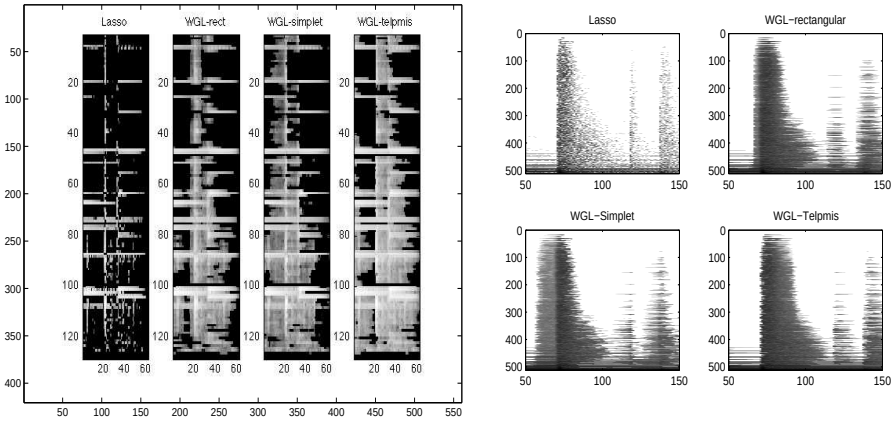

Fig. 3. Iterated WGL shrinkage results for different shapes (i.e. weightings) of the neighborhood on an excerpt including piano, double-bass and drums and on an snare drum hit excerpt. For both signals, the result of unsmoothed shrinkage, Rectangular, Simplet (= simple tent, starting at 1 and then linearly decaying to zero), Telpmis (= time-reversed simple tent) weighted smoothing is shown.

More examples for representations obtained from structured sparsity constraints, together with the corresponding sound files, can be found on `http://homepage.univie.ac.at/monika.doerfler/StrucAudio.html`.

3 Discussion and Future Work

In this contribution we showed how, even by visual inspection, the choice of various representations that exploit prior knowledge about a signal (class) of interest, can influence the resulting analysis. It will and should be the topic of further, and necessarily interdisciplinary, research to scrutinize the influence of these choices on the performance of higher-level processing steps. Some preliminary steps in this direction have been pursued within the research project *Audio Miner*, cf. `http://www.ofai.at/research/impml/projects/audiominer.html` and [9, 13, 8] and shown promising results. We strongly believe, that using appropriate, still concise, representations of the original data is important to avoid biased results in higher-level processing steps.

Acknowledgments. The author gratefully acknowledges the many discussions with all AudioMiner team members concerning the intricate topic of sound objects and appreciates the thoughtful remarks of two anonymous reviewers.

References

1. P. Balazs, M. Dörfler, F. Jaillet, N. Holighaus, and G. A. Velasco. Theory, implementation and applications of nonstationary Gabor Frames. *J. Comput. Appl. Math.*, 236(6):14811496, 2011.
2. J. Brown. Calculation of a constant Q spectral transform. *J. Acoust. Soc. Amer.*, 89(1):425434, 1991.
3. A. Chebira and J. Kovacevic. Life Beyond Bases: The Advent of Frames (Part I and II). *IEEE Signal Processing Magazine*, 24(4/5):86–104,115–125, 2007.
4. I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.
5. 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.*, 59(6):797–829, 2006.
6. M. Dörfler. Time-frequency Analysis for Music Signals. A Mathematical Approach. *Journal of New Music Research*, 30(1):3–12, 2001.
7. H. G. Feichtinger and T. Strohmer. *Gabor Analysis and Algorithms. Theory and Applications.* Birkhäuser, 1998.
8. M. Gasser, A. Flexer, and T. Grill. On Computing Morphological Similarity of Audio Signals,. *Proceedings of the 8th Sound and Music Computing Conference , Padova, Italy*, 2011.
9. A. Holzapfel, G. Velasco, N. Holighaus, M. Dörfler, and A. Flexer. Advantages of nonstationary Gabor transforms in beat tracking. In *Proceedings of MIRUM11.*, November 2011.
10. A. J. E. M. Janssen and T. Strohmer. Characterization and computation of canonical tight windows for Gabor frames. *J. Fourier Anal. Appl.*, 8(1):1–28, January 2002.
11. M. Kowalski and B. Torrésani. Structured Sparsity: from Mixed Norms to Structured Shrinkage. *SPARS'09 - Signal Processing with Adaptive Sparse Structured Representations (2009)*, 2009.
12. B. Logan. Mel frequency cepstral coefficients for music modeling. In *ISMIR,Plymouth, Ma (USA)*, 2000.
13. K. Siedenburg and M. Dörfler. Audio denoising by generalized time-frequency thresholding. In *Proceedings of the AES 45th Conference on Applications of Time-Frequency Processing,Helsinki, Finland*, 2012.
14. D. Smalley. Spectromorphology: explaining sound-shapes. *Organised Sound*, 2(2), August 1997.
15. G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill. Constructing an invertible constant-Q transform with non-stationary Gabor frames. *Proceedings of DAFX11*, Paris, 2011.