Removing Components from a Time-Frequency Representation

Monika Dörfler, Peter Balazs and Florent Jaillet

Acoustics Research Institute, A-1040 Vienna, Austria

March 17, 2009

Introduction

Time-frequency representations such as the spectrogram or short-time Fourier transform seem to be well suited to the task of removing 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. However, the trivial approach of just deleting the corresponding component in the time-frequency representation leads to artifacts such as "ghost-tones", musical noise and "phasing effects". We compare different approaches to tackle this problem: To gain some insight in the nature of the problem, we consider the optimal separating mask for known signal components. A soft-thresholding procedure is applied and compared to a model promoting sparsity in the representation. The latter can lead to favorable results by yielding a feasible approximation of the optimal mask. The results are presented visually and acoustically during the presentation.

Time-frequency representations

Time-frequency analysis aims at providing simultaneous information on a signal's time- and frequency-content. In order to clarify this idea, time-frequency representations are often compared to a music-score, which, in fact, very efficiently conveys the information which frequency, or rather pitch, should sound at which instant. For the mathematical understanding, however, this comparison may be misleading, as, according to Heisenberg's uncertainty principle, an exact separation of signal components in the time-frequency domain ia not possible. In particular, when it comes to removing certain components from a signal's time-frequency representation, it is usually not easy to delete the components' contribution to the time-frequency coefficients, for several reasons. First of all, no window function can be band limited and have a compact support at the same time. This statement is equivalent to saying that no ideal low-pass FIR filter can exist. Secondly, the phase information of the time-frequency representation seems to be of fundamental importance, yet is not fully understood, also see [8] in this volume. We next proceed to define the basic mathematical notions.

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

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

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, [6], 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. Under certain conditions, usually fulfilled in practice, Gabor theory yields a convenient reconstruction method by using either a dual window in the synthesis step or a tight window for both analysis and synthesis, see [4] for the corresponding formulas. Writing $\tilde{\varphi}$ for the dual window, we obtain the following perfect reconstruction formula:

$$x[n] = \sum_{k=1}^{N} \sum_{m=1}^{M} \mathcal{V}_\varphi x(ka, mb)\tilde{\varphi}[n-ka]e^{\frac{2\pi i n m}{M}}. \quad (2)$$

In fact, the Gabor transform is commonly used in acoustics, but often in a disguised form, called sliding window transform, which, in fact, equals the classical Gabor transform up to a phase factor, see [4]. In applications, the obtained coefficients are usually modified before reconstruction, yielding an output of the form:

$$Gx[n] = \sum_{k=1}^{N} \sum_{m=1}^{M} \mathbf{m}(m,k)\mathcal{V}_\varphi x(ka, mb)\tilde{\varphi}[n-ka]e^{\frac{2\pi i n m}{M}}, \quad (3)$$

where $\mathbf{m}$ denotes the mask which is applied to the coefficients.

From a mathematical point of view, this approach generates the interesting situation, that we reconstruct a signal from coefficients which are not the canonical coefficients of any existing original signal. Figure 1 illustrates this effect. It is easily seen that even in this fairly simple synthesized example, the separation of signal (chirp) and noise is not as easy as it might seem at first sight. The next section shows how optimal reconstruction is obtained in this situation.

The optimal mask

When trying to design an appropriate mask for signal separation, we have to keep in mind that phase plays a central role in the interpretation of (sampled) short-time Fourier transforms. In particular, if we know the signals
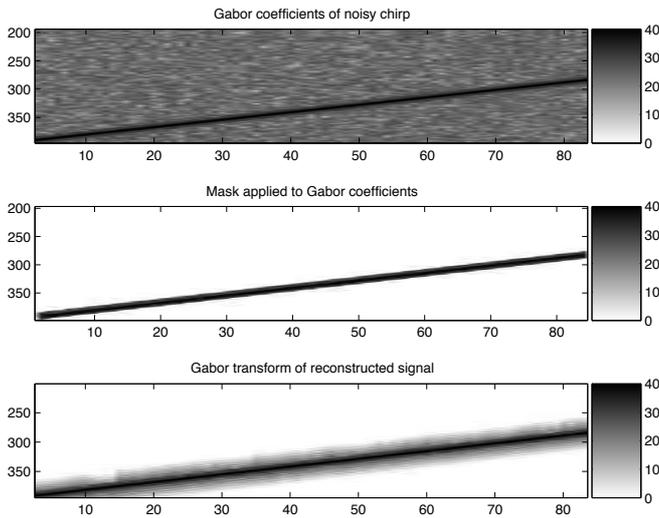
Figure 1: Gabor coefficients of a synthesized signal (chirp + noise). It can be seen that the reconstruction introduces smearing due to the spread of the window's FFT

at hand, we may calculate an optimal, though highly signal dependent mask as follows. Let us assume that we know (a good estimate of) our noise signal n, the signal of interest is denoted by x and $\hat{x} = x + n$ is the observed signal. Then, if we use, whenever $|\mathcal{V}_\varphi \hat{x}(ka, mb)|^2 > 0$ for all k, m

$$\mathbf{m}_{opt}(k, m) = \frac{\mathcal{V}_\varphi x(ka, mb) \cdot \overline{\mathcal{V}_\varphi \hat{x}(ka, mb)}}{|\mathcal{V}_\varphi \hat{x}(ka, mb)|^2}, \qquad (4)$$

we obtain, by a straight-forward multiplication

$$\mathcal{V}_\varphi x(ka, mb) = \mathbf{m}_{opt}(m, k) \cdot \mathcal{V}_\varphi \hat{x}(ka, mb). \qquad (5)$$

Obviously, this observation is of minor practical interest, as the signal x may not be observed directly. However, we may take a look at the behavior of $\mathbf{m}_{opt}$ in order to gain a better understanding of the process of masking in the time-frequency domain.

We consider the synthesized signal from Figure 1 again, this time calculating the optimal mask according to (4) and we look at the inverse effect, see Figure 2. Cutting out the chirp according to the mask depicted in Figure 1, the "common" approach to masking out components, we look at the Gabor coefficient of the resulting reconstruction, second display. The third display, then, shows the modulus of the optimal mask and the Gabor coefficients of the resulting signal, pure noise, can be seen in the last display. Interestingly, if the specific realization of the noise changes, the resulting noise-signal still does not reveal the original (chirp) signal, as opposed to the result from applying a "common" masking procedure.[1]

As stated above, using an "optimal" mask, can hardly be an approach of practical interest. However, it seems reasonable to reconsider the classical approach of changing only the modulus of the Gabor coefficients without touching the phase. As a starting point, we suggest extensive studies on the combination and separation of simple signals in order to achieve a better

---

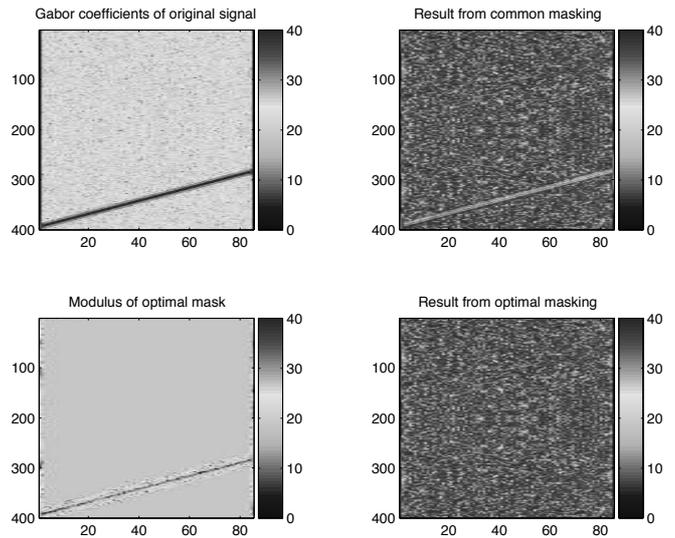[1] Audio examples will be provided during the presentation.

Figure 2: Gabor coefficients of a synthesized signal (chirp + noise), top left, of a signal obtained from commonly used masking as in Figure 1, top right, (modulus of) optimal mask, bottom left, and Gabor coefficients of resulting signal, bottom right.

understanding of masking of time-frequency components. It is interesting to note, that the coefficients obtained by optimal masking according to (4) yield the Gabor coefficients of some signal.[2]

Soft thresholding

For denoising, i.e. the separation of noise and signal components, in the frequency domain, various models based on a statistical approach have been suggested, compare [7]. Here, by using an estimate of the power spectrum $S_n(k, m)$ of the noise for each frequency bin k and every time-point m, different soft-thresholding procedures are realized. For simplicity of notation, let us write, $\mathcal{V}_\varphi \hat{x}(k, m) = \mathcal{V}_\varphi \hat{x}(ka, mb)$. A suppression rule based on a "pseudo-Wiener" noise reduction is given by:

$$\mathcal{V}_\varphi x(k, m) = \rho(k, m) \cdot \mathcal{V}_\varphi \hat{x}(k, m), \qquad (6)$$

where

$$\rho(k, m) = \begin{cases} \frac{|\mathcal{V}_\varphi \hat{x}(k,m)|^2 - S_n(k,m)}{|\mathcal{V}_\varphi \hat{x}(k,m)|^2}, & |\mathcal{V}_\varphi \hat{x}(k,m)|^2 > S_n(k,m) \\ 0 & \text{else} \end{cases}$$
$$(7)$$

Variants of the above formula exist, leading, e.g., to the power subtraction method, if $\sqrt{\rho(k, m)}$ is used in place of $\rho(k, m)$. We display the result of denoising according to (6) and (7) in Figure 3. In this experiment, we consider a music signal (piano, bass, drums), synthetically corrupted by Gaussian white noise. We estimate the average noise level S_n from the Gabor coefficients of n and $\rho(k, m)$ is then obtained directly from the Gabor coefficients of the noisy signal. The first display of Figure 3 shows the Gabor coefficients of

---

[2] As mentioned before, this is by no means obvious for a general mask applied to Gabor coefficients, as the coefficients have to satisfy a reproducing kernel equation, see [5] for details.

the noisy signal (redundancy is 4), while the coefficients selected by the thresholding procedure can be seen in the second display. The last display points out once more, that the Gabor coefficients of the reconstructed signal are not equal to the masked coefficients. It
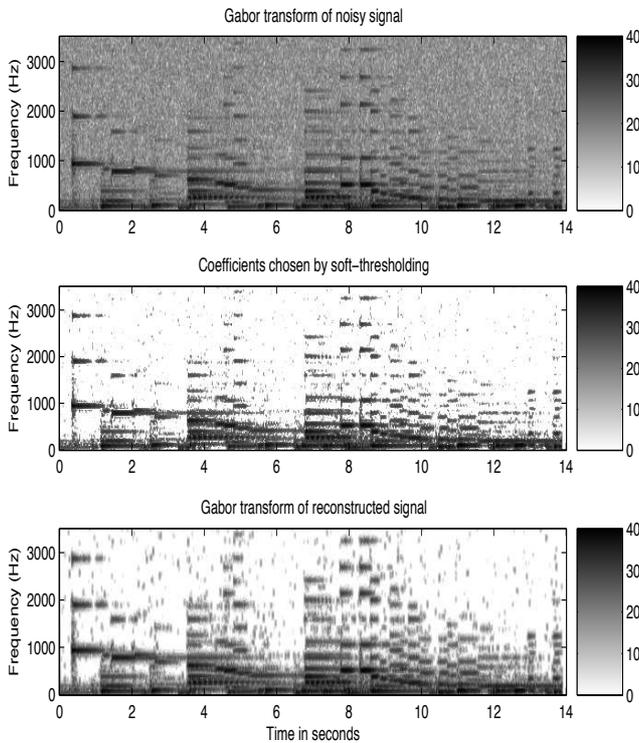

Figure 3: Denoising by soft thresholding.

is also obvious and well-known, that one of the most common problems related to thresholding methods is the regular occurrence of artifacts, most notably, of musical noise, which arises from the randomness inherent in the estimation of the signal power spectrum used. Several advanced suppression rules have been suggested to tackle this problem, see [10, 7] and references therein.

Sparse representations

We next suggest a novel approach to the separation of various signal components in the time-frequency domain. Partly, the problems arise from the fact that coefficients belonging to a particular, possibly isolated or well-concentrated signal component is smeared as a consequence of the analysis window's properties. in this way, components which are expected to be separated in time-frequency may have essentially over-lapping Gabor coefficients.

On the other hand, if we are convinced, that the signal components of interest have a sparse, at least approximative, representation in the atomic systems we use, then we may avoid the problem by looking 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.[3] In the present situation, we are going to minimize the following expression:

$$\Delta(x) = \|\sum_{k,m} c_{k,m}\tilde{\varphi}_{km} - \hat{x}\|_2^2 + \mu\|\mathbf{c}\|_{\ell^1} \qquad (8)$$

where $\tilde{\varphi}_{km} = \tilde{\varphi}[n - ka]e^{\frac{2\pi i n m}{M}}$ and $\|\mathbf{c}\|_{\ell^1} = \sum_{k,m}|c_{k,m}|$ is the ℓ^1-norm of the coefficient sequence. The solution can be found by an iterative algorithm called Landweber iterations [2]. We next present the results of two experiments. First, we 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 samples length and calculate the Gabor coefficients, shown in the first display of Figure 4. The second display, then, shows the coefficients resulting from ℓ^1-penalization on the expansion coefficients according to (8). It is immediately obvious, that the algorithm visually separates the two signal components. Note that reconstruction from these coefficients is possible, but not perfect, as may be seen in Figure 5. In fact, we face a trade-off between accuracy of reconstruction and sparsity of representation, which is reflected in the choice of μ in (8). However, the sparse representation allows the separation of signal components by designing a mask that suppresses undesired signal components more easily than the usual, more redundant representation.

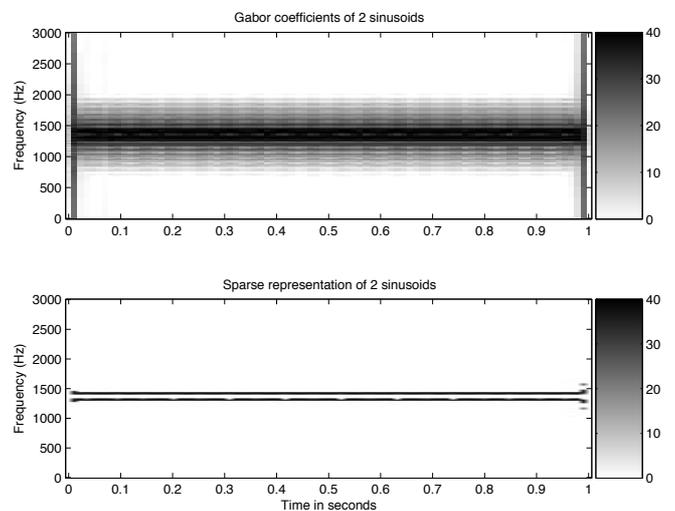For comparison, we also apply the Landweber iteration

Figure 4: Gabor coefficients and sparse representation of two close sinusoids

algorithm to the audio signal considered in the previous section. The results are displayed in Figure 6. Quite obviously, we encounter problems similar to those resulting from soft-thresholding. However, in the present approach, we have not made any a priori assumptions about the signal, in particular, about the statistical properties of the noise. We point out, that sparsity approaches bear the potential of more sophisticated

---

[3]Note, that it has been proved that certain situations ℓ^1-minimization in fact yields the optimally sparse solution, see [1].
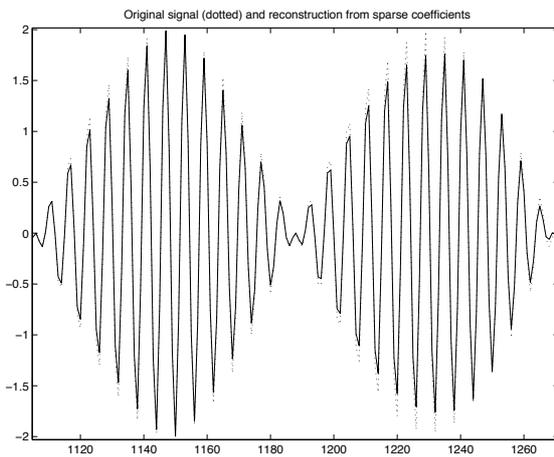
Figure 5: Original signal (dotted) and reconstruction from sparse coefficients

penalization terms leading to results better adapted to the signals at hand, compare [9]. Also note that the resulting coefficients heavily depend on the choice of μ, which is usually a difficult task.
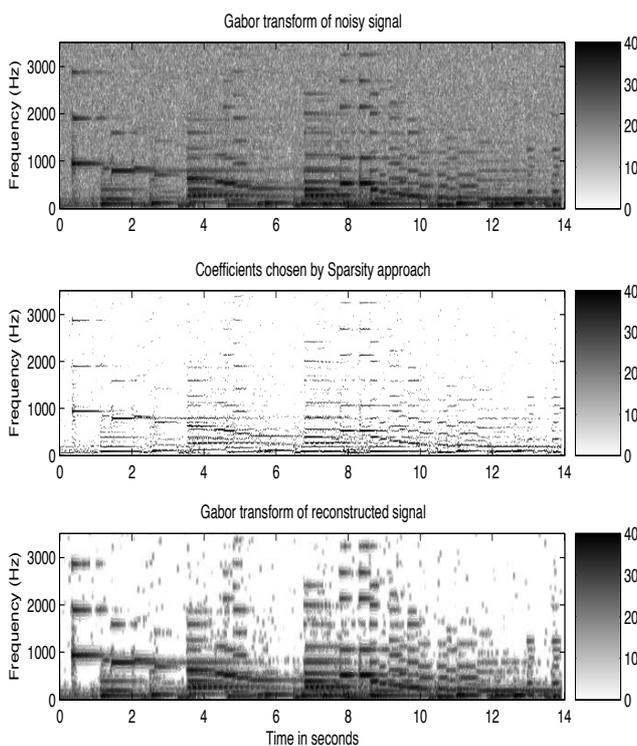

Figure 6: Denoising by sparsity approach.

Summary and Outlook

We summarized and compared various approaches to the problem of separating possibly over-lapping components in the time-frequency domain. While (soft) thresholding is a commonly used approach to tackle this task, we also mentioned signal separation by means of a sparsity approach. We believe, that the latter bears the potential to perform better than thresholding procedures in certain situations. Note that so far we have not performed any (pre-)denoising on the signal. Also, in order to

achieve favorable performance, the procedure has to be further refined. An additional problem connected to sparsity methods is the slow convergence of the iterative algorithms involved. Alleviations have been suggested, see, e.g. [3]. We also suggested to look at the optimal mask for separation of signal components, in order to gain insight in the nature of masking procedures in the time-frequency domain. We pointed out that, in contrast to common practice, approaches involving complex-valued masks may be well worth considering.

Acknowledgments

References

[1] S. S. Chen, David L. Donoho, and M. A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1999.

[2] 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.

[3] I. Daubechies, Massimo Fornasier, and I. Loris. Accelerated projected gradient methods for linear inverse problems with sparsity constraints. *J. Fourier Anal. Appl.*, to appear.

[4] M. Dörfler. Time-frequency Analysis for Music Signals. A Mathematical Approach. *Journal of New Music Research*, 30(1):3–12, 2001.

[5] M. Dörfler and B. Torresani. Representation of operators in the time-frequency domain and generalized Gabor multipliers. *Submitted, arXiv:0809.2698*, 2008.

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

[7] Simon J. Godsill and Peter J. W. Rayner. *Digital Audio Restoration.* Springer, 1998.

[8] F. Jaillet, P. Balazs, and M. Dörfler. On the structure of the phase around the zeros of the short-time Fourier transform. In *Proceedings of the International Conference on Acoustics, Rotterdam, March 23-26*, 2009.

[9] M. Kowalski and B. Torrésani. Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients. *Signal, Image and Video Processing, to appear*, 2009.

[10] P. Wolfe and S. Godsill. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. *EURASIP J. Appl. Signal Process. Special Issue: Audio for Multimedia Communications.*, 2003(10):1043–1051, 2003.