Arbitrary Phase Vocoders by means of Warping

Gianpaolo Evangelista, Monika Dörfler and Ewa Matusiak*

(Received March 3, 2010)

1 Introduction

Time-frequency representations play a central role in the analysis, synthesis, coding and processing of sound signals. In this context, the most commonly used representation is the Phase Vocoder [Dol86], which derives from sampling the Short-Time Fourier Transform (STFT), usually at uniformly spaced time and frequency sampling points in the time-frequency plane. This results in a covering or tessellation of the time-frequency plane with *atoms* or *logons* of equal duration and uniform nominal bandwidth. However, non-uniform resolution is desirable in several applications. For example, the analysis and synthesis frequency bands can be adapted to a perceptual scale, achieving clear advantages in synthesis and coding due to the direct psycho-acoustic relevance of each component. In synthesis-by-analysis schemes, the frequency bands can be adapted to characteristics of the signal suggested, for example, by the frequencies of the partials of the tones, which in many instruments, such as the piano in the low register or percussions, are not harmonically related. Another example is the representation of transient and stationary parts of sounds: close to the wideband onsets of sounds one definitely desires to allocate finer time resolution while in stationary segments, where the innovation rate of the signal is low, one would like to allocate coarser time resolution at the advantage of a finer frequency resolution, e.g., to be able to precisely track vibrato or glissando.

It is rather easy to construct non-uniform time windows that overlap-add to 1. For example, in the case of the half-length overlapping windows in Fig. 1, one can symmetrically alter the two overlapping segments of the windows so that these continue to overlap-add to 1, leading to windows with asymmetrical trailing and leading portions, as shown in Fig. 2. The necessary alteration is easily achieved by proportionally stretching or shrinking the overlapping segments. Equivalently, the time axis or, rather, its significance with respect to the signal and windows, can always be remapped by means of a one-to-one map, rescheduling time instants to other time instants in a process

*M. Dörfler and E. Matusiak were supported by the WWTF project *Audiominer* (MA09-24).

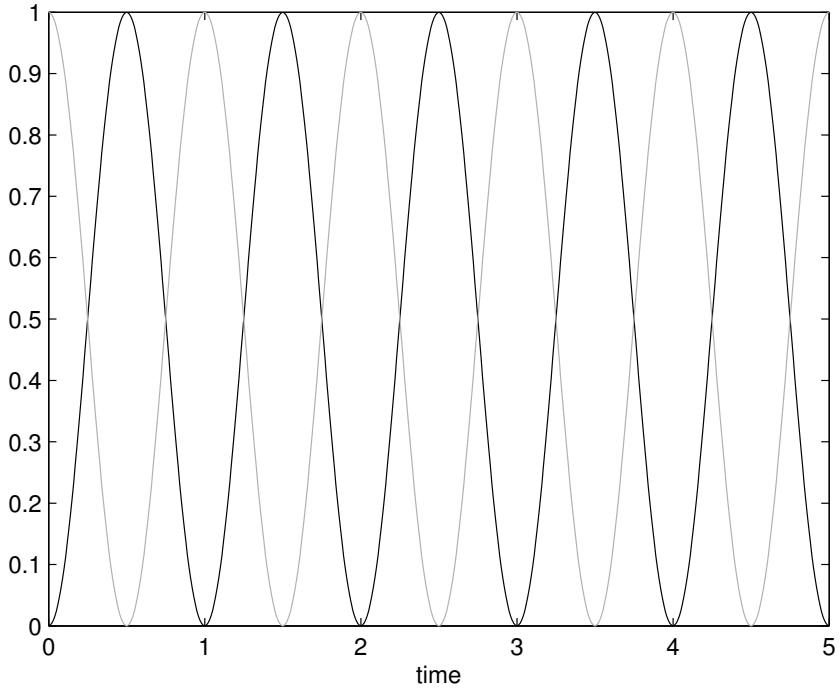

Fig. 1: Uniform time windows overlapping-add to 1.

known as time warping. In the previous example, this results in a piecewise linear time warping characteristic shown in Fig. 3. In general, the time warping map can be any increasing function of the real axis. Time adaptation can be driven by transient detection and can be smoothly performed according to a prescribed curve.

Similarly, non-uniform frequency bands can always be thought of as obtained from uniform bands through a frequency map, i.e. a monotonically increasing function remapping the frequency axis, as shown in Fig. 4. In certain cases, e.g. critical bands, the frequency map is given by experimentally fitted curves. In other cases, such as in the vibration of stiff strings or bars, the frequency map is derived from wave dispersion characteristics. Just like light travels with frequency dependent velocity in certain media, so does mechanical displacement propagate in thick strings or rods. Often the map is only specified at a finite number of points; a continuous curve can always be obtained by interpolation of the fitting points in an arbitrary fashion; therefore, the map can be assumed to be smooth. The application of a frequency map is known in filter design as frequency warping.

The process of adapting time and frequency resolutions of the analysis-synthesis scheme of the Vocoder can be thought of as a deformation of the time-frequency plane according to signal and/or perception. The atoms of the resulting time-frequency

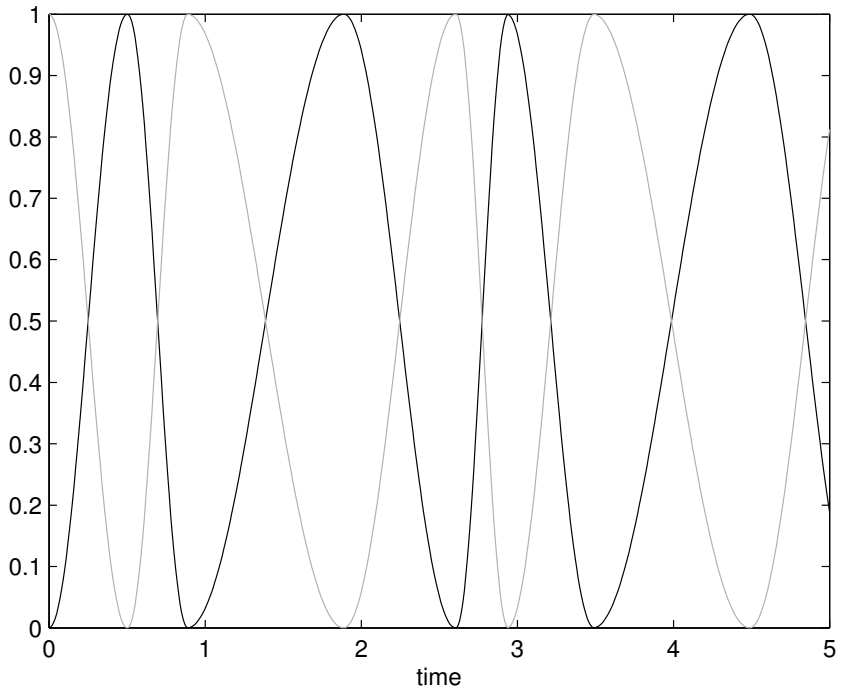

Fig. 2: Example of non-uniform time windows overlapping-add to 1.

representation differ in (essential) duration and bandwidth. Since time and frequency are not independent variables, care must be taken in that frequency warping also affects time just as time warping also affects frequency.

Moreover, especially in sound processing and transformation applications, one usually requires perfect reconstruction. However, if one operates with bounded and invertible deformations of the time-frequency plane, variable time-frequency atoms can be obtained from uniform atoms completely representing the signal, which automatically provide perfect reconstruction.

In this paper, based on recently developed results [Vel+11; EC07; EDM12], we address the problem of building perfect reconstruction structures for the time-frequency representation of signals that allow for arbitrary selection of bands specified according to a frequency map. Mathematically this amounts to constructing flexible frames that allow for parametric selection of the frequency bands of their atoms. Frames are sets of functions that completely specify any signal. They are overcomplete, just like using three or more coordinates to represent a 2D vector (but notice that the space of signals has infinite dimensions).

The paper is organized as follows. In this section we continue by reviewing the STFT and its sampling, illustrating the basics of Gabor and warped Gabor frames and their

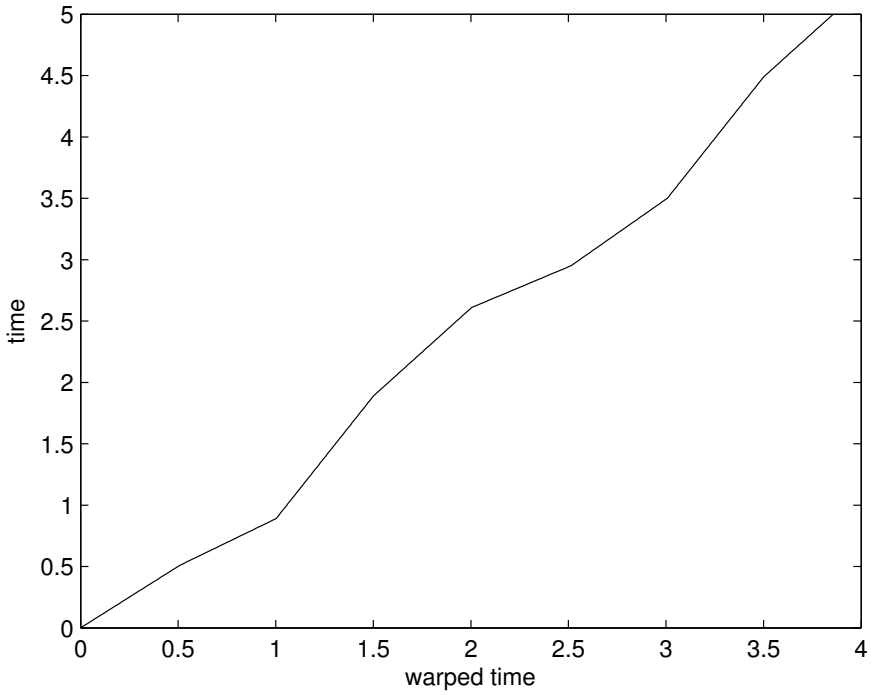

Fig. 3: Piecewise linear time warping map.

associated dual frames to be employed in the reconstruction of the signal. In Section 2 we introduce the new concept of nonuniform Gabor frames and their dual frames designed by means of frequency warping. In Section 3 we illustrate some applications of the methods and in Section 4 we draw our conclusions.

1.1 The Three Souls of the Short-Time Fourier Transform

One way to characterize a sound signal in time-frequency is to take short and possibly overlapping time segments of the signal and analyze them in frequency. The Short-Time Fourier Transform (STFT) is obtained by sliding in time a window over the signal and by taking the Fourier transform of each windowed portion of the signal. In general, the window does not need to have finite duration and could be any function with a certain time decay. For a continuous-time signal s and a real window g the STFT takes on the form of the following integral:

$$S_g(\tau,\nu) = [\mathbf{Q}_g s](\tau,\nu) = \int s(t)g(t-\tau)e^{-j2\pi\nu(t-\tau)}dt, \tag{1}$$

where $j = \sqrt{-1}$ is the imaginary symbol. The *STFT operator* $\mathbf{Q}_g$ brings the time signal s to its description S_g in a time-frequency plane, where τ represents time and

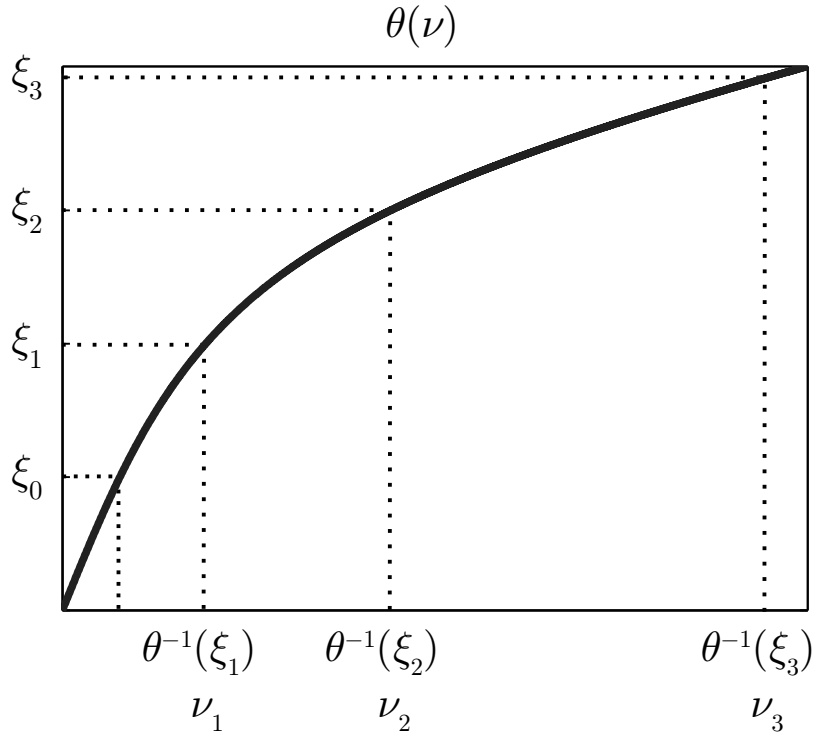

Fig. 4: Example of frequency warping map, mapping uniform bands (ordinate) into nonuniform bands (abscissa).

ν frequency. With respect to the classical definition, in (1) we have introduced the irrelevant but useful phase factor $e^{j2\pi\nu\tau}$. With this phase factor the Fourier transform contained in the STFT is reset at the center of the window, i.e. the instant $t = \tau$, for any τ, rather than being referred to the time $t = 0$.

The STFT is equivalently obtained by filtering the signal s with a continuum of bandpass filters whose impulse responses are obtained by modulating the time reversed window $\breve{g}(t) = g(-t)$, i.e.

$$S_g(\tau,\nu) = s(\tau) * \breve{g}_\nu(\tau) = \int s(t)\breve{g}(\tau - t)e^{j2\pi\nu(\tau-t)}dt, \tag{2}$$

where the symbol $*$ denotes convolution and

$$\breve{g}_\nu(t) = \mathbf{M}_\nu \breve{g}(t) = \breve{g}(t)e^{j2\pi\nu t}. \tag{3}$$

The symbol $\mathbf{M}_\nu$ is referred to as the *modulation operator*, which multiplies a function by a complex sinusoid of frequency ν.

Convolution in time corresponds to multiplication in the frequency domain. Therefore, denoting by $\hat{S}_g(\varphi,\nu)$ the Fourier transform of the STFT $S_g(\tau,\nu)$ with respect to

the first argument τ, we have:

$$\hat{S}_g(\varphi, \nu) = \hat{s}(\varphi)\overline{\hat{g}(\varphi - \nu)}, \tag{4}$$

where we have used the facts that the Fourier transform of the time-reversed window equals the complex conjugate of the Fourier transform of the window and that modulation in time corresponds to frequency shift of the Fourier transform. Thus, an alternative way of computing the STFT is to take the Fourier transform of the signal, multiply it for the conjugate Fourier transform of the modulated window and then compute the inverse Fourier transform. Most often the window g is chosen to be symmetric, i.e. $g(t) = g(-t)$, in which case g and $\breve{g}$ coincide and $\hat{g}$ is a real function.

Defining the *time-shift operator* $\mathbf{T}_\tau$ such that $\mathbf{T}_\tau f(t) = f(t - \tau)$, it is possible to write (1) as follows:

$$S_g(\tau, \nu) = \int s(t)\overline{\mathbf{T}_\tau \mathbf{M}_\nu g(t)}dt, \tag{5}$$

where the overbar symbol denotes complex conjugation: $\overline{a + jb} = a - jb$. Defining the scalar product in the space of finite energy signals (the space $L^2(\mathbb{R})$) as follows:

$$\langle f, g \rangle = \int f(t)\overline{g(t)}dt, \tag{6}$$

from (5) one can see that

$$S_g(\tau, \nu) = \langle s, \mathbf{T}_\tau \mathbf{M}_\nu g \rangle , \tag{7}$$

which makes it possible to interpret the STFT as the scalar product of the signal with the time-shifted modulated versions $\mathbf{T}_\tau \mathbf{M}_\nu g$ of the window g.

Thus, the STFT has at least three souls: it is the Fourier transform of windowed portions of the signal, it is the convolution of the signal with modulated versions of the window and it is the scalar product of the signal with time-shifted modulated versions of the window. These three souls can be exploited in the applications, for the computation of the STFT and for its sampling, as shown in the next section.

1.2 Sampling the Short-Time Fourier Transform: Gabor Frames

The STFT leads to a 2D representation of a 1D signal. Due to the addition of an extra dimension, the representation is very likely to be redundant. For example, if the window is bandlimited, the output of each filter whose impulse response is a modulated version of the window is bandlimited; therefore it can be sampled in time. As a result, the STFT can be reconstructed from its values at a discrete set of time instants τ_n. Similarly, if the window has finite length, each windowed portion of the signal has finite duration; therefore the STFT can be reconstructed from its samples at a discrete set of frequency points. Several other possibilities are available.

Sampling does not necessarily eliminate redundancy, it may just reduce it. The redundancy of the transform also implies that not all 2D functions are valid STFT of a signal: the values of the STFT are interdependent and cannot be assigned arbitrarily.

Similar considerations apply to the problem of reconstructing the signal itself from the knowledge of its STFT. In fact, since the STFT is the Fourier transform of windowed segments of the signal, taking the inverse Fourier transform of the STFT will leave us with windowed portions of the signal, a lot of them! Even if time is sampled in the time-frequency plane, reconstruction of the signal is still possible with minor requirements on the window provided that the time samples are sufficiently dense.

In more general terms, one would like to sample both time and frequency in the time-frequency plane without losing information. The scalar product soul (7) of the STFT comes in handy in this type of questions. Given the set of projection coefficients

$$S_g(\tau_n, \nu_q) = \left\langle s, \mathbf{T}_{\tau_n}\mathbf{M}_{\nu_q} g \right\rangle, \qquad n, q \in \mathbb{Z}, \tag{8}$$

obtained by sampling the STFT on a grid of points $\{(\tau_n, \nu_q)\,|n, q \in \mathbb{Z}\}$ in the time-frequency plane, one would like to know if it is possible to reconstruct the signal. Which are the classes of windows that guarantee perfect reconstruction? Which are feasible sampling grids? Which is the reconstruction algorithm? Given that the set of functions $\left\{\mathbf{T}_{\tau_n}\mathbf{M}_{\nu_q} g\right\}_{n,q\in\mathbb{Z}}$ is not generally orthogonal and it might not even be complete, these questions raise a mathematical problem, which we will first explore in finite dimensional spaces of vectors by way of an example drawn from linear algebra.

1.2.1 Redundant Representations in Finite Dimensions and Duality

In finite dimensional euclidean vector spaces one can reconstruct any vector from its projections over a sufficient number – at least equal to the dimension of the space – of well chosen directions (for example the Cartesian axes). The projection coefficients are given by the scalar products of the vector over the vectors describing the directions. As a matter of fact, reconstruction is possible even if the vectors are not orthogonal (basis and dual basis or biorthogonal bases) and even if, obviously, the set of directions is redundant (for example three suitable directions in a 2D space).

Consider the example shown in Fig. 5(a), where one would like to specify a two-dimensional vector $\mathbf{v}$ in terms of the scalar products taken along the three directions ψ_1, ψ_2 and ψ_3. Clearly, any two of these directions are linearly independent so that any subset of two directions would suffice to specify $\mathbf{v}$. In particular, the set $\{\psi_1, \psi_2\}$ forms an orthogonal basis so that one could immediately write $\mathbf{v} = c_1\psi_1 + c_2\psi_2$, where $c_k = \langle \mathbf{v}, \psi_k \rangle = \psi_k^T \mathbf{v}$ are the scalar products of the vector along the first two directions, where the symbol ^{T} denote transposition. Writing the components of all vectors in terms of the Cartesian coordinate system given by the axes ψ_1 and ψ_2, one can put this trivial result in matrix-vector form:

$$\psi_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; \psi_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \; \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \psi_1^T \\ \psi_2^T \end{bmatrix} \mathbf{v} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{v}. \tag{9}$$

Thus, $\mathbf{c} = \mathbf{H}\mathbf{v}$, where, in this case, $\mathbf{H}$ is the 2x2 identity matrix $\mathbf{I}_2$. Clearly, the original vector $\mathbf{v}$ can be recovered by matrix inversion:

$$\mathbf{v} = \mathbf{H}^{-1}\mathbf{c} = \begin{bmatrix} \tilde{\psi}_1^{(o)} & \tilde{\psi}_2^{(o)} \end{bmatrix} \mathbf{c} = c_1\tilde{\psi}_1^{(o)} + c_2\tilde{\psi}_2^{(o)}, \tag{10}$$

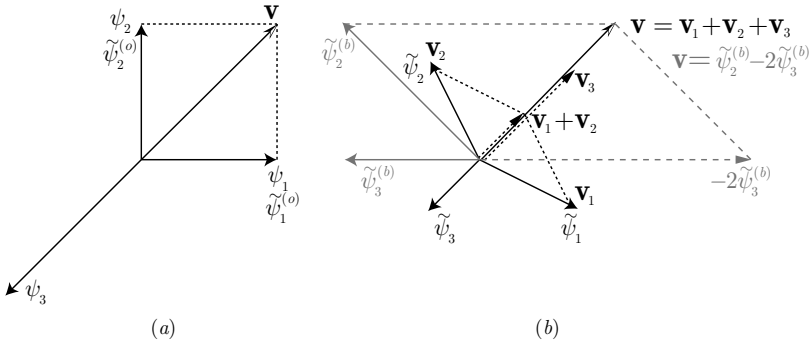

Fig. 5: Example of redundant representation in a 2D vector space; (a) Original vector and three possible directions for projection and (b) Reconstruction of the vector with dual basis (gray) and with dual frame (black)

where we have denoted by $\tilde{\psi}_1^{(o)}$ and $\tilde{\psi}_2^{(o)}$ the columns of $\mathbf{H}^{-1}$. However, in this case, $\mathbf{H}^{-1} = \mathbf{H} = \mathbf{H}^T = \mathbf{I}_2$ so that the directions for extracting the analysis coefficients $\mathbf{c}$ and those to recover the vector $\mathbf{v}$ from these coincide: $\tilde{\psi}_1^{(o)} = \psi_1$ and $\tilde{\psi}_2^{(o)} = \psi_2$ for orthogonal bases.

Consider as representative, instead, the scalar products of $\mathbf{v}$ along the directions $\{\psi_2, \psi_3\}$. These two directions are not orthogonal but still linearly independent. In matrix-vector form we have:

$$\psi_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ \psi_3 = \begin{bmatrix} -1 \\ -1 \end{bmatrix},\ \mathbf{c} = \begin{bmatrix} c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \psi_2^T \\ \psi_3^T \end{bmatrix} \mathbf{v} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} \mathbf{v} \tag{11}$$

where, as before, we wrote the components of all vectors in terms of the Cartesian coordinate system given by the axes ψ_1 and ψ_2. Here again we have $\mathbf{c} = \mathbf{H}\mathbf{v}$, but $\mathbf{H}$ is not the 2x2 identity matrix. The original vector $\mathbf{v}$ can still be recovered by matrix inversion: $\mathbf{v} = \mathbf{H}^{-1}\mathbf{c}$, but, in this case, $\mathbf{H}^{-1} \neq \mathbf{H}^T$. In fact,

$$\mathbf{H}^{-1} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}^{-1} = \begin{bmatrix} -1 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \tilde{\psi}_2^{(b)} & \tilde{\psi}_3^{(b)} \end{bmatrix}. \tag{12}$$

Thus,

$$\mathbf{v} = c_2\tilde{\psi}_2^{(b)} + c_3\tilde{\psi}_3^{(b)} \neq c_2\psi_2 + c_3\psi_3 \tag{13}$$

and one can refer the two sets $\{\psi_2, \psi_3\}$ and $\{\tilde{\psi}_2^{(b)}, \tilde{\psi}_3^{(b)}\}$, respectively, as *basis* and *dual basis*. The analysis coefficients are obtained by taking the scalar products of the vector along the basis elements and the reconstruction is possible via expansion over the dual basis. The dual basis and the relative reconstruction of the vector $\mathbf{v} = [1, 1]^T$, where from (11) $c_2 = 1$ and $c_3 = -2$, are represented by the gray lines and vectors in Fig. 5(b).

Since they are derived, respectively, from the rows of the matrix $\mathbf{H}$ and from the columns of its inverse, i.e.,

$$\begin{bmatrix} \tilde{\psi}_2^{(b)} & \tilde{\psi}_3^{(b)} \end{bmatrix} \begin{bmatrix} \psi_2^T \\ \psi_3^T \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{14}$$

it is easy to see that basis and dual basis satisfy the cross-orthogonality property: $\left\langle \psi_n, \tilde{\psi}_m^{(b)} \right\rangle = \delta_{n,m}$, $n, m = 2, 3$, where $\delta_{n,n} = 1$ and $\delta_{n,m} = 0$ for $n \neq m$; for this reason they are called *biorthogonal bases*. They form bases for the same vector space and their role can even be interchanged while preserving perfect reconstruction.

Consider now the set of all three directions $\{\psi_1, \psi_2, \psi_3\}$ shown in Fig. 5(a). Taking the scalar products of the vector $\mathbf{v}$ over these directions yields:

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \psi_1^T \\ \psi_2^T \\ \psi_3^T \end{bmatrix} \mathbf{v} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} \mathbf{v} = \mathbf{H}\mathbf{v}. \tag{15}$$

In this case, the matrix $\mathbf{H}$ is 3x2 and, not being square, it does not have an inverse in the usual sense. However, provided that the determinant of the 2x2 matrix $\mathbf{H}^T\mathbf{H}$ is nonzero, one can find a 2x3 matrix $\tilde{\mathbf{H}}$ such that $\tilde{\mathbf{H}}\mathbf{H} = \mathbf{I}_2$. In fact, one can let $\tilde{\mathbf{H}}$ be the left pseudoinverse of the matrix $\mathbf{H}$, defined as $\tilde{\mathbf{H}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$. We have: $\tilde{\mathbf{H}}\mathbf{H} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{H} = \mathbf{I}_2$. Therefore, the elements to reconstruct $\mathbf{v}$ from the coefficients $\mathbf{c}$ can be identified as the columns of the pseudoinverse matrix $\tilde{\mathbf{H}}$. In our example we have:

$$\begin{bmatrix} \tilde{\psi}_1 & \tilde{\psi}_2 & \tilde{\psi}_3 \end{bmatrix} = \tilde{\mathbf{H}} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} = \tfrac{1}{3} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \end{bmatrix}, \tag{16}$$

which yields the components in the Cartesian basis $\{\psi_1, \psi_2\}$ of the three directions $\{\tilde{\psi}_1, \tilde{\psi}_2, \tilde{\psi}_3\}$ shown in Fig. 5(b). In the same figure, the original vector $\mathbf{v} = [1, 1]^T$ is constructed as the sum of the vectors $\mathbf{v}_k = c_k\tilde{\psi}_k$, $k = 1, 2, 3$, where, from (15), the coefficients are $c_1 = 1$, $c_2 = 1$ and $c_2 = -2$.

The redundant set of directions $\{\psi_1, \psi_2, \psi_3\}$ used to compute the analysis coefficients in (15) constitutes what is referred to as a *frame* for the vector space and the set $\{\tilde{\psi}_1, \tilde{\psi}_2, \tilde{\psi}_3\}$ generated in (16) for the reconstruction constitutes its *dual frame*. Frames and dual frames will be more formally introduced in the next section.

Clearly, the frame analysis coefficients c_k are not unique. In fact, the same vector in the figure could be expressed as well by the coefficients $c_1 = 1 + x$, $c_2 = 1 + x$ and $c_2 = -2 + x$ for any x. We conclude that if the number of representative directions is larger than the dimension of the space, the representation of the vector is not unique as the directions are not independent (in a 2D space with three representative directions one can always express at least one of the directions in terms of the other two). Moreover, even the reconstruction algorithm is not unique, i.e., one does not need to construct the dual frame in order to reconstruct the vector, since, as we have seen previously, one could for example discard one of the scalar products and use only c_1 and c_2 to reconstruct the vector using the orthogonal basis as in (10) or use c_2 and c_3 only to reconstruct the vector using the dual biorthogonal basis as in (13), just to name two alternative ways but the possibilities are infinite.

Even in a 2D space with available projections in multiple directions one needs to be careful in the selection of the directions. While the vector $\mathbf{v}$ in Fig. 5 could be expressed

in terms of the direction ψ_3 only, all vectors not aligned with $\mathbf{v}$ require additional representative elements. If all the representative directions are aligned, not all vectors can be described by their projections. In fact, all components orthogonal to the aligned directions come out of the picture: any vector orthogonal to the aligned directions will have zero projections, all of them! To avoid this problem one can enforce the condition that no nonzero vector in the space has zero projection on all the selected directions or, what is the same, that the sum of the magnitudes of the scalar products of any vector with all the representative directions is not smaller than a fraction of the norm of the vector. In the finite dimensional case, this is equivalent to enforce that the projection matrix $\mathbf{H}$ has at least rank equal to the dimensionality of the space, which in turn guarantees that the matrix $\mathbf{S} = \mathbf{H}^T\mathbf{H}$ is not singular so that the left psuedoinverse $\tilde{\mathbf{H}}$ exists. Note that in the scalar vector notation, for our three-element frame one can write:

$$\mathbf{S}\mathbf{v} = \mathbf{H}^T\mathbf{H}\mathbf{v} = \begin{bmatrix} \psi_1 & \psi_2 & \psi_3 \end{bmatrix} \begin{bmatrix} \psi_1^T \\ \psi_2^T \\ \psi_3^T \end{bmatrix} \mathbf{v} = \sum_{k=1}^{3} \langle \mathbf{v}, \psi_k \rangle \psi_k \tag{17}$$

so that the existence of the pseudoinverse is linked to the invertibility of an operator $\mathbf{S}$ acting on the vectors of the space. This operator is referred to as the *frame operator*. In terms of this operator, (16) becomes:

$$\begin{bmatrix} \tilde{\psi}_1 & \tilde{\psi}_2 & \tilde{\psi}_3 \end{bmatrix} = \tilde{\mathbf{H}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T = \mathbf{S}^{-1}\mathbf{H}^T = \mathbf{S}^{-1}\begin{bmatrix} \psi_1 & \psi_2 & \psi_3 \end{bmatrix}. \tag{18}$$

Thus, each element of the dual frame can be obtained by applying the inverse of the operator $\mathbf{S}$ to each frame element.

The space of finite energy signals has infinite dimension. Signal expansions, e.g., Fourier series for a time-limited signal, are computed by taking signal projections, i.e., scalar products of the signal over a countable number of representative elements (e.g., harmonic complex exponentials). When the representative elements are not orthogonal according to the scalar product given in (6), projections over elements other than the representative elements is required. In this case the elements and their duals work in symbiosis in order to analyze and synthesize the signal. For example, for a non-orthogonal basis the right algorithm is projection on the basis and expansion over the dual basis. However, when the representation elements are redundant, then the algorithm is not unique since there are various sets of coefficients and even various sets of synthesis elements that can equally represent the signal. For the same reason, the expansion coefficients are not independent, which means that not all finite energy sequences are valid coefficients to represent any signal in the space. The method of projections on a set of elements and expansion over a dual set of elements still works but one must be aware that this is just one of the infinite number of ways to obtain the right expansion coefficients and a perfect reconstruction scheme for the signal [Mal98].

1.2.2 Gabor Frames

In the STFT case, we are interested in representative elements of the type

$$\mathcal{G}^{\sharp}(g, \tau_n, \nu_q) = \{\mathbf{T}_{\tau_n}\mathbf{M}_{\nu_q} g : n, q \in \mathbb{Z}\}, \tag{19}$$

where g is a non-zero window function, which we refer to as the *generalized Gabor system*. Conventional Gabor systems [Gab46] stem from uniform sampling of the STFT on a time-frequency grid of points or *lattice* $\{(na, qb)\,|n, q \in \mathbb{Z}\}$, where a and b are parameters controlling the density of the grid points. In the classical definition, time-shift and modulation operators are interchanged but this can always be achieved with irrelevant phase factors, as we did in the definition of STFT (1).

Depending on the grid, the generalized Gabor system can be redundant or insufficient to completely represent signals. In order to ascertain perfect reconstruction, we need the concept of frame which generalizes the concept of basis. Frames are sets of redundant elements that are still able to represent any finite energy signal. A sequence of functions $\{\psi_l\}_{l\in I}$ in the Hilbert space $\mathcal{H}$ is called a *frame*, if there exist positive constants A and B (called lower and upper frame bounds, respectively) such that

$$A\|f\|^2 \leq \sum_{l\in I} |\langle f, \psi_l\rangle|^2 \leq B\|f\|^2 \quad \forall f \in \mathcal{H}, \tag{20}$$

where $\|f\|^2 = \langle f, f\rangle$ is the norm square or total energy of the signal. As stated, the finite and strictly positive frame bounds A and B must be common to all signals f in the Hilbert space, e.g., they can be found as the greatest lower bound and the least upper bound, respectively, of the sum in (20) divided by the norm of the signal as f, such that $\|f\| > 0$, varies over the entire signal space. In the above inequality we recognize that the lower bound A guarantees that no signal has zero projection on all the representative elements. Moreover, in infinite dimensional spaces, like the signal space, one has to additionally require that the projections are finite or, more strictly for the expansion sums to have proper convergence properties, that the sum of the magnitude squares of the projection coefficients is finite, which is guaranteed by the upper bound B in (20). In orthonormal bases one has Parseval's equality stating that the sum of the magnitude square projection coefficients equals the energy of the signal, i.e. $\sum_{l\in I} |\langle f, \psi_l\rangle|^2 = \|f\|^2$. In frames, Parseval equality is only approximate, i.e. $\sum_{l\in I} |\langle f, \psi_l\rangle|^2 \simeq \|f\|^2$ with multiplicative margins A and B. The requirements that B is finite guarantees that the expansion of any finite energy signal in terms of the frame does not blow up. If $A = B$, then $\{\psi_l\}_{l\in I}$ is a *tight frame*, which energy-wise has the same behavior as that of an orthogonal basis.

One can show that the existence of *dual frames*, which can be used for reconstruction, is equivalent to the existence of frame bounds $0 < A, B < \infty$. In turn, this is equivalent to the boundedness and invertibility of the frame operator

$$\mathbf{S}f = \sum_{l} \langle f, \psi_l\rangle \psi_l. \tag{21}$$

The frame operator plays exactly the same role as the matrix operator $\mathbf{H}^T\mathbf{H}$ that we used in the finite dimensional frame example of the previous section; see (17). The *canonical dual frame* $(\tilde{\psi}_l)$, is found by applying the inverse of $\mathbf{S}$ to the original frame elements, i.e. $\tilde{\psi}_l = \mathbf{S}^{-1}\psi_l$ for all l. This is the infinite dimensional analogon of (18) in the finite dimensional frame example of the previous section. For all $f \in \mathcal{H}$ we then have the following reconstruction formulae:

$$f = \sum_l \langle f, \psi_l \rangle \tilde{\psi}_l = \sum_l \langle f, \tilde{\psi}_l \rangle \psi_l.$$

A central property of conventional Gabor frames is the fact that the dual frame of a Gabor frame is again a Gabor frame, generated by the *dual window* $\tilde{g} = \mathbf{S}^{-1}g$ and the same lattice, i.e. the set of time-frequency points $\{(na, qb)\,|n, q \in \mathbb{Z}\}$. Note that the property that the dual system is again a system with the same structure is a particular property of Gabor frames. As we will see, this property is shared by nonstationary Gabor frames in the painless setting, described in Section 2.1.

1.3 Warping Gabor Frames

Any unitary operation on a frame results in a new frame with the same frame bounds A and B [BJ95]. A unitary operator $\mathbf{U}$, the function space analogon of unitary matrices, preserves the scalar product, i.e. $\langle \mathbf{U}f, \mathbf{U}g \rangle = \langle f, g \rangle$ and, in particular, it preserves energy (norm square). Unitary operators can be applied to Gabor frames to obtain new frames. Depending on the operator, the resulting frames are not necessarily of the Gabor type, as the atoms are not generated by time-shifting and modulating a single window function.

A frequency warping operator is completely characterized by a function composition operator in the frequency domain

$$\hat{s}_w = \hat{s} \circ \theta, \tag{22}$$

where θ is the warping map, which transforms the Fourier transform $\hat{s} = \mathcal{F}s$ of a signal s into the Fourier transform $\hat{s}_w = \mathcal{F}s_w$ of another signal s_w, where $\mathcal{F}$ is the Fourier transform operator. If the map θ is one-to-one and almost everywhere differentiable then a unitary form of the warping operator can be defined by the frequency domain action

$$\hat{s}_w(\nu) = \left[\widehat{\mathbf{U}_\theta s}\right](\nu) = \sqrt{\left|\tfrac{d\theta}{d\nu}\right|}\hat{s}(\theta(\nu)). \tag{23}$$

Unitary warping ensures that while bands are stretched their amplitudes are reduced so that the area under the magnitude square Fourier transform is the same as that of the original, i.e., energy is preserved. In particular, when applied to a Gabor frame, the unitary warping operator $\mathbf{U}_\theta$ generates the frequency warped frame $\{\varphi_{n,q}\}_{n,q\in\mathbb{Z}}$ and dual frame $\{\gamma_{n,q}\}_{n,q\in\mathbb{Z}}$:

$$\begin{aligned} \varphi_{n,q} &= \mathbf{U}_\theta \mathbf{T}_{na} \mathbf{M}_{qb} g, \\ \gamma_{n,q} &= \mathbf{U}_\theta \mathbf{T}_{na} \mathbf{M}_{qb} \tilde{g}. \end{aligned} \tag{24}$$

The computation of signal expansions over warped frames requires the generation of sets of warped delayed and modulated windows. Since by warping the time shift operator is transformed into a frequency dependent shift, the windows are not simply translated. Moreover, in principle, the warped windows have infinite support in the time domain even when the original window has finite support.

Alternately, in the computation, one can inversely warp the signal [EC98] and project it onto a conventional Gabor frame. The direct computation of frequency warping consists in taking the Fourier transform of the signal, reassigning frequencies by means of the warping map and taking the inverse Fourier transform.

Filter chain structures for the time-domain computation of frequency warping are also available [EC98]. However, one has to keep in mind that frequency warping is in general a non-causal operation. Approximations of the frequency warping operator, which allow for online computation, are presented in [EC07; Eva08].

Taking the Fourier transform of the first set of functions in (24) we notice that

$$\hat{\phi}_{n,q}(\nu) = \sqrt{\tfrac{d\theta}{d\nu}}\hat{g}\left(\theta(\nu) - qb\right) e^{-j2\pi na\theta(\nu)}. \tag{25}$$

The first two factors correspond to the Fourier transform of the unitarily warped modulated window $\sqrt{\frac{d\theta}{d\nu}}\hat{g}\left(\theta(\nu) - qb\right)$. The last factor shows frequency dependent time shifts: after warping, the original uniform multiple of a time shifts are altered by the phase delay $\theta(\nu)/\nu$. These in-band frequency dependent delays tend to hide or destroy the time structure of the signals, which is a negative factor in signal visualization and synthesis. By means of discrete-time inverse frequency warping acting on the STFT time index n one can revert the frequency dependent delays to in-band constant delays. However, since discrete-time frequency warping implies periodic warping maps, exactly redressing of the delays is possible only in particular cases, e.g. when the analysis window g is bandlimited.

Using the bandlimitedness of the windows assumption, also referred to as the *painless case*, in the next section we introduce families of frequency warped frames that directly show constant in-band delays.

2 Warped Frames with Constant In-Band Delay

In this section we describe frames with arbitrarily assignable bandwidths, which are inspired by the warped frames but have better or more intuitive properties for sound and music computing, i.e. avoiding the presence of in-band frequency dependent delays resulting from frequency warping the time shift operator.

2.1 Building Generalized Gabor Frames with Arbitrary Frequency Bands

As described in Section 1.2, Gabor frames are obtained by uniformly time shifting and modulating a unique time window φ. Here we consider a generalization of Gabor

frames that admits a different window φ_q and different time shift τ_q for each frequency channel indexed by q. In other words, the time-frequency atoms of the representation are the functions

$$\varphi_{n,q}(t) = \mathbf{T}_{n\tau_q}\varphi_q(t) \qquad n, q \in \mathbb{Z} \tag{26}$$

where the modulation is implicit in the windows, i.e., as q varies, the Fourier transforms of the windows $\hat{\varphi}_q$ occupy different frequency bands. Moreover, the bandwidths are allowed to be different and the center frequencies of the bands are not necessarily harmonically related. A frame with frequency channel dependent time shift was introduced in [EC07] and employed in a computationally efficient approximation for warping signals. Properties and applications of generalized Gabor frames having nonuniform frequency resolution were studied in [Vel+11].

The frequency bands of the functions $\hat{\varphi}_q$ are allowed to overlap. Intuitively, a necessary condition for the invertibility of the representation is that there are no gaps or zeros in the total superposition of frequency bands. As we will see, together with the boundedness of the superposition, this is precisely the condition for the set of functions in (26) to form a frame. In this way, using suitable synthesis windows one can achieve perfect reconstruction by means of overlap-add in the frequency domain.

Considering the frequency domain interpretation (4) of the STFT in Section 1.1, it is clear that if the windows are bandlimited – which requires them to have infinite time support – then for any analysis frequency ν the transform produces bandlimited components. Clearly, these components can be sampled at a sufficient rate not smaller than the total bandwidths without hindering perfect reconstruction. Sampling continues to be possible when the bandwidths of the various components are different as in (26).

Formally, in order to ascertain perfect reconstruction, one has to show that (26) constitutes a frame. As remarked in Section 1.2, this is equivalent to study the invertibility and boundedness of the frame operator

$$\mathbf{S}f = \sum_{n,q} \langle f, \varphi_{n,q} \rangle \varphi_{n,q} \tag{27}$$

associated with (26). Taking the Fourier transform of each side in (27), one arrives at the following Fourier representation for the frame operator:

$$\widehat{\mathbf{S}f}(\nu) = \sum_q \tfrac{1}{\tau_q} \sum_m \hat{f}\left(\nu - \tfrac{m}{\tau_q}\right) \overline{\hat{\varphi}_q\left(\nu - \tfrac{m}{\tau_q}\right)} \hat{\varphi}_q(\nu) \tag{28}$$

in which the Fourier transform of the signal $\hat{f}(\nu)$ appears together with its frequency aliased versions $\hat{f}\left(\nu - \frac{m}{\tau_q}\right)$ for $m \neq 0$.

If the frame operator is invertible one can show that upper and lower frame bounds can be determined and that a dual frame $\{\gamma_{n,q}\}_{n,q\in\mathbb{Z}}$ can be provided as follows

$$\gamma_{n,q} = \mathbf{S}^{-1}\varphi_{n,q} \tag{29}$$

In fact, in these circumstances, we have

$$\begin{aligned} f = \mathbf{S}^{-1}\mathbf{S}f &= \sum_{n,q} \langle f, \varphi_{n,q}\rangle \mathbf{S}^{-1}\varphi_{n,q} \\ &= \sum_{n,q} \langle f, \varphi_{n,q}\rangle \gamma_{n,q} \end{aligned} \tag{30}$$

In the frequency domain, inversion of the frame operator requires means to cancel the frequency aliased versions of $\hat{f}$, which may be complex to achieve in the general case. In the so called "painless" case [DGM86], the windows are chosen to have compact support in the frequency domain, i.e. bandlimited, and the time shifts τ_q are chosen so that in (28) the product

$$\overline{\hat{\varphi}_q\left(\nu - \tfrac{m}{\tau_q}\right)}\hat{\varphi}_q(\nu) = 0 \quad \text{for } m \neq 0 \tag{31}$$

That is, the aliased versions of each window have no overlap with the window itself. This is simply achieved by selecting

$$\tau_q \leqslant \frac{1}{B_q} \tag{32}$$

where B_q is the total bandwidth or length of the support of $\hat{\varphi}_q(\nu)$, where we assume that the support is an interval. As announced, in each band, the sampling rate $1/\tau_q$ has to be not smaller than the total bandwidth B_q.

In the painless case, the Fourier representation of the frame operator (28) becomes

$$\widehat{\mathbf{S}f}(\nu) = \hat{f}(\nu) \sum_q \tfrac{1}{\tau_q} |\hat{\varphi}_q(\nu)|^2 \tag{33}$$

This shows that the frame operator is diagonalized by the Fourier transform in the painless case:

$$\widehat{\mathbf{S}f} = \hat{\lambda}\hat{f}, \tag{34}$$

with eigenvalues

$$\hat{\lambda} = \sum_q \tfrac{1}{\tau_q} |\hat{\varphi}_q|^2 \tag{35}$$

See [Dör01; DGM86; Bal+11] for detailed proofs of the diagonality of the frame operator in the described setting.

As shown in the next section, the bandlimited windows assumption provides great simplifications for the definition of frames with arbitrary band allocation.

2.2 Dual Warped Frames

For use in the reconstruction formula, a dual frame associated with (26) can be generated by inverting the frame operator, according to (29). While in general the

inversion of the operator $\mathbf{S}$ poses a problem in numerical realization, we know that under certain conditions, often fulfilled in practical applications, $\mathbf{S}$ is diagonal, in time or frequency domain, as described in the previous section.

We employ windows with adaptive, compact bandwidth and choose the time-shift parameters dependent on the bandwidth of each window: the time-sampling points have to be chosen dense enough to guarantee (32). From (34), since the frame operator associated with (26) is diagonalized by the Fourier transform then it assumes the following form:

$$\mathbf{S}f = \mathcal{F}^{-1}[\hat{\lambda}\hat{f}] = \mathcal{F}^{-1}[\hat{\lambda}] * f. \tag{36}$$

From (34) and (35) it follows immediately that the frame operator is invertible whenever there exist real numbers numbers A and B such that the inequalities

$$0 < A \leq \sum_q \frac{1}{\tau_q}|\hat{\varphi}_q|^2 \leq B < \infty \tag{37}$$

hold almost everywhere. From (36) it follows that the inverse of the frame operator has the following form:

$$\mathbf{S}^{-1}f = \mathcal{F}^{-1}[\hat{f}/\hat{\lambda}] = \mathcal{F}^{-1}[1/\hat{\lambda}] * f. \tag{38}$$

For the Fourier transform of the frame elements we have:

$$\hat{\varphi}_{n,q}(\nu) = \mathcal{F}\left[T_{n\tau_q}\varphi_q\right](\nu) = e^{-j2\pi\nu n\tau_q}\hat{\varphi}_q(\nu) = \mathbf{M}_{-n\tau_q}\hat{\varphi}_q(\nu). \tag{39}$$

Thus, applying (29), the dual frame is given by the elements

$$\gamma_{n,q} := \mathcal{F}^{-1}\left[\mathbf{M}_{-n\tau_q}\left(\hat{\varphi}_q/\hat{\lambda}\right)\right] = \mathbf{T}_{n\tau_q}\mathcal{F}^{-1}\left[\hat{\varphi}_q/\hat{\lambda}\right]. \tag{40}$$

As a result, in order to obtain the synthesis windows γ_q, one needs to filter the analysis windows with the frequency response $1/\hat{\lambda}$. As for the analysis, the synthesis frame elements are obtained by frequency channel dependent time-shifting by integer multiples of τ_q.

Based on the implementation of nonstationary Gabor frames performing adaptivity in the time domain, the above framework permits a fast realization of frequency-adaptive Gabor frames directly in the frequency domain, by considering the Fourier transform of the input signal. The transform coefficients $S_{n,q} = \langle f, \varphi_{n,q}\rangle$ take the form

$$S_{n,q} = \langle \hat{f}, \mathbf{M}_{-n\tau_q}\hat{\varphi}_q\rangle, \tag{41}$$

and can be calculated, for each q, from the FFT of the signal with a number of operations solely determined by the support of $\hat{\varphi}_q$. Similarly, reconstruction is realized by applying the windows $\mathbf{M}_{-n\tau_q}\hat{\gamma}_q$, where $\hat{\gamma}_q = \hat{\varphi}_q/\hat{\lambda}$, in a simple overlap-add process in the frequency domain. Whenever perfect reconstruction is necessary, the transform parameters must be chosen, such that conditions (32) and (37) are satisfied.

2.3 Designing Frames with Arbitrary Band Allocations by Means of Frequency Warping

In this section a design procedure for frames with arbitrary allocation of frequency bands is described, which is based on frequency warping. In [Vel+11] the design of "painless" frames with compact support in the frequency domain and arbitrary band allocation was approached by scaling the Von Hann window in the frequency domain. The main advantage of the warping approach is that the bands can be allocated following a curve, the warping map, that can be derived from physical or perceptual characteristics of the signal. Another advantage is that by frequency warping one can obtain tight frames with arbitrary band allocation. While tight frames can also be obtained from the original design by reassigning the windows as follows

$$\hat{\varphi}_q \leftarrow \frac{\hat{\varphi}_q}{\sqrt{\sum_k \frac{1}{\tau_k} \left|\hat{\varphi}_k\right|^2}},$$

the denominator introduces in-band ripple due to the imperfect overlap of the original windows at transition bands. The design based on warping eliminates the need for re-normalization and allows for smooth transition bands. It can be performed starting from any tight uniform Gabor frame.

To approach the frequency warping based design, consider a nonnegative symmetric window $\hat{h}(\nu)$ in the frequency domain satisfying the requirement

$$\sum_q \hat{h}^2\left(\nu - bq\right) = 1 \tag{42}$$

for a certain $b > 0$. Several well-known windows satisfy (42). For example one can select the square root of the Von Hann window, i.e. the cosine window

$$\hat{h}(\nu) = \begin{cases} \sqrt{\frac{2}{K}} \cos \frac{\pi\nu}{\beta} & \text{if } -\frac{\beta}{2} \leqslant \nu < +\frac{\beta}{2} \\ 0 & \text{otherwise} \end{cases} \tag{43}$$

where $\beta > 0$ is the total bandwidth of the window and $K > 1$ is an integer. The cosine window satisfies (42) for $b = \beta/K$.

The frequency support of each of the windows $\hat{h}_q(\nu) = \hat{h}(\nu - bq)$ is the interval $\left[bq - \frac{\beta}{2}, bq + \frac{\beta}{2}\right]$.

In order to construct windows with nonuniform bandwidth one can start by prescribing a monotonically increasing, one-to-one, map θ of the frequency axis, which we assume henceforth to be an almost everywhere differentiable odd function of frequency. In practical applications, the map can be inspired by physical or perceptual characteristics. Otherwise, if only the desired center band frequencies ν_q are prescribed, one can build a smooth continuous map by interpolation from the pairs (ν_q, bq).

To each of the functions $h_q = \mathcal{F}^{-1}[\hat{h}_q]$ we apply a nonunitary warping operator $\mathbf{W}_\theta$ built as the operator $\mathbf{U}_\theta$ in (23) without the square root derivative factor. As a result, the

uniformly spaced windows $\hat{h}_q$ are transformed to the nonuniformly spaced frequency domain windows

$$\hat{g}_q(\nu) = \widehat{\mathbf{W}_\theta h_q}(\nu) = \hat{h}_q(\theta(\nu)) = \hat{h}(\theta(\nu) - bq) \tag{44}$$

Since (42) holds for any ν then warping preserves the constant overlap-add property:

$$\sum_q \hat{g}_q^2(\nu) = \sum_q \hat{h}^2(\theta(\nu) - bq) = 1 \tag{45}$$

The frequency domain shapes of the original windows are smoothly altered by the warping map so that the nonuniformly stretched windows blend nicely into each other.

The center frequencies ν_q of the warped windows are the solutions of the equations $\theta(\nu) - bq = 0$ that is:

$$\nu_q = \theta^{-1}(bq) \tag{46}$$

since the map is invertible. Similarly, the band edge frequencies $\nu_q^\pm$ are solutions of the equations $\theta(\nu) - bq = \pm\beta/2$, i.e.,

$$\nu_q^\pm = \theta^{-1}\left(bq \pm \tfrac{\beta}{2}\right) \tag{47}$$

Comparing (45) with (35) and (37), we see that by letting

$$\hat{\varphi}_q = \sqrt{\tau_q}\hat{g}_q \tag{48}$$

with the warped windows one can achieve

$$\hat{\lambda}(\nu) = \sum_q \tfrac{1}{\tau_q} |\hat{\varphi}_q(\nu)|^2 = 1 \tag{49}$$

That is, tight frames with unit frame bounds $A = B = 1$ can be generated with the warped uniform windows, in which case the dual frame $\{\gamma_{n,q}\}_{n,q\in\mathbb{Z}}$ coincides with the frame $\{\varphi_{n,q}\}_{n,q\in\mathbb{Z}}$.

The atoms of the frame and dual frame can be generated by time-shifting the inverse Fourier transforms of the windows as in (26). Since the supports of the warped windows are the intervals

$$\left[\theta^{-1}\left(bq - \tfrac{\beta}{2}\right), \theta^{-1}\left(bq + \tfrac{\beta}{2}\right)\right] \tag{50}$$

in order to fulfill (32) one needs to select

$$\tau_q \leqslant \frac{1}{\theta^{-1}\left(bq + \frac{\beta}{2}\right) - \theta^{-1}\left(bq - \frac{\beta}{2}\right)} \tag{51}$$

We remark that, if the bandwidth ν of the original window h is small, then

$$\begin{aligned} \theta^{-1}\left(bq + \tfrac{\beta}{2}\right) - \theta^{-1}\left(bq - \tfrac{\beta}{2}\right) &\approx \beta \left.\frac{d\theta^{-1}}{d\nu}\right|_{\nu=bq} \\ &= \beta \left(\left.\frac{d\theta}{d\nu}\right|_{\nu=\theta^{-1}(bq)}\right)^{-1} \end{aligned} \tag{52}$$

Thus, for the largest values of allowed time-shifts we have

$$\tau_q \approx \frac{\theta'(\theta^{-1}(bq))}{\beta} = \frac{\theta'(\theta^{-1}(bq))}{Kb} \tag{53}$$

Hence, if the map is identical, $\theta(\nu) = \nu$, we have $\tau_q b \approx 1/K$ as in the uniform Gabor case with redundancy overlap factor K.

In the warped case, the time-shifts are proportional to the approximate frequency dependent time-stretching of the narrow band windows due to frequency warping, which are given by group delay θ' evaluated at the center bands. Moreover, the normalization factors in (48) can be interpreted as proportional to the square root of the derivative of the warping map θ evaluated at the center bands of the warped windows, which approximately restores the normalization factor of the unitary warping operator (23) to the nonunitarily warped windows.

3 Applications and Examples

The flexibility in time-frequency tiling achieved by the arbitrary band vocoder described in this paper paves the way to several applications in sound and music computing, coding and music information retrieval. The tightness of the frames guarantees that the total energy of the signal is equal to the sum of the squares of the analysis coefficients and that no multiplicative bias is introduced in the analysis that is then taken away by the synthesis.

3.1 Adaptation to Desired Frequency Scales

The allocations of the bands of the frequency channels is performed by means of a frequency map θ, mapping nonuniform bands into uniform ones. It is convenient to normalize the map so that the desired center frequencies are mapped to integers denoting the channel indices. The map can be derived from perceptual, physical or musical scales.

An example of adapted frequency band characteristics is shown in Fig. 6, where the map is directly obtained from the perceptual Bark scale [Tra90], resulting in 25 channels with bandwidths increasing with frequency in the frequency range 0 to 22 KHz. It is apparent how the bands are not simply obtained by constant scaling: the frequency characteristics are more stretched for the upper band portions than for the lower ones.

Another example is reported in Fig. 7 where the frequency tiling is adapted to a $1/3$ of octave scale, resulting in 33 channels. The arbitrariness of the map allows one to build frames tuned to any scale and pitch resolution.

3.2 Frequency Channels

Given the integer N obtained by rounding the maximum value of the normalized warping map, the bandwidth parameter b in (42) is selected by dividing the maximum frequency of the total range by N.

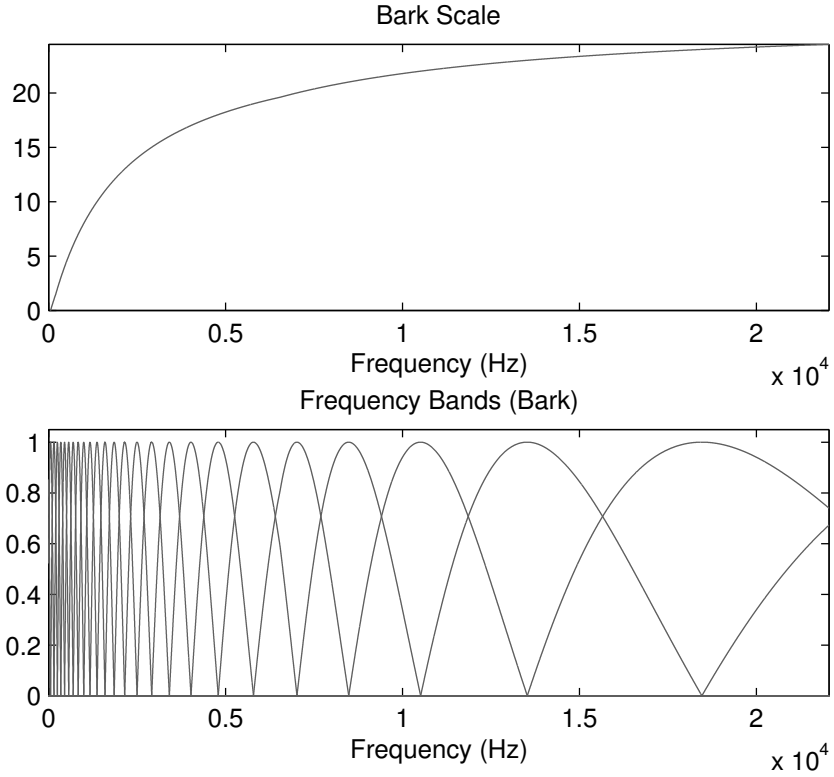

Fig. 6: Warping the frame frequency channels according to Bark scale (top): resulting frequency band characteristics (bottom).

The actual number of frequency channels depends on the overlap factor K. For $K = 2$, useful in most applications, the final number of channels is given by $N + 1$, which includes the 0 frequency channel and the highest frequency channel that has support overlapping with the frequency range of the signal. When needed, the frequency response of the 0 frequency channel is obtained by fixing a lower bandwidth and summing together all the frequency windows having center frequency below the lower bandwidth.

For real signals, negative frequency channels can be obtained by complex conjugation of the corresponding positive frequency channel and do not need to be computed.

3.3 Implementation

In our implementation we employed an algorithm directly derived from (41) in which the FFT of the signal is first computed and the scalar products are computed in the frequency domain over the frequency support of the windows $\hat{\varphi}_q$. This algorithm

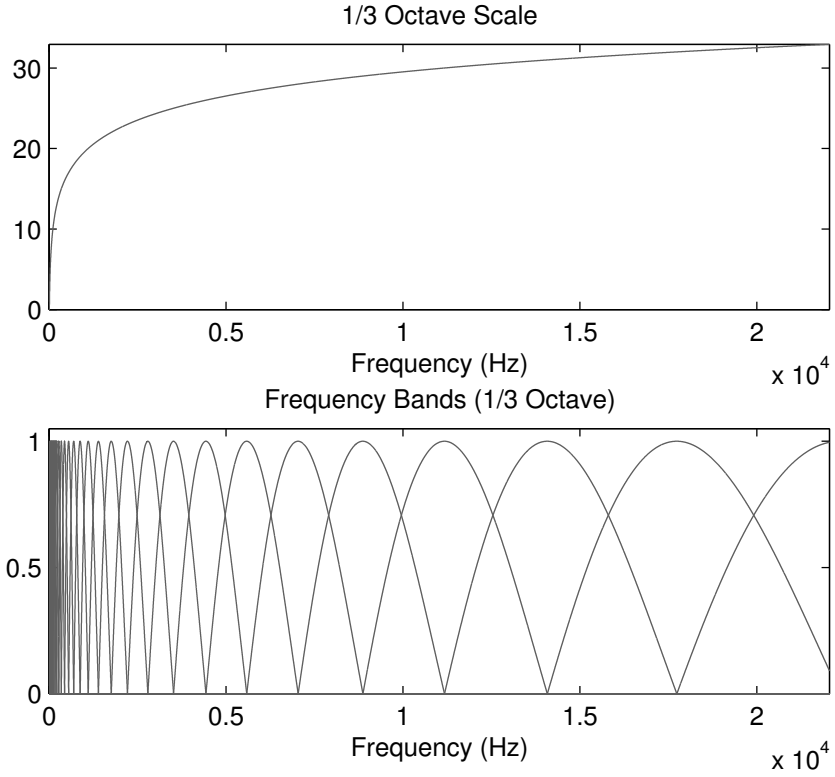

Fig. 7: Warping the frame frequency channels according to a $1/3$ of octave scale (top): resulting frequency band characteristics (bottom).

is shown in [Vel+11] to have complexity $O(L \log L)$ for both analysis and synthesis, where L is the length of the signal.

Implementation is therefore very simple and it parallels in the frequency domain what is usually performed in the time domain, in terms of windowing for the analysis and overlap-add for the synthesis. At different time instants the windows only differ by a phase factor, which is easily obtained by multiplication.

3.4 Nonuniform Spectrograms

In order to display time-frequency spectrograms with synchronous time scale for all bands, it is convenient to compute the transform coefficients with time oversampling, choosing the sampling intervals τ_q in (51) to be all equal to their minimum value τ. This is always possible for finite bandwidth signals. In this case, at each time step one needs to modulate the DFT of the signal, multiplying it by $e^{j2\pi\nu\tau}$. Incidentally, should we modulate, at this step, by the factor $e^{j2\pi b\theta(\nu)\tau}$ instead, we would obtain a frequency domain implementation of the expansion over the warped Gabor frames in (24).

Here the frequency ν takes the discrete values $k\nu_s/L$, where $k \in \{0, \dots, L/2-1\}$, obtained from the FFT calculation, in which ν_s is the sampling rate of the signal. The modulated DFT of the signal is stored and reused in the next time step calculation.

The spectrogram based on the 12-tone equally tempered scale of a singing musical phrase is shown in Fig. 8, resulting in 118 frequency channels at $\nu_s = 44.1$ KHz. For comparison, the uniform frequency band spectrogram of the same signal based on the same number of channels is shown in Fig. 9, using the same distribution of frequency bins. It is apparent how, at same computational cost, the 12-tone spectrogram better captures the musical score by zooming into the part of the time-frequency domain, where most energy of the signal is concentrated.

For another comparison, the spectrogram based on the warped Gabor frames (24) is reported in Fig. 10. One can notice how the presence of the dispersive delays destroys the time structure of the score.

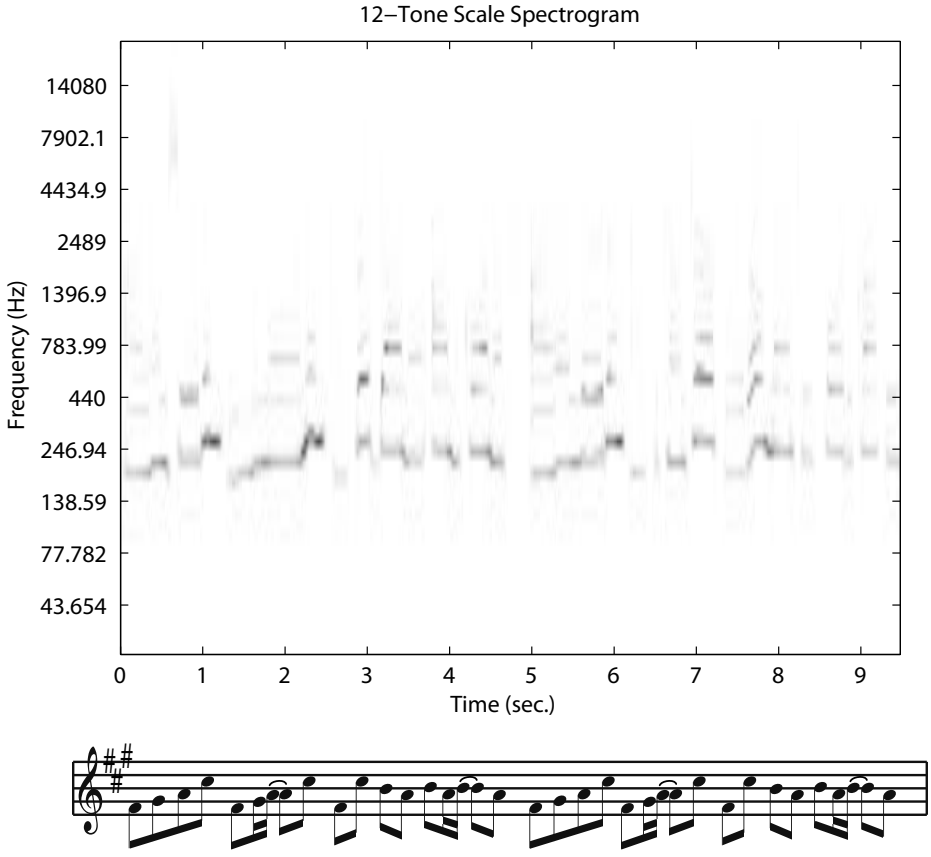

Fig. 8: Nonuniform 12-tone scale spectrogram of the singing phrase represented in the score line [from *Tom's Diner*, Suzanne Vega], in which it is possible to track in tempered time-frequency scale the score and even the glissando introduced by the singer.

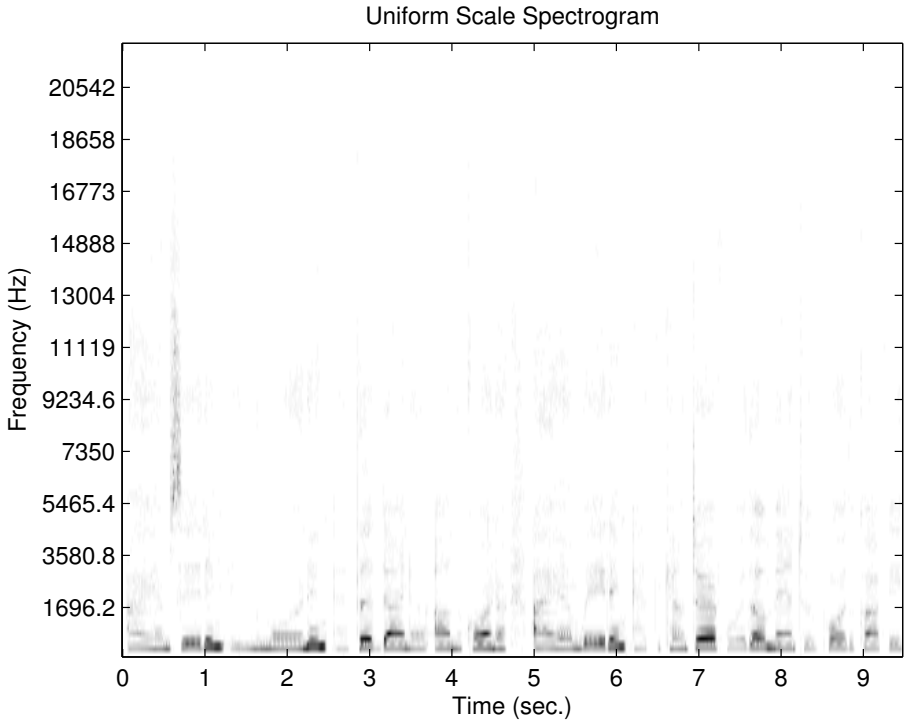

Fig. 9: Uniform frequency scale spectrogram of singing phrase.

As another application example suitable for coding applications, the nonuniform spectrogram based on $1/4$ Bark scale (4 channels per Bark) of a complex rock music excerpt is shown in Fig. 11. For comparison, the uniform spectrogram is plotted below in Fig. 12. It is apparent how the perceptually relevant information is distributed in time-frequency by using the nonuniform spectrogram.

3.5 Sound Computing Applications

The flexibility of the signal representation illustrated in this paper inspires a number of creative uses. For example, in the analysis and synthesis of sounds with nonharmonic partials one can adapt the frequency bands to capture the frequency content of the main partials, together with that of other bands lying in-between the given partials. If the frequency channels corresponding to the partials are suppressed, one can study the noise or fluctuations of the signal.

For example, in piano tones in the low register, one can extract the hammer noise by re-synthesizing all the nonpartial bands. Vice versa, in denoising or audio restoration applications, one may want to suppress the side bands of the partials, which contain disturbing unmasked noise.

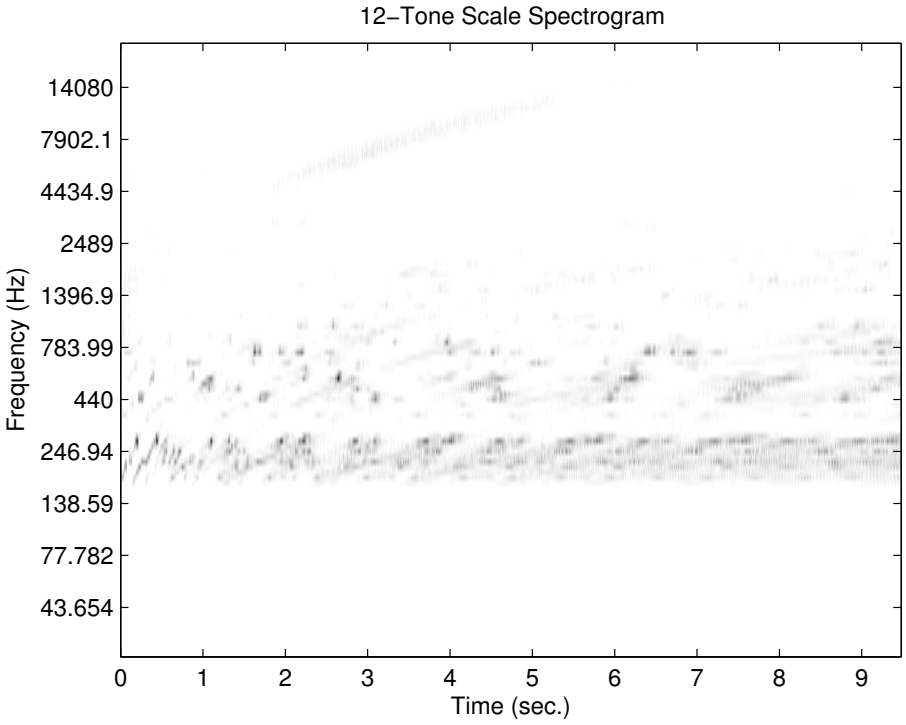

Fig. 10: Warped Gabor spectrogram of singing phrase.

In an analysis-synthesis scheme similar to the one illustrated in [PE07], within the proposed signal representation one can allocate narrow bands centered on the frequencies of the signal's partials and wider left and right sidebands of the partials. This is realized by means of a pitch dependent smoothed staircase warping map, with higher slope around the frequencies of the partials and lower slope at the sidebands. The sidebands represent fluctuations of the partials that can be often modeled as modulated $1/f$ noise, while the partials are represented by low-rate pitch and amplitude information. Perceptually, the presence of the fluctuations is relevant, while the details of the fluctuations are less relevant, which allows for coarser modeling of the sidebands. The flexibility of the presented scheme allows for accurate tuning of the representation bands, which is far less rigid than the allocation in [PE07].

In other applications as sound effects, one can use uniform bands for the analysis coupled with nonuniform bands for the synthesis with same coefficients. Thus, the frequency content of the signal in the analysis bands is displaced to other frequency bands in the synthesis, resulting in band stretching and modulation. In this way one obtains an efficient algorithm for frequency warping, similar to the one presented in [EC07].

In conventional uses of the phase vocoder, such as time stretching, frequency shifting

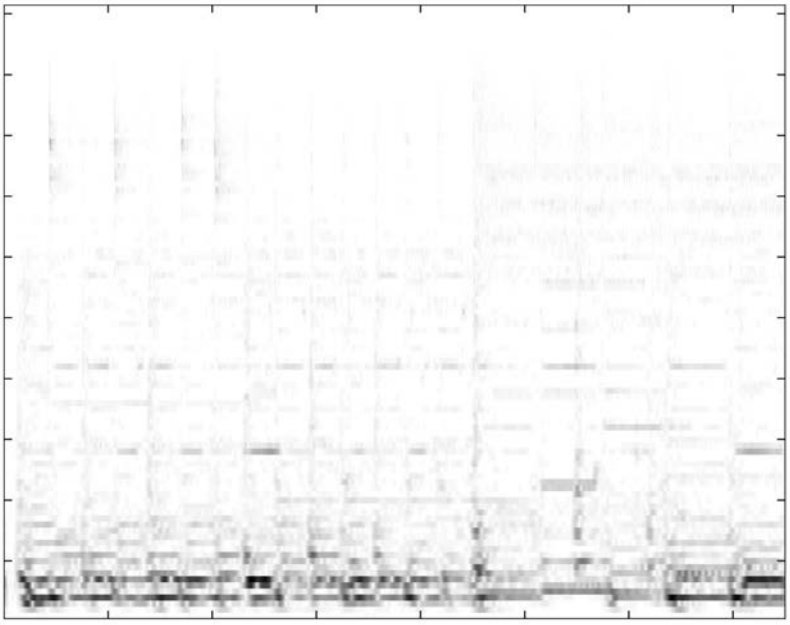

Fig. 11: Bark scale spectrogram of rock music excerpt.

and harmonizer, one is often interested in tracking the partials, which requires estimation of the instantaneous frequencies. This is usually performed by time differencing the phase of the transform coefficients, as evaluated at adjacent time instants. For the estimation to be successful one needs to have sufficiently narrow-band analysis atoms so that interference of distinct signal partials is low. The proposed transform allows us to design the frequency resolution arbitrarily, assigning, e.g., higher resolution around the expected frequencies of the partials, which for most signals are in the lower frequency portion of the spectrum.

Even in the constant Q case, one can set the resolution to arbitrary fractions of a tone, as sufficient for frequency tracking. Mixed mode is also possible, e.g., by assigning uniform resolution at low frequencies and constant Q at high frequencies. Compared to orthogonal wavelets, whose resolution in the simplest and most popular case is one octave throughout the frequency spectrum, this is a major improvement. The mixed mode allows us to combine the benefits of conventional phase vocoders and of wavelets in terms of an arbitrarily configurable frequency channel allocation. Moreover, time sampling can directly reflect the bandwidth allocation for each channel. The sampling theorem exactly holds in view of the compactly supported windows in the frequency domain.

In conventional phase vocoders based on a compactly supported window in the time domain, frequency leakage of a single sinusoidal component of the signal occurs

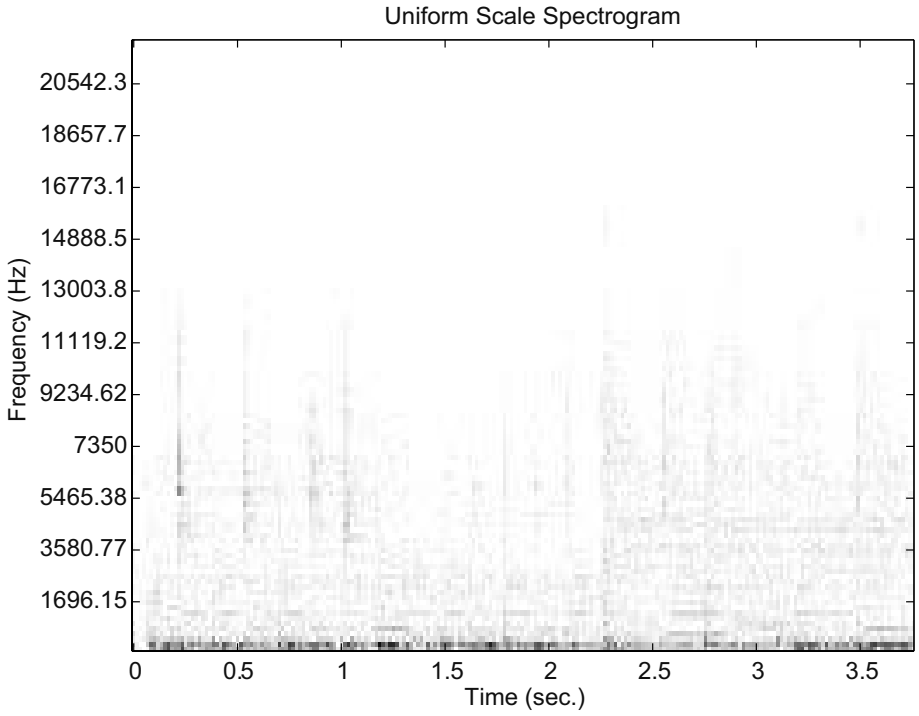

Fig. 12: Uniform spectrogram of rock music excerpt.

throughout the spectrum. Since the analysis atoms of the proposed representation have compact support in the frequency domain, frequency leakage is exactly confined so that interference can only occur from signal components falling in adjacent bands. Both amplitude and phase estimation of the partials benefit from this localization property.

Additional examples together with the corresponding sound files are available at `http://homepage.univie.ac.at/monika.doerfler/WarpFrames.html`

This page will be continuously updated as new examples and applications of the proposed framework become available.

4 Conclusions and further work

In this paper we have explored the design of perfect reconstruction phase vocoders based on nonuniform Gabor frames with arbitrary band allocation.

By means of a warping map, uniform frequency bands are mapped into nonuniform frequency bands, while keeping constant the sampling rate within each band. The representation is shown to be useful in several applications adding flexibility to the well known and ubiquitous concept of phase vocoder in sound and music computing. These

range from visualization of musical data to audio effects, feature extraction, coding and synthesis.

Some important aspects of the phase vocoder, in particular concerning phase estimation of the partials in each band, have been pointed out but not analytically addressed in the current contribution. These aspects are particularly useful and important in sound transformation (stretching, transposition) and their adaptation to the flexible framework introduced will be presented elsewhere.

Methods allowing for nearly constant delay within each band even when the supports of the bands are not finite have been developed and will be the object of a forthcoming publication. These methods are relevant for real-time computation for example to allow the windows to have finite support in the time domain. In future work we will further consider the use of nearly perfect reconstruction nonuniform representations that are suitable for real time computation and achieve a compromise between time and frequency localization. Furthermore, a time-varying band allocation method generalizing the scheme in [Eva08] is under investigation.

References

[BJ95] R. G. Baraniuk e D. L. Jones. “Unitary Equivalence : A New Twist on Signal Processing”. In: *IEEE Trans. Signal Processing* 43.10 (1995), pp. 2269–2282.

[DGM86] Ingrid Daubechies, A. Grossmann e Y. Meyer. “Painless nonorthogonal expansions”. In: *J. Math. Phys.* 27.5 (1986), pp. 1271–1283.

[Dol86] M. Dolson. “The phase vocoder: a tutorial”. In: *Computer Musical Journal* 10.4 (1986), pp. 11–27.

[EC07] Gianpaolo Evangelista e Sergio Cavaliere. “Real-Time and Efficient Algorithms for Frequency Warping Based on Local Approximations of Warping Operators”. In: *Proc. of Digital Audio Effects Conf. (DAFx ‘07)*. Bordeaux, France, 2007, pp. 269–276.

[EC98] Gianpaolo Evangelista e Sergio Cavaliere. “Frequency Warped Filter Banks and Wavelet Transform: A Discrete-Time Approach Via Laguerre Expansions”. In: *IEEE Trans. on Signal Processing* 46.10 (1998), pp. 2638–2650.

[EDM12] Gianpaolo Evangelista, Monika Dörfler e Ewa Matusiak. “Phase Vocoders With Arbitrary Frequency Band Selection”. In: *Proceedings of the 9th Sound and Music Computing Conference*. Copenhagen, Denmark, 2012, pp. 442–449.

[Eva08] Gianpaolo Evangelista. “Modified Phase Vocoder Scheme for Dynamic Frequency Warping”. In: *Proc. of IEEE 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP 2008)*. St. Julians, Malta, 2008, pp. 1291–1296.

[Gab46] D. Gabor. “Theory of communications”. In: *J. IEE* III.93 (1946), pp. 429–457.

[Mal98] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press London, 1998.

[PE07] Pietro Polotti e Gianpaolo Evangelista. “Fractal additive Synthesis: a Deterministic/Stochastic Model for Sound Synthesis by Analysis”. In: *IEEE Signal Processing Magazine* 24.2 (2007), pp. 105–115.

[Tra90] Hartmut Traunmüller. “Analytical expressions for the tonotopic sensory scale”. In: *The Journal of the Acoustical Society of America* 88.1 (1990), pp. 97–100.

[Bal+11] Peter Balazs et al. “Theory, implementation and applications of nonstationary Gabor Frames”. In: *J. Comput. Appl. Math.* 236.6 (2011), pp. 1481–1496.

[Dör01] M. Dörfler. “Time-frequency Analysis for Music Signals. A Mathematical Approach”. In: *Journal of New Music Research* 30.1 (2001), pp. 3–12.

[Vel+11] Gino Angelo Velasco et al. “Constructing an invertible constant-Q transform with non-stationary Gabor frames”. In: *Proc. of the Digital Audio Effects Conf. (DAFx-11)*. Paris, France, 2011, pp. 93–99.