

# WHAT TIME-FREQUENCY ANALYSIS CAN DO TO MUSIC SIGNALS

(and what it can't do . . .)

Monika Dörfler\*

NuHAG, Institut für Mathematik, Universität Wien  
e-mail: monika.doerfler@univie.ac.at

April 20, 2004

## 1 Introduction

It seems to be almost common knowledge that there are strong connections between mathematics and music. Both mathematics and music appear as highly structured disciplines with a desire for transparency, lucidity and a certain kind of simplicity. Musical phrases obeying strict structures, though often not consciously understood as such by the listener, are perceived as particularly pleasing. The rather rigorous rules of counterpoint, as, e.g., in the compositions of J. S. Bach and his contemporaries, are a typical example, a more recent one is given by the harmonic regularities in jazz harmony.<sup>1</sup> On the other hand, Bach as well as all famous improvisers in jazz don't recoil from a controlled breach of the rules – leading to the advent of new perspectives and connections. The same principles hold analogously for mathematics: first, the given rules of a mathematical theory have to be thoroughly studied and practised in order to be able to finally make a step beyond their limits.

Also in a historical sense, the mutual influences between mathematics and music are manifold. Mathematical diagrams such as the cartesian coordinates developed under the influence of models from music notation, where staves with the x-coordinate representing time and the y-coordinate representing frequency started to be used in the eleventh century – long before they were introduced in geometry. Diagrams resulting from time-frequency analysis, the topic of this article, can even be interpreted as a generalized music notation, as we will explicate below.

Another example for the importance of influences from music theory in mathematics is the logarithm, which was used by music theorists in an intuitive way long before the introduction of its abstract notion in mathematics. Apparently, the human mind tends to apply similar models for mapping both musical and mathematical findings to a descriptive medium. However, the mapping can never be an "invertible" one. The description of a highly complex processes, such as the performance *or* the perception of music is bound to be an approximation. Written music is only a faint image of the sound in the composer's head which needs musicians to be filled with sound, listeners to be filled with relevance. As a consequence, mathematical formalism as such can only be a fragment in a communication chain for the understanding

---

\*Supported by the FWF, Project P14485

<sup>1</sup>Mark Levine, the author of several important textbooks on jazz harmony and theory, gives a detailed account of harmonic structures in improvisation in [1].

of data as complex as music.<sup>2</sup> In the next we section briefly describe two possible tasks which may arise in a mathematical analysis of music and point out to which parts of the complex accomplishment of solving these problems the mathematics of time-frequency analysis can contribute.

## 2 Problems

- **Transcription**

By transcription we mean the extraction of an acceptable notation from performed music, i.e., from a sound signal, which has been digitalized and can thus be numerically processed. Transcription is a task which is well accomplished by trained humans. However, the automatic transcription problem is still unsolved for any class of fairly complex music signals. In fact, tackling the task of transcription requires the consideration of several levels of processing. First of all, an appropriate signal representation must be achieved, often by time-frequency methods from signal analysis. Pattern recognition methods or methods from artificial intelligence then yield information on onset timing, [3, 4], and fundamental frequencies, [5, 6, 7], resulting in data in a format similar to MIDI<sup>3</sup>. Finally, it seems to be a highly non-trivial issue in artificial intelligence to obtain written music from MIDI data, see, e.g., [3, 9].

In fact, mathematics can mainly contribute to the low-level part of processing concerning the quality of the time-frequency methods used. In the sequel we describe mathematical means to achieve a useful representation of music signals by means of time-frequency analysis. In particular, we introduce the theory of *Gabor analysis*, which can be closely related to the concept of music notation. Furthermore ... a representation which we call *multiple Gabor systems*.

- **Time-varying filtering**

Music signals are prototypical examples for time-varying signals. Hence, in order to extract, for example, the part of a musical piece which is performed by a certain instrument or group of instruments, from a given sound signal, time-variant filtering methods have to be applied. As opposed to linear time-invariant filtering, in which the spectrum is multiplied by a transfer function, in time-varying filtering the transfer function explicitly depends on time. We discuss a specific tool for time-varying filtering called *Gabor multipliers*. An overview of existing methods for time-varying filtering based on joint time-frequency distributions has been given in [10]. Gabor multipliers are a generalization of short-time Fourier transform multipliers, which are discussed in the above-mentioned article.

Naturally, this article can only scratch the surface of possible applications of mathematics, and in particular of time-frequency analysis, in music.

## 3 Modelling musical sound

If a sound signal is digitally recorded, we obtain a series of measurements of the corresponding air pressure. A priori, we have no knowledge about the signal, but we can have a look at the waveform, see Figure 2, where the waveform of a simple music signal is depicted. Immediately, the periodic nature of the signal strikes us. Typically, for music signals, we can assume at least sinusoidal components, corresponding to the main modes of vibration of the instrument producing the sound. in addition to these sinusoidal and

---

<sup>2</sup>See [2, Chapter 15] for a treatise of the musical communication chain and its modeling.

<sup>3</sup>MIDI (short for Musical Instrument Digital Interface) is a music industry standard communications protocol that lets MIDI instruments and sequencers (or computers running sequencer software) talk to each other to play and record music. More and more of the music heard every day is written with and played by MIDI sequencers, see[8].

nearly periodic components, however, we'll have to expect stochastic or more noisy parts of the signal. Hence, a common model for music signals  $s(t)$  is the following, see [11], for instance.

$$s(t) = \sum_{r=1}^R A_r(t) \cos[\theta_r(t)] + e(t), \quad (1)$$

where  $A_r(t)$  and  $\theta_r(t)$  are the instantaneous amplitude and phase of the  $r$ th sinusoid, respectively and  $e(t)$  is the noise component at time  $t$ . Now, in a next step, we can ask the obvious question which frequencies are contained in the given signal  $s(t)$ . The *musical* notion of pitch is of course related to the physical concept of frequency, however in a rather ambiguous manner. The pitch of a note is normally given by a bunch of related frequencies, corresponding to a fundamental and its overtones or harmonics. Often, but by no means always, the fundamental frequency corresponds to the perceived pitch. The frequencies in a given signal, however, correspond to the Fourier spectrum, given by the **Fourier transform**, which we define as:

$$\hat{s}(\omega) = \int_{\mathbb{R}^d} s(t) e^{-2\pi i \omega t} dt, \quad (2)$$

or, for discrete (digital) signals:

$$\hat{s}(k) = \sum_{m=0}^{M-1} s(m) e^{-\frac{2\pi i}{M} km} dt, \quad k = 0, \dots, M-1, \quad (3)$$

such that, by the inverse Fourier transform, the signal  $s(t)$  is again a sum of weighted components, in this representation a sum of pure sinusoids,  $e^{\frac{2\pi i}{M} km} = \cos(\frac{2\pi}{M} km) + i \sin(\frac{2\pi}{M} km)$ ,  $k = 0, \dots, M-1$ :

$$s(m) = \frac{1}{M} \sum_{k=0}^{M-1} \hat{s}(k) e^{\frac{2\pi i}{M} km} dt, \quad m = 0, \dots, M-1. \quad (4)$$

Now, have a look at Figure 2 again. The melody, which is shown on top of the page, played by a piano, was recorded and analysed. The second plot shows the Fourier transform  $\hat{s}(t)$  of the melody. Here, we can discern the distinct frequencies, corresponding to the sinusoid which represented by the peaks in the spectrum. However, this representation does not give any information about the localization of the sinusoids in time. We do not even know whether some peaks might contribute to separate notes at different time instants.<sup>4</sup>

It becomes clear that in the analysis of music signals we have to look for a representation which considers localisation in time *and* in frequency, which jointly yields information about *when* a note was played and *which* note was played. The theory providing tools for this kind of analysis is called *time-frequency analysis*, see [10, 12, 13, 14, 15], among many others. The most important tools in time-frequency analysis are the short-time Fourier transform and the spectrogram as well as *Gabor analysis*, [16], which can be understood as a generalisation of the short-time Fourier transform. In this article, we are going to discuss the basic principles of Gabor analysis, a method of time-frequency analysis which has an interpretation particularly close to the description of music.

The main idea in Gabor analysis is to expand the signal  $s(t)$  into a series of time-frequency concentrated building blocks (*atoms*), which are constructed from a single building block by translation (*time shift*) and modulation (*frequency shift*).

**Definition 1 (Time-frequency shifts)**  $T_k f(t) := f(t+k)$  is called translation operator or time shift.

$M_l f(t) := e^{-\frac{2\pi i l t}{L}}$ ,  $l \in \mathbb{Z}$  is called modulation operator or frequency shift.

The composition of these operators,  $M_l T_k$ , is a time-frequency shift operator.



Figure 1: *Melody*

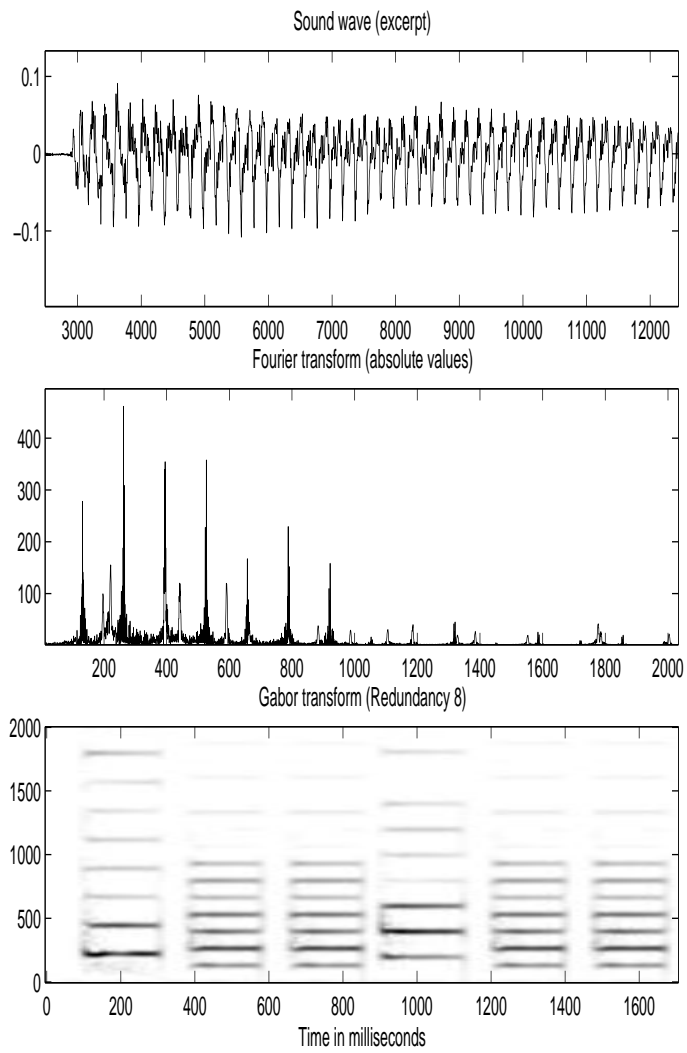


Figure 2: *Different representations of the melody*

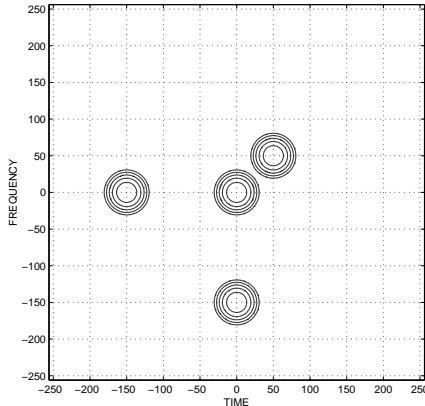


Figure 3: Time-frequency shifted versions of a Gauss function in the time-frequency plane

Figure 3 shows time-frequency shifted versions of a basic building block  $g$ , in this example a Gauss function, in the time-frequency plane. Now, the Gabor representation of  $s(t)$  is given as a weighted sum of time-frequency shifted versions of a basic building block  $g$  by

$$s(t) = \sum_{n,m \in \mathbb{Z}} c_{m,n} M_{mb} T_{ka} g(t), \quad (5)$$

where the coefficients  $c_{m,n}$  express the contribution of the building block at time  $na$  and frequency  $mb$ . As opposed to the wave-form representation in (1) or the representation as a sum of pure frequencies in (4), the Gabor representation can be intuitively understood as the composition of building blocks which correspond, like the notes in a music score, to certain events concentrated in specific areas of the time-frequency plane. The third plot in Figure 2 shows the Gabor coefficients of the signal  $s(t)$  corresponding to the melody in Figure 1. Although the harmonics, or overtones, are also present here, the melody can clearly be recognized.

**Remark 1** *The correspondence between written score music and the Gabor representation is only conceptual. In both cases certain events which are localised in time and frequency are depicted by a symbol in the time-frequency plane.<sup>5</sup> However, a musical note corresponds to a signal which, by itself, has to be represented by more than one Gabor atoms. The harmonics of the example signal shown in Figure 2, e.g., are sustained in time and hence have positive coefficients for more than one time-frequency concentrated building block. Recall that the basic building blocks have an elliptic shape, see Figure 3.*

## 4 Basic Gabor Theory

The definition of the Gabor representation as given in (5) immediately evokes many questions:

- **Which kind of function should be used as a basic building block  $g$ ?**

As a first choice, one might like to use a rectangular window as basic building block. This corresponds to cutting the given signal into pieces and looking at the Fourier transform of each of the pieces. This choice guarantees no loss of information and an easy reconstruction of the signal. On the other hand, the rectangular window is not even continuous and thus possesses a very slowly decaying Fourier transform. This means that the building blocks are not concentrated in frequency. Hence, we have to choose *smooth* windows. A typical example is the Gaussian window  $g(t) = e^{-\pi t^2} \in \mathbf{L}^2$ . It is invariant

<sup>4</sup>Harmonically closely related tones, for instance a quint apart, have a major part of their harmonics in common.

<sup>5</sup>Here, we interpret the music staves as a schematic time-frequency plane.

under Fourier transform and has minimal “extension in the time-frequency plane” in the sense that it achieves equality in Heisenberg’s uncertainty principle inequality.<sup>6</sup>

- **How are the parameters  $a$  and  $b$  chosen?**

The time-shift parameter  $a$  and the frequency-shift parameter  $b$  define the density of the lattice along which the window  $g$  is being shifted. Naturally, if the lattice is too coarse, information might be lost. This corresponds in a way to restricting the notes which are allowed in a piece of music to be no less than a major second apart or to have no smaller time value than a quaver. In fact, from theory we know that the product  $a \cdot b$  must be less than or equal to 1. The case  $a \cdot b = 1$  is called critical and it is, as its name suggests, a critical choice, leading to numerical instability and unsatisfying results in many applications. Hence, we should usually choose  $a$  and  $b$  to satisfy  $a \cdot b < 1$ . In application, we often use constants with  $a \cdot b = \frac{1}{4}$  or  $\frac{1}{8}$ , leading to a redundancy of 4 and 8, respectively. The Gabor coefficients of our example signal in Figure 2 were calculated with a redundancy of 8.

- **Can all possible signals be written as a Gabor series (5)?**

When thinking about natural signals such as music, the first idea is to think of them as continuous signals. Of course they can’t have infinite energy, so it is appropriate to assume they are members of  $\mathbf{L}^2(\mathbb{R})$ , the space of square-integrable functions  $\mathbb{R} \rightarrow \mathbb{C}$ . So here we ask the question under which conditions

$$M_{mb}T_{ka}g(t), \text{ where } m, k \in \mathbb{Z} \tag{6}$$

forms a complete system of building blocks. The theory of *frames* provides the right answer. Whenever there exist constants  $A, B > 0$  so that

$$A\|s\|^2 \leq \sum_{k,m \in \mathbb{Z}} |\langle s, M_{mb}T_{ka}g \rangle|^2 \leq B\|s\|^2 \text{ for all } s \in \mathbf{L}^2(\mathbb{R}),$$

the system (6) is called a *Gabor frame*. As an immediate consequence, any function  $s \in \mathbf{L}^2$  can be written as a Gabor series (5).

- **How can the coefficients be calculated?**

If the family (6) of time-frequency shifted atoms form a frame for a class of signals, there exists a second, dual frame  $M_{mb}T_{ka}\gamma(t)$ ,  $m, n \in \mathbb{Z}$ , such that a valid set of coefficients for (5) can be obtained as

$$c_{m,n}(s) = (\langle s, M_{mb}T_{ka}\gamma \rangle)_{m,n}. \tag{7}$$

Note that it is a very special and very useful property of Gabor frames, that the dual frame can be obtained from another basic building block  $\gamma$  by time-frequency shifts along the same lattice. Obviously, like this, only  $\gamma$  has to be calculated, which is numerically much more efficient than calculating each member of the dual frame separately.

Detailed information about all the above-mentioned topics can be found in [16] and [15]. A introduction to the connections between signal-processing of audio signals and Gabor analysis was given in [14].

---

<sup>6</sup>Heisenberg’s inequality states that for all functions  $f \in \mathbf{L}^2(\mathbb{R})$  and for all points  $(t_0, w_0)$  in the time-frequency plane

$$\|f\|_2^2 \leq 4\pi\|(t - t_0)f(t)\|_2\|(w - w_0)\hat{f}(w)\|_2$$

where equality is achieved only by functions of the form

$$g(t) = Ce^{2\pi itw_0}e^{-s(t-t_0)^2}, \quad C \in \mathbb{C}, s > 0$$

which are modulated and translated (i.e. time-frequency shifted) Gaussians.

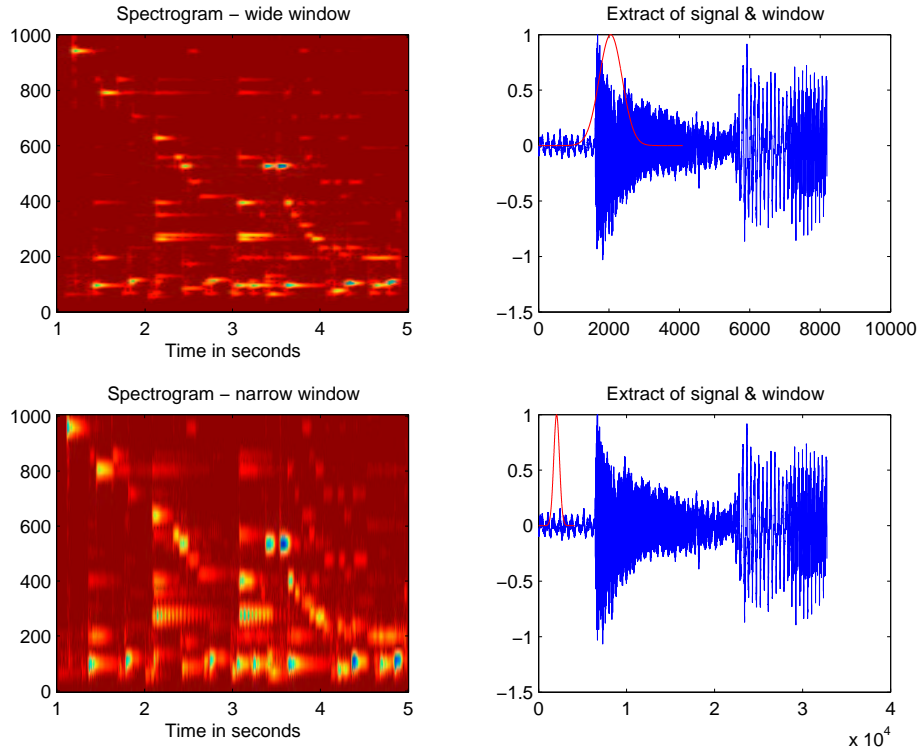


Figure 4: *Two windows and resulting STFT of music signal.*

## 5 Uncertainty

A function, e.g., a window function  $g$ , can not simultaneously be arbitrarily concentrated in time and frequency. This is the main statement of Heisenberg's uncertainty principle which was already mentioned in Section 4. This fact has important consequences for time-frequency analysis. The choice of the window  $g$  basically decides about the resolution quality that can be achieved in time and frequency. Coarsely speaking, wide windows yield a good concentration in frequency, whereas, obviously, narrow windows yield more precise information about localisation in time.

Figure 4 illustrates the effect of the usage of windows of different widths. The plots show the Gabor transform of a short segment from a piece performed by a piano, a double-bass and a drummer. The signal was analysed with two Gaussian windows of different widths. From the resulting spectrograms it becomes obvious that especially in low frequency regions, the wider window yields a far better representation of the signal components. Signal components which are close in frequency become blurred if the analysing window is too narrow hence yielding a bad frequency resolution. Note that the two Gabor systems generating the results have the *same* redundancy. On the other hand, in high frequency regions, we are interested in precise time resolution, hence, windows with a small essential support in the time-domain should be used. However, classical time-frequency analysis with redundant systems is restricted by using systems with the *same* time-frequency resolution for the whole time-frequency plane. Intuitively, what we are looking for is a generalisation of a tiling of the time-frequency plane allowing for more flexibility. In the case of music signals, for example, transients are important for several reasons. They give important cues for onset timing, and they carry information about instrument timbre (often instrument perception hinges on the perception of transients.) As another example, in low-frequency regions, very fine frequency resolution is required, because notes in this region lay the harmonic basis, musically speaking. This is especially true in music such as Jazz, where the function of the bass determines the harmonic structure and function of the whole piece.

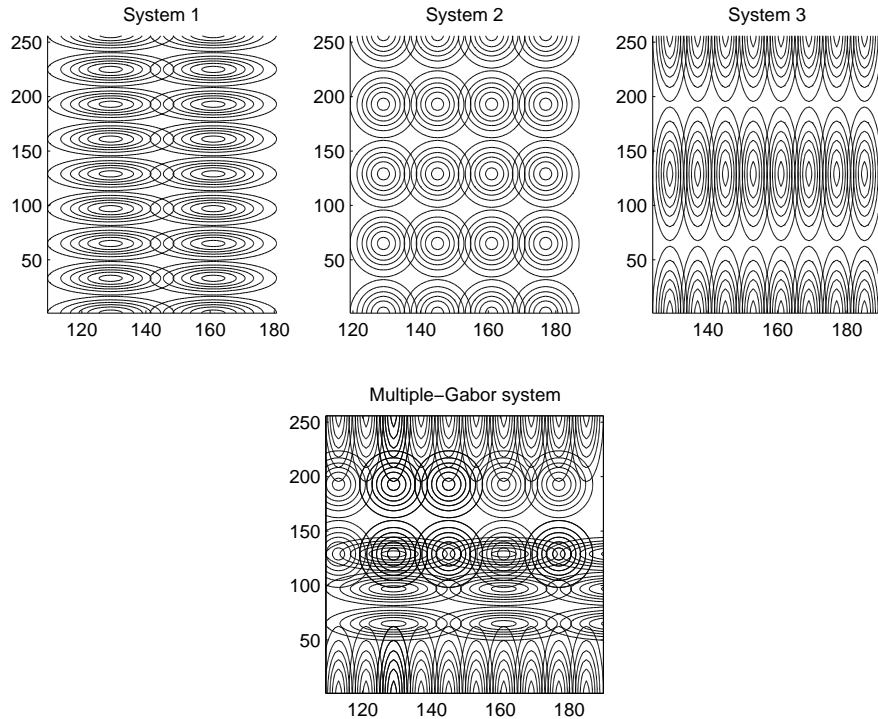


Figure 5: *Constructing multiple-Gabor frames*

## 5.1 Multiple Gabor Frames

A new approach allowing to locally use Gabor systems which are well suited to the characteristics of a given signal or class of signals has been suggested in [17]. It has been shown that frames can be constructed by using sets of building blocks corresponding to the frequency region they cover. These new frames will be called *multiple-Gabor frames*, emphasising that various different Gabor frames are used to generate the new system. Figure 5 schematically illustrates the construction of a multiple-Gabor frame from various distinct Gabor frames. The discussion of technical details of multiple Gabor frames is beyond the scope of this contribution and will be found in future work.

## 6 Time-Variant Filtering

Consider the task of extracting signal components which possess certain properties, e.g., belonging to a given instrument, from a complex audio signal. This obviously is not only a non-trivial problem, but also a highly nonlinear one, if the dependence on the specific signal under consideration is observed. The process of extracting parts of a given signal can be described by an operator  $\mathbf{G}(f)$ :

$$f \longrightarrow \mathbf{G}(f)f,$$

where  $\mathbf{G}(f)$  denotes the time-varying filtering system depending on the properties of the signal  $f$  under inspection. Ignoring the dependance on  $f$ , however, we can fix the amount of filtering in time and frequency and  $\mathbf{G}$  becomes a linear operator. Here, we want to study the properties of a specific class of time-varying filter operators, the so-called Gabor multipliers. A similar approach has been studied in [18].

**Definition 2 (Gabor multiplier)** *Let  $g_1$  and  $g_2$  be two functions in  $S_0(\mathbb{R}^d)$ ,  $\Lambda \subset \mathbb{R}^d \times \widehat{\mathbb{R}^d}$  a TF-lattice and  $\mathbf{m} = (m(\lambda))_{\lambda \in \Lambda}$  a bounded, complex-valued sequence in  $\ell^\infty(\Lambda)$ . Then the **Gabor multiplier** associated*

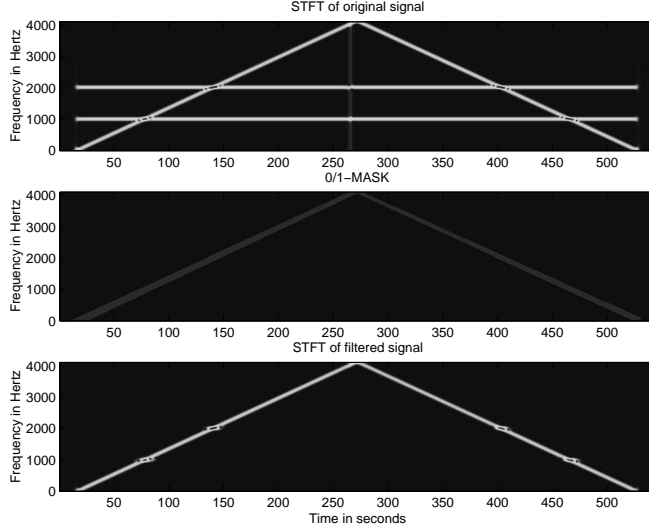


Figure 6: *Gabor multiplier applied to synthesized signal*

Figure 7: *Gabor multiplier applied to real signal*

to the triple  $(g_1, g_2, \Lambda)$  is given as

$$G_{\mathbf{m}}(f) = \mathbf{G}_{g_1, g_2, \Lambda, \mathbf{m}}(f) = \sum_{\lambda \in \Lambda} m(\lambda) \langle f, \pi(\lambda)g_1 \rangle \pi(\lambda)g_2. \quad (8)$$

The subscripts  $g_1, g_2$  and  $\Lambda$  will be omitted if they are not crucial for the discussion. If  $g_1 = g_2$ , we write  $\mathbf{G}_{g_1, g_2, \Lambda, \mathbf{m}} = \mathbf{G}_{g, \Lambda, \mathbf{m}}$ . Unless otherwise stated,  $g_1 = g_2$  will be assumed for most of the rest of this chapter. Hence,  $f \mapsto \langle f, \pi(\lambda)g \rangle \pi(\lambda)g$  are the projections (orthogonal projections if  $\|g\|_2 = 1$ ) onto the one-dimensional subspaces of  $\mathbf{L}^2$  generated by the time-frequency shifted versions of  $g$ .

Figure 6 shows an example for the application of a Gabor multiplier. A signal comprising a chirp, two sinusoids at distinct frequencies and a Dirac impulse has been synthesized. The short-time Fourier transform (STFT) of the composite signal is shown in the first plot. The second plot shows the multiplier  $\mathbf{m}$ , which is a 0/1-mask in this example. The chirp has been filtered out, which can be seen in the third plot. This example shows that with Gabor multipliers time-varying filtering can be realized.

A second example, obtained from a music signal, is shown in Figure 7. The signal is a short segment from a piece performed by a double-bass and drums. The first plot shows the STFT of the signal. The note played by the bass is well recognizable with its harmonics in the low frequency region. The second plot shows a 0/1-mask which was used to extract the bass note. The STFT of the resulting signal is shown in the last plot.<sup>7</sup>Clearly, the noisy components, originating mainly from the drums, have been eliminated. It is hard to evaluate, however, the quality of the extracted signal component. Human perception must be seen as the ultimate criterion in audio signal processing, as differences between signal which may seem crucial in a mathematical sense, can be inaudible. For the given example (Figure 7), the result of the extraction seems satisfying to the listener, i.e., the instrument, the pitch and the quality of the extracted note are perfectly recognizable.

<sup>7</sup>Here, the mask was found by visual inspection. However, we hope that for an automatized extraction routine, image recognition tools might be applied to find corresponding signal components.

The close examination of the properties of Gabor multipliers is motivated by the desire to quantify as accurately as possible the dependence of the action of  $\mathbf{G}_m$  on the shape of the windows, the lattice parameter and the symbol  $\mathbf{m}$  itself.

## References

- [1] M. Levine. *The Jazz Theory Book*. Sher Music Co., Petaluma, CA, 1995.
- [2] G. Assayag, H. G. Feichtinger, and J. F. Rodrigues, editors. *Mathematics and Music*. Diderot Forum, Lisbon-Paris-Vienna. Springer, 2002.
- [3] S. Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001.
- [4] A.T. Cemgil, P. Desain, and B. Kappen. Rhythm quantization for transcription. *Computer Music Journal*, 24:2:60–76, Summer 2000.
- [5] P.J. Walmsley, S.J. Godsill, and P.J.W. Rayner. Bayesian graphical models for polyphonic pitch tracking. In H. G. Feichtinger and M. Dörfler, editors, *Proceedings of the Diderot Forum on Mathematics and Music: Computational and Mathematical Methods in Music*, pages 353–366, Vienna, 1999.
- [6] D. Byrd and T. Crawford. Problems of music information retrieval in the real world. *Information Processing and Management*, 38(2):249–272, March 2002.
- [7] J. Fitch and W. Shabana. A wavelet-based pitch detector for musical signals. In Jan Tro and Mikael Larsson, editors, *Proceedings of DAFx99*. Department of Telecommunications, Acoustics Group, Norwegian University of Science and Technology, December 1999.
- [8] C. Roads. *The Computer Music Tutorial*. Computer Science. MIT press, cambridge, MA, 1996.
- [9] E. Cambouropoulos. From MIDI to traditional musical notation. In *Proceedings of the AAAI Workshop on Artificial Intelligence and Music: Towards Formal Models for Composition, Performance and Analysis*.
- [10] J. Jeong and W.J. Williams. *Time-Varying Filtering and Signal Synthesis, Time-Frequency Signal Analysis Methods and Applications*, chapter 17, pages 389–405. Longman Cheshire, Wiley Halsted Press, Boualem Boashash edition, 1992.
- [11] X. Serra. Musical sound modeling. In A. Picciali C. Roads, S.T. Pope and G. De Poli, editors, *Musical Signal Processing*. Swets and Zeitlinger, Belgium, 1997.
- [12] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, USA, 1998.
- [13] S. Qian and D. Chen. *Joint Time-Frequency Analysis: Method and Application*. Prentice Hall, Englewood Cliffs, NJ, 1996.
- [14] M. Dörfler. Time-frequency analysis for music signals: A mathematical approach. *Journal of New Music Research*, 30(1):3–12, 2001.
- [15] K. Gröchenig. *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston, 2001.

- [16] H. G. Feichtinger and H. G. Strohmer, editors. *Gabor Analysis and Algorithms: Theory and Applications*. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston, 1998.
- [17] P. J. Wolfe, M. Dörfler, and S. J. Godsill. Multi-Gabor dictionaries for audio time-frequency analysis. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 43–46, Mohonk, NY, 2001.
- [18] S. Raz and S. Farkash. Time variant filtering via the Gabor representation. In L. Torres, E. Masgrau, and M.A. Lagunas, editors, *Proceedings of the EUSIPCO-90, European Signal Processing Conference*, pages 509 – 512, Barcelona, Spain, Sept. 1990.