Audio Pitch Shifting Using the Constant-Q Transform

CHRISTIAN SCHÖRKHUBER[1], ANSSI KLAPURI[2], AND ALOIS SONTACCHI,[1] *AES Member*
(schoerkhuber@iem.at) (anssi@ovelin.com) (sontacchi@iem.at)

[1]*Institute of Electronic Music and Acoustics, Graz, Austria*
[2]*Ovelin, Helsinki*

Pitch shifting of polyphonic music is usually performed by manipulating the time-frequency representation of the input signal. Most approaches proposed in the past are based on the Fourier transform although its linear frequency bin spacing is known to be inadequate to some degree for analyzing and processing music signals. Recently invertible constant-Q transforms (CQT) featuring high Q-factors have been proposed exhibiting a more suitable geometrical bin spacing.

In this paper a frequency-domain pitch shifting approach based on the CQT is proposed. The CQT is specifically attractive for pitch shifting because it can be implemented by frequency translation (shifting partials along the frequency axis) as opposed to spectral stretching in the Fourier transform domain. Furthermore, the high time resolution of CQT at high frequencies improves transient preservation. Audio examples are provided to illustrate the results achieved with the proposed method.

1 INTRODUCTION

Pitch shifting is a digital audio effect that changes the pitch of a sound or music signal without altering its duration. That is, all frequencies are scaled by a constant factor. Applications of pitch shifting include pitch correction of musical performances in the recording studio, transposing songs to a desired key, and voice modification, to mention a few examples.

A common approach to pitch shifting is based on a two-stage process: first, the time-scale of the input signal is modified (time-stretching), and then the output signal is resampled to restore the signal's original time-base but having shifted its frequency content. In general, time scaling can either be performed in the time domain [1] or the time-frequency domain. Approaches operating in the time-frequency domain are often based on the phase vocoder where time scaling is achieved by altering the analysis or synthesis frame hop size [2]. In [3] a constant frame-hop phase vocoder is proposed where time scaling is achieved by copying or deleting frames. Several improvements to phase-vocoder-based time scaling have been proposed to reduce phasiness [4] [5] and transient smearing [6].

Alternatively, several implementations have been proposed that perform pitch shifting directly without a time-stretching stage, mainly based on the phase vocoder [7] and on synchronous overlap-add (SOLA) [8,9]. The former operates in the time-frequency domain while the latter operates in the time-domain. Usually time domain approaches are more efficient computationally while time-frequency-domain approaches based on the time scaling achieve higher quality for polyphonic music signals.

Typically phase-vocoder approaches suffer from artifacts that are perceived as phasiness and transient smearing. Both problems stem from the loss of "horizontal" phase coherence between frames and/or loss of "vertical" phase coherence within frames [2]. The standard implementation to reduce typical phase vocoder artifacts uses instantaneous frequency estimation and phase unwrapping to establish phase coherence between frames and a phase locking scheme to (partly) retain within-frame phase coherence. The phase locking process comprises a spectral peak picking stage to define regions within each frame that are dominated by single sinusoidal components (peaks). The phase values within these regions of influence are assumed to be dominated by the peak's phase and are thus "locked" to its phase value after the phase propagation between frames has been performed.

Audio quality of the phase-vocoder-based pitch-shifter depends strongly on whether the implicit assumptions regarding the time-frequency representation of a particular input signal are valid. The phase update process is only correct when the input signal can be modeled as a sum of a small number of slowly varying sinusoids. Furthermore, it is assumed that each sinusoidal component excites a

discrete peak in the spectrum and that the spectrum can be divided into *independent* regions that are dominated by a single sinusoidal component (see Section 2). Especially for dense polyphonic music these assumptions do not always hold since phase-vocoder-based pitch-shifting implementations usually operate on the Short Time Fourier Transform (STFT) representation. A well known disadvantage of the STFT is the rigid time-frequency resolution trade-off providing a constant absolute frequency resolution throughout the range of audible frequencies. In contrast to this we know that due to both musical and auditory aspects a frequency resolution is preferred that increases from high to low frequencies (and vice versa for time resolution). Using the STFT representation, two closely spaced sinusoids in low-frequency regions might excite only one spectral peak. On the other hand, the time-resolution at high-frequency regions might be too coarse to capture quick temporal changes.

For phase-vocoder-based *time-scaling* implementations these issues have been addressed by several authors. In [3] a constant frame-rate phase vocoder has been proposed where multiple windows are used in parallel. That is, the input signal is subdivided in three frequency bands and a different STFT window length is used in each band (multiresolution discrete Fourier transform). In [5] the authors propose to use a multiresolution technique at the peak picking stage where the peak detection function is made frequency dependent to address the fact that closely spaced spectral peaks need to be processed separately in low-frequency regions but can be combined in high-frequency regions. This notion stems from the fact that the frequency resolution of the human auditory system is approximately inversely proportional to frequency and that most audio signals exhibit a non-uniform distribution of their partials.

Unfortunately, for phase-vocoder-based pitch-scaling implementations [7] the application of multiple Fourier transform resolutions in parallel is not straightforward. In this approach each spectral peak (representing a sinusoidal component) and its neighborhood (region of influence) is translated up or down in frequency according to a given pitch-shifting factor. The abrupt time-frequency resolution changes around the subband boundaries in the multiresolution technique do not allow coefficient shifts across the boundaries.

Another considerable inconvenience is the fact that a different frequency translation (shift) has to be applied on each spectral peak, because the STFT frequency bins are linearly spaced whereas scaling all frequencies by a constant factor α corresponds to a constant shift on log-frequency scale. In dense polyphonic music signals, that leads to higher computational complexity and requires interpolation for fractional frequency-bin translations.

Despite the above challenges, frequency-domain pitch shifting has several advantages over the two-stage time-scaling/resampling implementation: The computational complexity is independent of the scaling factor and sinusoidal components can be shifted independently. That is, single notes in the input signal can be altered while leaving others untouched. Such needs arise for example when pitch-correcting the sound of one instrument within a polyphonic signal.

Applying the constant-Q transform (CQT) in place of the STFT for time-frequency-domain pitch shifting provides a solution to all of the aforementioned disadvantages of this approach. Constant-Q transform refers to a technique that transforms a time-domain signal $x(n)$ into the time-frequency domain so that the center frequencies of the frequency bins are geometrically spaced and their Q-factors are all equal. In effect, this means that the frequency resolution is better for low frequencies and the time resolution is better for high frequencies. The CQT is essentially a wavelet transform, but here the term CQT is preferred since it emphasizes the fact that we are considering transforms with relatively high Q-factors, equivalent to 12–96 bins per octave. This renders many of the conventional wavelet transform techniques inadequate; for example, methods based on iterated filterbanks would require filtering the input signal hundreds of times [10]. The CQT was proposed in [11] but has been playing only a minor role in the fields of music analysis and music processing since then. The main reasons for this were its complexity when broadband music signals are considered and the fact that it lacked an inverse transform that would allow reconstruction of the original signal from its transform coefficients.

In [12] we proposed solutions to these problems providing a Matlab toolbox for efficient computation of the CQT coefficients and reasonable quality reconstruction (around 55 dB signal-to-noise ratio) of the original input signal. Recently another approach to invertible constant-Q transforms featuring high Q-factors has been proposed, yielding even perfect reconstruction [13]. The suggested transform is based on frame theory [14] and utilizes nonstationary Gabor frames [15] to achieve geometrically spaced frequency bins and the property of invertibility. In [16] a discrete-time wavelet transform with tunable Q-factor based on a real-valued dilation factor has been presented, which is implemented using a perfect reconstruction over-sampled filter bank with real-valued sampling factors.

One of the major benefits of using the CQT for frequency-domain pitch shifting is that processing the magnitude spectrum becomes a trivial operation: translating CQT coefficients up or down in frequency corresponds to *scaling* the frequencies with a constant factor. In other words, the entire magnitude spectrum can be translated by the same amount and usually interpolation of the spectral components can be avoided. Second, the time-frequency resolution of CQT varies as a function of frequency (similarly to the human auditory system), which makes it easier to achieve high audio quality. More specifically, transients are largely represented by the high time resolution magnitude spectrum information at high frequencies and need not be encoded in the phase spectrum. On the other hand, distinguishing nearby frequency components (and their regions of influence) at low frequencies is facilitated by the high frequency resolution at the lower end of the spectrum.

With efficient invertible CQT implementations now at hand we present a frequency-domain pitch-shifting

algorithm based on the CQT representation of music signals.[1]

In Section 2 we briefly describe how time-frequency-domain pitch shifting is performed using the Fourier transform and discuss some of its drawbacks. In Section 3 we summarize some aspects of the CQT implementation we proposed in [12] that are crucial for the present CQT-based pitch-shifting algorithm outlined in Section 4. In [18] we provide audio examples to illustrate the achieved quality of the proposed approach (see Section 5).

2 BACKGROUND: STFT-BASED FREQUENCY-DOMAIN PITCH SHIFTING

In order to understand the problems that arise in STFT-based pitch shifting and to provide background for the proposed method, let us first diagnose the frequency-domain pitch shifting implemented using the STFT.

2.1 Implementation

Pitch shifting using the STFT representation of an audio signal as proposed in [7] is performed in four steps:

1. Peak detection: The simplest scheme consists of declaring that a bin is a peak if its magnitude is larger than that of its two (or four) nearest neighbors. It is assumed that each detected peak represents a sinusoidal component.
2. Define regions of influence: The region of influence is the sub-band around a spectral peak in which it is assumed that all phase values are dominated by the peak's phase. The boundaries of these sub-bands can be defined halfway between two peaks or at the lowest magnitude bin between two peaks.
3. Coefficient shift: Peaks and their regions of influence are shifted by frequency $\Delta f_m = f_m(\alpha - 1)$, where f_m is the frequency of peak m and α is the pitch-scaling factor. As observed in [7], if the relative amplitudes and phases of the bins around a sinusoidal peak are preserved during the translation, then the time-domain signal corresponding to the shifted peak is simply a sinusoid at a different frequency, modulated by the same analysis window.
4. Phase update: Since the frequencies of underlying sinusoids have been changed during the coefficient shift, phase-coherence from one frame to the next is lost. To avoid artifacts due to horizontal phase inconsistency, phase values need to be updated.

As discussed in Section 1 the standard techniques to retain phase coherence in phase vocoder implementations are instantaneous frequency estimation and phase unwrapping for horizontal phase coherence and phase-locking for vertical phase coherence. Since the loss of phase coherence due to a synthesis frame hop size that differs from the analysis frame hop size (time scaling) is analogous to loss of phase coherence due to a frequency shift, the same phase update techniques could be used [2]. However, in [7] a very simple phase update scheme is proposed that does not involve instantaneous frequency estimation and can thus be implemented very efficiently. If a peak is shifted by Δf, the difference of the peak's phase between two successive frames must be increased or decreased by an amount consistent with the modified frequency of the underlying sinusoid. For a constant-frequency sinusoidal component, phase coherence can be achieved by simply multiplying each STFT coefficient in the region of influence by the complex

$$Z_u = e^{i\frac{2\pi R}{f_s}\Delta f_{m,u}} \tag{1}$$

where R is the frame hop size, $\Delta f_{m,u}$ is the frequency difference due to shifting peak m in frame u, and f_s is the sampling frequency. These phase rotations have to be accumulated from one frame to the next, that is

$$Z_{u+1} = Z_u e^{i\frac{2\pi R}{f_s}\Delta f_{m,u}}. \tag{2}$$

Under the assumption that all phase values in the region of influence are dominated by the peak's phase, horizontal and vertical phase coherence can thus be retained exactly for a constant-frequency sinusoid. Due to the linear spacing of STFT frequency bins, the phase rotation that has to be applied to the STFT coefficients in this approach is independent of the exact frequency of the sinusoid.

2.2 Drawbacks

The simple phase update process outlined above exploits the fact that DFT frequency bins are uniformly distributed along the frequency dimension. On the other hand, as discussed in Section 1, the rigid time-frequency resolution of the DFT is known to be disadvantageous for processing broadband audio signals. To justify the assumption that each STFT coefficient within a frame can be assigned to one single peak determining its phase value, very long windows need to be applied at lower frequencies. However, this would result in an unacceptably low time resolution for higher frequencies, as rapid temporal changes usually occur at higher frequencies.

In the implementation outlined above we assumed the desired frequency shift Δf to be known. Obviously this is hardly the case in pitch-shifting applications as frequencies usually need to be scaled by a factor α rather than shifted by a constant frequency Δf. Hence, for each spectral peak the distinct frequency shift corresponding to the desired scaling factor needs to be determined. To avoid relative detuning between different sources or partials in this process, the instantaneous frequencies of spectral peaks need to be estimated that neutralizes the benefit of the simplified phase update approach. Alternatively, the center frequencies of the peak bins could be employed as frequency estimates that again calls for very long STFT windows for lower frequency partials. Another drawback stemming from the linear frequency-bin spacing is the fact that in most cases the desired frequency shifts correspond to a fractional number of DFT bins. That is, usually STFT coefficients need to be interpolated to avoid detuning due to rounding of Δf_m.

[1] This paper has previously been presented at DAFx conference [17].

When frequency-domain pitch shifting is applied to the entire input signal rather than individual sinusoidal components, these issues cause a considerable increase of computational complexity and decrease the overall quality of the pitch-shifted output signal.

In the remainder of this paper we will outline the implementation of a frequency-domain pitch-shifting algorithm based on the constant-Q transform in place of the STFT, providing solutions to the above described problems. The pitch-shifting algorithm we propose is based on the CQT implementation described in [12]. Hence, in the next section we will briefly summarize the basic concepts of the CQT implementation. Note that also other implementations [13] could be used as long as they meet the constraints on the time-frequency sampling scheme and vertical phase relations outlined in Section 4.

3 CONSTANT-Q TRANSFORM

To implement frequency-domain pitch shifting using the CQT toolbox we described in [12] it is not necessary to be aware of all implementation details. Therefore here we will only give an overview of the basic properties of the CQT, introduce user definable parameters, and mention implementation aspects that are crucial for understanding how pitch shifting can be performed using the CQT representation.

3.1 Signal Model

The CQT transform $X^{\mathsf{CQ}}(k, n)$ of a discrete time-domain signal $x(n)$ is defined by

$$X^{\mathsf{CQ}}(k, n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j - n + N_k/2) \tag{3}$$

where $k = 1, 2, \ldots, K$ indexes the frequency bins of the CQT, $\lfloor \cdot \rfloor$ denotes rounding toward negative infinity, and $a_k^*(n)$ denotes the complex conjugate of $a_k(n)$. The basis functions $a_k(n)$ are complex-valued waveforms, here also called time-frequency *atoms*, and are defined by

$$a_k(n) = \frac{1}{C} w\left(\frac{n}{N_k}\right) \exp\left[i\left(2\pi n \frac{f_k}{f_s} + \Phi_k\right)\right] \tag{4}$$

where f_k is the center frequency of bin k, f_s denotes the sampling rate, and $w(t)$ is a continuous window function (here we use the Hann window), sampled at points determined by $\frac{n}{N_k}$. The window function is zero outside the range $t \in [0, 1]$. Φ_k is a phase offset and $\Phi_k = 0$ for the transform we proposed in [12]. The scaling factor C is given by

$$C = \sum_{l=-\lfloor N_k/2 \rfloor}^{\lfloor N_k/2 \rfloor} w\left(\frac{l + N_k/2}{N_k}\right). \tag{5}$$

The window lengths $N_k \in \mathbb{R}$ in (3)–(5) are real-valued and inversely proportional to f_k in order to have the same Q-factor for all bins k. Since a bin spacing corresponding to the equal temperament[2] is desired, the center frequencies f_k obey

$$f_k = f_1 2^{\frac{k-1}{B}} \tag{6}$$

where f_1 is the center frequency of the lowest-frequency bin (due to the desired logarithmic frequency resolution there is no DC bin), and B determines the number of bins per octave. For $B = 12$ each CQT bin corresponds to one semitone, however, higher values are usually appropriate (in the audio examples, we used $B = 48$). In practice, B is the most important parameter of choice when using the CQT, because it determines the time-frequency resolution trade-off. The corresponding window lengths $N_k \in \mathbb{R}$ are given by

$$N_k = \frac{f_s}{f_k(2^{\frac{1}{B}} - 1)}. \tag{7}$$

It is not computationally reasonable to calculate the coefficients $X^{\mathsf{CQ}}(k, n)$ at all positions n of the input signal. To enable signal reconstruction from the CQT coefficients, successive atoms can be placed H_k samples apart ("hop size"). In order to analyze all parts of the signal properly and to achieve reasonable signal reconstruction, values $0 < H_k \lessapprox \frac{1}{2} N_k$ are meaningful.

3.2 Efficient Computation

Since the direct evaluation of (3) is quite expensive computationally we reduced the complexity by computing the CQT coefficients in the frequency domain and performed the CQT separately for each octave. To understand the data structure that is produced by the CQT only the latter is important. Fig. 1 shows an overview of the octave-by-octave computation of the CQT transform, downsampling the input signal by factor 2 when proceeding to the next octave. To facilitate accessing and manipulating CQT coefficients we used a constant hop size $H_k = H$ for all k within the one-octave CQT. Fig. 2 shows the resulting sampling of CQT bins in the time-frequency domain. Note that although all frequency bins within an octave are equally sampled, the window lengths N_k are unique for each frequency bin as defined by (7).

3.3 Inverse Transform

Using the toolbox presented in [12] the original signal can be efficiently reconstructed from its CQT coefficients with reasonable quality applying the processing scheme depicted in Fig. 1 in reverse order. The signal to noise ratio (SNR) between the original signal and the reconstruction error depends on the CQT resolution, the shape of window function $w(n)$ and the redundancy of the transform.[3] For

[2] An equal temperament is a musical temperament where every pair of adjacent notes has an identical frequency ratio.

[3] The redundancy factor $\Upsilon = \frac{2C_{\mathrm{CQT}}}{C_{\mathrm{IN}}}$, where C_{CQT} is the number of CQT coefficients and C_{IN} is the number of input samples.

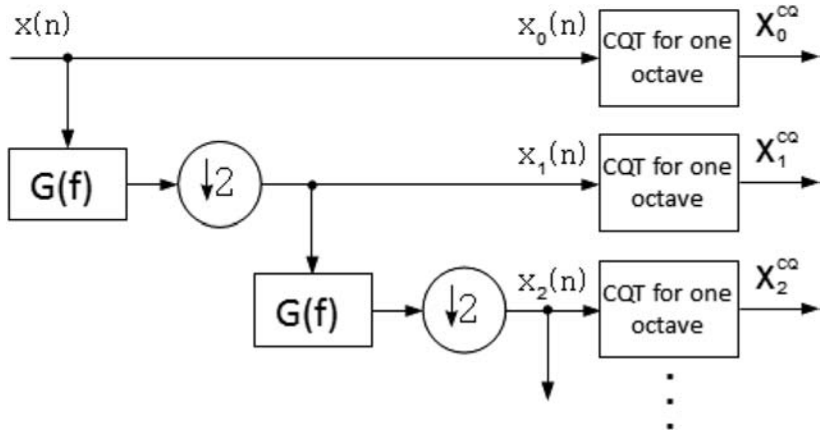

Fig. 1. An overview of computing the CQT one octave at the time. Here $G(f)$ is a lowpass filter and $\downarrow 2$ denotes downsampling by factor two.

$B = 48$ and redundancy factors around four or five, the reconstruction quality is around 55 dB SNR.

4 CQT-BASED FREQUENCY-DOMAIN PITCH SHIFTING

Given the efficient and invertible implementation of the constant-Q transform outlined above, we will now describe how pitch shifting based on the CQT representation of the input signal can be implemented with ease.

The frequency of a spectral peak in the CQT representation can be scaled by a factor α by translating (shifting) the corresponding CQT coefficient by r CQT bins. For the CQT resolution B (bins per octave) the shift in CQT bins is given by $r = B\log_2(\alpha)$, where r is independent of the frequency of the spectral peak. Furthermore, for the case of chromatic pitch transpositions or transpositions by a fraction of a semitone (e.g., $\frac{1}{8}$-tones for $B = 48$), respectively, r is an integer and no coefficient interpolation is needed. In the following chromatic pitch transpositions will be discussed, however, using simple interpolation arbitrary pitch-scaling factors can be implemented.

4.1 Time-Frequency Sampling Grid

The feasibility of shifting CQT coefficients up or down in frequency, however, does not only depend on the placement of sampling points in the frequency domain but also on their placement in time domain. In Fig. 3a an exemplary sampling grid of the time-frequency plane is depicted that produces minimal redundancy of the CQT representation while still being invertible. Such sampling grids are exhibited by CQT implementations where the hop size H_k from one atom to the next (along the time axis) is strictly increasing for increasing frequency-bin indices k. That is,

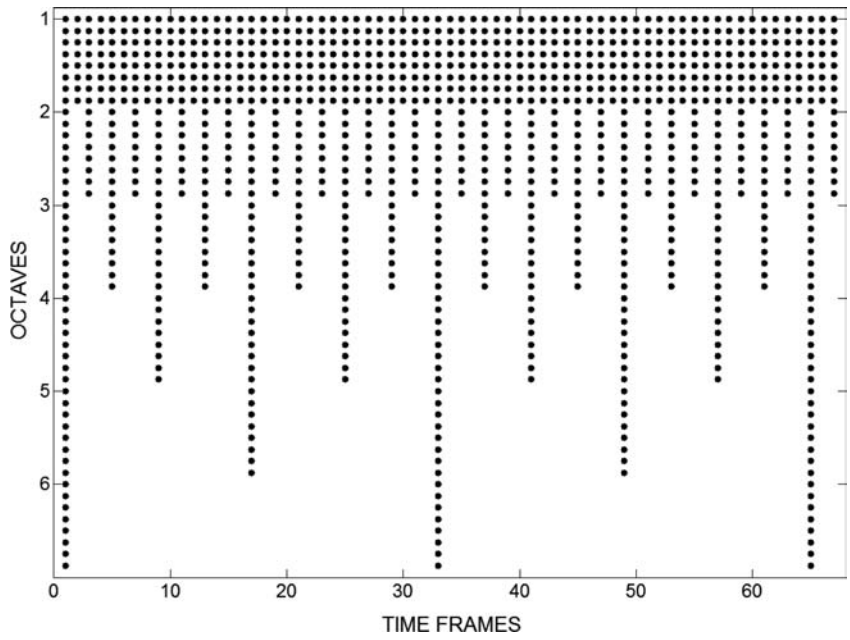

Fig. 2. Points in the time-(log-)frequency plane where $X^{CQ}(k, n)$ is evaluated.

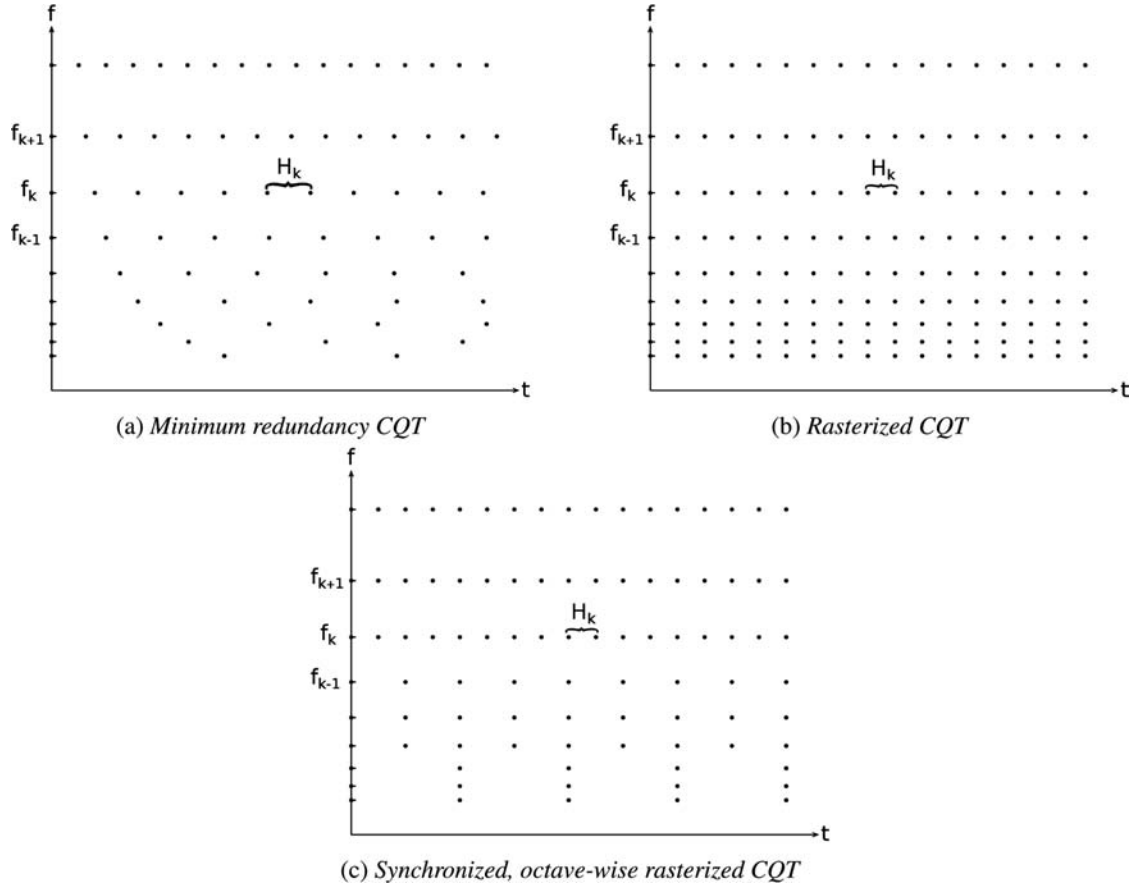

Fig. 3. Different CQT time-frequency sampling schemes. f_k denotes the k^{th} bin center frequency in Hz and H_k is the hop size for channel k in samples.

the overlap factor between successive window functions in time domain is constant while window lengths N_k are decreasing with increasing k. Although producing minimum redundancy, in Fig. 3a it can be observed that CQT coefficients cannot be shifted along the frequency dimension without changing their position in time (except for shifts corresponding to one octave). This also means that coefficients need to be skipped or interpolated since the number of coefficients changes from one frequency channel to the next.

One way to overcome the above problem is to use a "rasterized" CQT representation where the hop sizes $H_k = H_K$, $\forall k \in \{1, 2, \ldots, K\}$, that is the hop sizes for all center frequencies are set to the smallest hop size in the representation. A time-frequency sampling grid thus obtained is depicted in Fig. 3b. This sampling scheme yields invertible CQT representations where coefficients can be arbitrarily shifted along the frequency dimension.[4] A major drawback of this sampling scheme, however, is that it produces a highly redundant CQT representation, e.g., for typical transform settings ($B = 48$, $f_1 = 43$ Hz, $f_K = 22050$ Hz, $f_s = 44100$ Hz) the redundancy increases by factor 6.3 compared to the minimum redundancy sampling scheme depicted in Fig. 3a.

The time-frequency sampling scheme we proposed in [12] can be seen as the middle ground between minimum redundancy and feasibility of coefficient shifts. As discussed in Section 3.2 the atom hop sizes H_k are set according to the highest frequency bin *within each octave*, that is, for each octave down H_k is multiplied by 2. The time-frequency sampling grid thus achieved is depicted in Fig. 2 and is reproduced in Fig. 3c for the sake of clarity. It can be observed that all CQT coefficients in this representation are temporally aligned, enabling coefficient shifts along the frequency dimension without altering their position in time. For typical transform settings, the redundancy of this CQT representation is only 1.4 times higher than the optimal value. This CQT representation, however, can only be used for pitch-shifting factors smaller than 1, that is, for coefficients shifts toward lower frequencies. For coefficient shifts toward higher frequencies errors due to missing time-frequency sampling points would occur around octave boundaries. To avoid these errors we propose to use an upsampled CQT representation where the number of sampling points in each frequency channel is doubled in all except the highest octave, without changing the window lengths N_k. That increases the redundancy by factor 1.5, ending up with a redundancy of the representation that is 2.1 times higher

[4]This sampling scheme is optional in the CQT implementations [12] and [13].

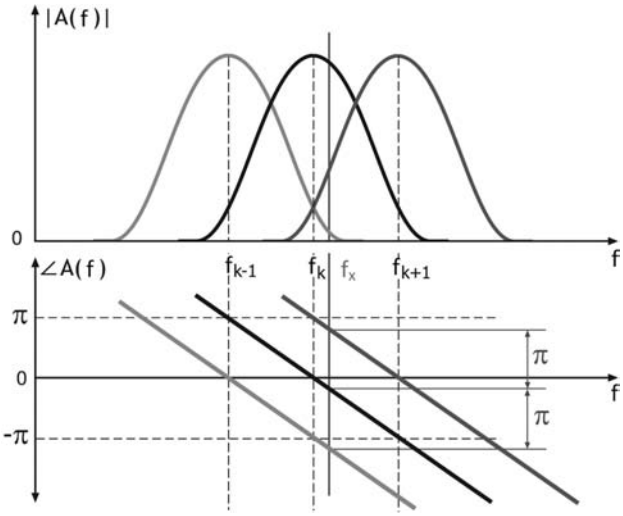

Fig. 4. Continuous magnitude and phase spectra of three adjacent windowed DFT basis functions.

than the optimal value and enables pitch-shifting factors up to 2. If even larger factors are desired, higher upsampling factors have to be applied.

4.2 Phase Coherence

Phase coherence from one time instant to the next is lost when CQT coefficients are shifted along the frequency dimension. Therefore, we need to introduce a phase update stage as is the case with all phase-vocoder-based approaches.

4.2.1 Vertical Phase Coherence

The phase locking scheme to retain vertical (within-frame) phase coherence is based on the assumption that the phase relationships between a peak bin and its neighbors are invariant under a frequency shift. For a constant-frequency sinusoid, this assumptions holds for the STFT representation in the absence of interfering signal components. The reason for this property of the STFT is that all DFT atoms are of equal lengths and are all centered at the exact same point within a frame. Hence, the group delays[5] of all DFT bins (band-pass filters) are constant and all equal. This implies that two neighboring DFT bins that are excited by the same sinusoid (within their main-lobes) will always exhibit a phase difference, which is independent of the sinusoid's frequency but is only determined by the common group delay. This fact can be appreciated from Fig. 4, where $A(f)$ is the continuous Fourier transform of the windowed basis function $a^{DFT}(n)$, f_k denotes the center frequency of bin k, and f_x is the frequency of the sinusoidal input signal. In this sketch frame-centered windows are assumed, hence the phase difference between neighboring bins is exactly π. Often it is preferred that neighboring bins excited by the same sinusoid share the same phase value. This is easily achieved by swapping the first and second half of the analysis frame after having windowed it with $w(n)$.

To establish the same property for the CQT, we have to ensure that all CQT atoms corresponding to the same time instance (*atom stack*) exhibit equal group delays. Hence, the CQT atoms in (2) need to meet two constraints: First, the (symmetric) continuous window function $w(t)$ has to be sampled so that there exists a sample N_c that is located exactly at the window center. Second, the phases of the CQT transform basis functions (atoms) $a_k(n)$ have to satisfy

$$\measuredangle a_k(N_c) \stackrel{!}{=} \text{const} \tag{8}$$

for all supported k. Both conditions are met when the CQT transform is implemented according to (3)–(4); however, since N_k is unique for each k in practice they are easily violated. For example, if a standard implementation of the (symmetric) Hann window is used the first condition is violated since odd length windows sample the continuous Hann window at the window center and even length windows sample the continuous Hann window around the center. Furthermore, standard window implementations violate the condition in (8) as they do not allow fractional window lengths.

That is, window functions need to be implemented that support fractional window lengths and exact window-center placement for any N_k in order to meet these conditions. Hence, we propose to use modified window functions that support arbitrary window lengths and sampling of the window center for all N_k. An implementation of a discrete-time Hann window thus modified is given by

$$w[n] = 0.5\left(1 - \cos\left(\frac{2\pi g_N[n]}{N}\right)\right) \tag{9}$$

where $N \in \Re^+$ is the window length, n is an integer and $0 \le n \le 2\lfloor\frac{N}{2}\rfloor$. $g[n]$ is a function that defines where the continuous Hann window is sampled and

$$g_N[n] = \frac{N}{2} - \left\lfloor\frac{N}{2}\right\rfloor + n. \tag{10}$$

In Fig. 5 four modified Hann windows with different window lengths are depicted. It can be observed that fractional window lengths are supported and that all windows can be exactly stacked at a common center. Note that due to the modified sampling of the continuous Hann window, $\omega[n]$ is always defined for $2\lfloor\frac{N}{2}\rfloor + 1$ samples.

Using the window implementation in (9)–(10), all CQT bins corresponding to the same time instance will exhibit equal group delays. For convenience, it is desirable that an impulse at the center of an atom stack exhibits real valued (zero-phase slope) CQT coefficients. This can be achieved by setting the phase offset Φ_k in (4) such that $\measuredangle a_k(N_c) = 0$, hence

$$\Phi_k = -\pi N_k \frac{f_k}{f_s} \tag{11}$$

Applying atoms $a_k(n)$ thus implemented, neighboring CQT bins excited by the same sinusoid (within their

[5]For the discrete-time discrete-frequency case, the group delay $\tau = -\Delta\Phi(f)/\Delta f$ where $\Delta\Phi(f)$ is the phase difference between two neighboring bins and Δf frequency difference between them [19].

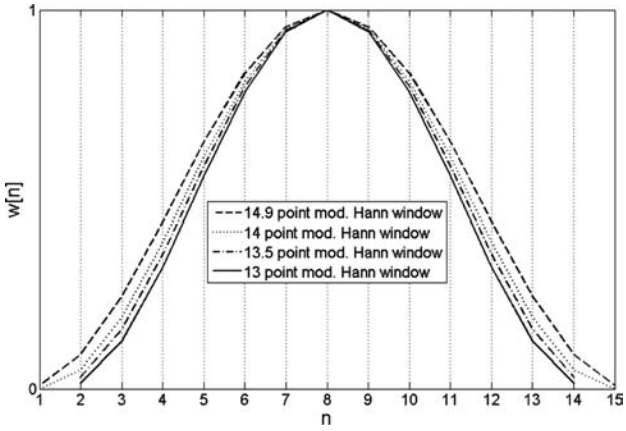

Fig. 5. Modified Hann windows with different (fractional) lengths centered at a common time instance.

main-lobes) will exhibit equal phase values and vertical phase coherence can be retained by phase-locking the translated CQT coefficients. This is done by setting the phases of all CQT coefficients within the region of influence of a peak to the peak phase.

4.2.2 Horizontal Phase Coherence

In Section 2 we outlined a simple phase update approach to retain horizontal phase coherence for DFT-based pitch shifting. Here we compare the DFT- and CQT-based approach concerning the frame-to-frame phase update process to point out the differences.

Let's assume that the input signal consists of only one constant-frequency sinusoid with frequency f_1. The center frequency of the corresponding peak bin is $\hat{f}_1$. The phase difference $\Delta\phi_1$ between two transform coefficients of consecutive time frames $u-1$ and u is given by $\Delta\phi_1 = 2\pi R\frac{f_1}{f_s}$, where R is the frame hop size and f_s is the sampling frequency.

DFT case: If we shift the entire input signal (or the region around the sinusoid) up by r DFT bins, the frequency of the sinusoid after the shift is $f_2 = r\frac{f_s}{N} + f_1$ and the center frequency of the corresponding peak bin is $\hat{f}_2 = r\frac{f_s}{N} + \hat{f}_1$, where N is the DFT size. The phase difference $\Delta\phi_2$ between two consecutive time frames after the shift is given by $\Delta\phi_2 = 2\pi R\frac{f_2}{f_s}$. In order to account for this, we need to add (cumulatively from frame to frame) a phase value Φ_{DFT} to each coefficient in time frame u, where

$$\Phi_{DFT} = \Delta\phi_2 - \Delta\phi_1 = \frac{2\pi R}{f_s}(f_2 - f_1) = \frac{2\pi R}{f_s}\Delta f. \tag{12}$$

This can be implemented by simply multiplying each coefficient by the complex

$$Z_u = e^{i\Phi_{DFT}}. \tag{13}$$

This means that in order to update the phase values we do not need to know the exact frequency of the shifted sinusoid. The phase value that needs to be added to each coefficient in time frame u is only determined by the frequency difference Δf and due to the linear frequency-bin spacing of the DFT

$$\Delta f = f_2 - f_1 = \hat{f}_2 - \hat{f}_1 = r\frac{f_s}{N} \tag{14}$$

and

$$\Phi_{DFT} = 2\pi r\frac{R}{N}. \tag{15}$$

That is, in the DFT domain pitch shifting a sinusoid by a given frequency shift Δf does not require instantaneous frequency estimation. However, in practice spectral components are not translated by a certain frequency but their frequency is scaled by a given factor. Hence the instantaneous frequency needs to be estimated in order do determine the desired frequency shift. Alternatively, the center frequency of the peak bin can be used as a frequency estimate, however, especially for small DFT sizes and low frequencies artifacts due to detuning will be introduced.

CQT case: If we shift the entire input signal (or the region around the sinusoid) up by r CQT bins, the frequency of the sinusoid after the shift is $f_2 = f_1 2^{r/B}$ and the center frequency of the corresponding peak bin is $\hat{f}_2 = \hat{f}_1 2^{r/B}$ where B is the CQT resolution in bins per octave. The phase difference $\Delta\phi_2$ between two consecutive time frames after the shift is given by $\Delta\phi_2 = 2\pi R\frac{f_2}{f_s}$. In order to account for this, we need to accumulate-add a phase value Φ_{CQT} to each coefficient, where

$$\Phi_{CQT} = \Delta\phi_2 - \Delta\phi_1 = \frac{2\pi R}{f_s}(f_2 - f_1) = \frac{2\pi R}{f_s}\Delta f. \tag{16}$$

Again we only need to know Δf to correctly perform phase update. However, in the CQT case

$$\Delta f = f_2 - f_1 = f_1\left(2^{r/B} - 1\right) \neq \hat{f}_2 - \hat{f}_1 \tag{17}$$

due to the geometrical frequency-bin spacing. That is, in the CQT domain we can achieve pitch shifting by a given factor by simply shifting all coefficients by r CQT bins, however, unlike the DFT case, Φ_{CQT} is not fully determined by r. Consequently, in order to correctly update the phase values in horizontal direction, we need to estimate the instantaneous frequency f_1 [2]. Alternatively, $\hat{f}_1$ can be used as an approximate for f_1, that is

$$\Phi_{CQT} \approx \frac{2\pi R}{f_s}\hat{f}_1\left(2^{r/B} - 1\right). \tag{18}$$

This approximation introduces slight frequency and amplitude modulations in the output signal, however, informal listening tests suggest that these errors are hardly (if at all) audible. One explanation for this is that according to [20], the human hearing is insensitive to amplitude and frequency modulations below certain thresholds.[6] Since the quality of the output signal is hardly degraded when frequency

[6]This fact has also been exploited in [21] to reduce phasiness for time-scaling applications of STFT-based phase vocoders.

estimation is omitted,[7] we recommend to do so as this improves not only computational efficiency but also robustness.

4.3 Implementation

Having described the general idea of CQT-based pitch shifting and the minor changes that we applied to the CQT toolbox, we will now outline how music signals can be transposed using the CQT:

1. Transform: Using an oversampled CQT representation (see Section 4.1), the input signal can be transposed in the range of ± 1 octave. The CQT coefficients are presented in the sparse matrix M^{CQT} as depicted in Fig. 2. The audio examples were generated using a CQT resolution of 48 bins per octave.
2. Translating CQT coefficients: The entire input signal is transposed by rotating (shifting) the rows of M^{CQT} up- or downwards. Using $B = 48$ a matrix shift by one row corresponds to a transposition of 25 cents (quarter of a semitone).
3. Phase update: Using a simple peak-detector for all valid CQT coefficients within the columns of M^{CQT}, each column (frame) is divided into several regions of influence. Horizontal and vertical phase coherence can be retained by multiplying all coefficients within the same region with the complex $Z_u = e^{i\,\Delta f_{k,u} H_k}$ where $\Delta f_{k,u}$ is the difference between the center frequencies of the old and the new peak bin in column u of the CQT and H_k is the atom hop size. The applied phase rotations need to be accumulated from one frame to the next.
4. Inverse Transform: Reconstruct the transposed output signal from the processed CQT representation.

5 RESULTS

In [18] we provide several audio examples to demonstrate the performance for pitch-shifting factors in the range of ± 1 octave for different music genres. From these samples it can be appreciated that, despite the very simple implementation (no instantaneous frequency estimation, no peak following [3], no trajectory heuristics [5], no transient detection [22]), the proposed algorithm produces very little artifacts for a wide range of scaling factors. Due to the logarithmic frequency resolution of the CQT, relative detuning between partials is avoided and closely spaced sinusoidal components at lower frequencies can be distinguished. In [5] a STFT-based phase vocoder for time scaling has been proposed that includes a frequency-dependent peak-picking stage to reduce artifacts in the output signal. Due to the geometrical bin spacing, frequency-dependent peak picking is inherent in the proposed CQT-based pitch-shifting technique.

As discussed above a major benefit of the proposed method is that polyphonic music signals can be transposed in the time-frequency domain by simply shifting the entire CQT up- or downwards (followed by a phase-update stage). An advantage of time-frequency-domain pitch-shifting approaches in general is that individual signal components (e.g., single notes) can be transposed in polyphonic music signals while leaving others unchanged. The CQT-based approach is specifically interesting for this application since harmonic structures can be easily detected (distribution of partials does not depend on the fundamental frequency) and interference among fundamental frequencies in lower frequency areas is reduced.

5.1 Transients

In general transients pose a problem for all phase-vocoder-based time- and pitch-scaling methods as the phase update process only considers slowly varying sinusoidal components, hence transients are usually softened (smeared) due to loss of vertical phase coherence at transients. For time-scaling algorithms based on the STFT phase vocoder different solutions to this problem have been suggested in the past. In [22] a transient preservation technique is proposed where phases are reset to their original values at transients. Another approach [3] suggests to set the time-scaling factor to 1 in transient regions (this is compensated by increasing the time-scaling factor in steady-state parts). Both approaches need to rely on a transient detection stage of some kind.

The CQT-based pitch-shifting approach mitigates problems with transients by providing a very good time resolution at high frequencies. Therefore transients are preserved simply due to the high time resolution of the magnitude CQT spectrum, without the need to encode the transients in vertically synchronous phase information. However, at low frequencies atom lengths N_k get very wide and the lack of vertical phase coherence gets increasingly audible toward lower frequencies. That is, low-frequency transients (e.g., electric/upright bass notes, bass drum hits) call for dedicated processing to reduce low frequency transient smearing.

As for the STFT case, this could be achieved by detecting transients and realigning the phase in transient regions [22]. To avoid the need for explicit transient detection, however, as transients do not carry tonal information we suggest to subtract lower frequency transients from the signal prior to pitch shifting and add these signal portions again after pitch shifting. Essentially this is a classic source separation problem known from singing voice separation [23] or note separation [24] tasks that in this case boils down to a general percussive/harmonic separation task. In [25] a simple and efficient percussive/harmonic separation approach has been proposed where transients are regarded as outliers along the time dimension and harmonic components are regarded as outliers along the frequency dimension in the STFT representation of the input signal (the approach proposed in [26] is based on a similar notion). Percussive and harmonic signal components are separated by applying 1-dimensional median filters both along the frequency- and the time dimension.

[7] Formal listening tests to prove this result are yet to be conducted.

To subtract low-frequency transients from the input signal we implemented a straightforward CQT adaptation of the technique proposed in [25] that only considers CQT coefficients below the frequency threshold f_{th} since higher-frequency transients do not need to be processed. In [18] we provide audio examples to demonstrate that using this efficient approach, low-frequency transients can be retained in the pitch-shifted output signal.

5.2 Limitations of the Method

To obtain a natural-sounding pitch-shifted output signal it is desired to retain the formant structure of the original as formants are independent of the fundamental frequencies. A common approach is to model the spectral envelope, use this model to flatten the magnitude spectrum of the input signal, and apply the original spectral envelope to the pitch-shifted signal. Several techniques to gain an estimate of the spectral envelope have been proposed, most prominently approaches based on linear prediction [27] or the real cepstrum [28].

Another approach is to apply pitch shifting in the time domain, where well-established techniques such as pitch-synchronous overlap-add produce rather natural quality by assuming that only one speaker is active at a time [8].

Up to this point we have not included any formant preservation technique in the CQT-based pitch-shifter, that is, the audio samples we provide in [18] are lacking naturalness (especially when singing voice or speech is considered). However, since formants approximately feature the constant-Q property (i.e., the bandwidth of high frequency formants is wider than for lower frequency formants), we expect the CQT-based pitch-shifting algorithm to exhibit some advantageous qualities for formant preservation approaches in future implementations.

For speech signals, a loss of naturalness does not only occur due to formant shifts but annoying artifacts are introduced due to lack of vertical phase coherence among partials. A possible explanation for these artifacts is the fact that the shape of the underlying glottal pulses changes when the phase relations between partials are altered. In [6] a shape-invariant phase vocoder for speech transformation is proposed that reduces these artifacts. We have not included techniques to retain inter-partial phase coherence, hence when speech signals are transposed applying large scaling factors, audible artifacts are introduced.

6 CONCLUSIONS

A frequency-domain pitch-shifting method was proposed that exploits the logarithmic frequency-bin spacing of the CQT. The presented technique enables pitch-scaling of monophonic and dense polyphonic music signals by applying a simple linear translation of the CQT representation followed by a phase update stage. High-quality pitch transpositions with large scaling factors can be achieved without estimating instantaneous frequencies of partials.

Use of the CQT representation provides natural solutions to some of the basic issues in pitch shifting: problems with transients are mitigated due to the very good time resolution of the CQT at high frequencies, while interference between tonal components at low frequencies is reduced due to the high frequency resolution at low frequencies.

Performing pitch shifting directly in the frequency domain allows one to process only parts of the signal while leaving other parts untouched. Another advantage is that computational complexity does not depend on the scaling factor, contrary to the time-stretching/resampling approach. Audio examples have been provided to demonstrate the achieved quality of the proposed algorithm for different scaling factors up to ± 1 octave.

7 REFERENCES

[1] E. Coyle, D. Dorran, and R. Lawlor, "A Comparison of Time-Domain Time-Scale Modification Algorithms," presented at the *120th Convention of the Audio Engineering Society* (2006 May), convention paper 6674.

[2] J. Laroche and M. Dolson, "Improved Phase Vocoder Time-Scale Modification of Audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332 (1999).

[3] J. Bonada, "Automatic Technique in Frequency Domain for Near-Lossless Time-Scale Modification of Audio," in *Proc. of International Computer Music Conference* (2000).

[4] D. Dorran and R. Lawlor, "Time-Scale Modification of Music Using a Synchronized Subband/Time-Domain Approach," in IEEE International Conference on Acoustis, Speech and Signal Processing (2004).

[5] T. Karrer, E. Lee, and J. Borchers, "Phavorit: A Phase Vocoder for Real-Time Interactive Time-Stretching," in *Proc. International Computer Music Conference (ICMC)* (2006), pp. 708–715.

[6] A. Röbel, "A Shape-Invariant Phase Vocoder for Speech Transformation," in *Proc. Digital Audio Effects (DAFx-10)*(2010).

[7] J. Laroche and M. Dolson, "New Phase-Vocoder Techniques Are Real-Time Pitch Shifting, Chorusing, Harmonizing, and Other Exotic Audio Modifications," *J. Audio Eng. Soc*, vol. 47, pp. 928–936 (1999 Nov.).

[8] E. Moulines and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," *Speech Communication*, vol. 9, no. 5, pp. 453–467 (1990).

[9] J. Laroche, "Autocorrelation Method for High-Quality Time/Pitch-Scaling," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE (1993), pp. 131–134.

[10] M. Vetterli and C. Herley, "Wavelets and Filter Banks: Theory and Design," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 40, no. 9, pp. 2207–2232 (1992).

[11] J. C. Brown, "Calculation of a Constant Q Spectral Transform," *J. Acous. Soc. America*, vol. 89, no. 1, pp. 425–434(1991).

[12] C. Schörkhuber and A. Klapuri, "Constant-Q Transform Toolbox for Music Processing," in *Proc. Sound and Music Computing Conference (SMC)* (2010).

[13] G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an Invertible Constant-Q Transform with Nonstationary Gabor Frames," in *Proc. Digital Audio Effects (DAFx-11)* (2011).

[14] J. Kovacevic and A. Chebira, "Life Beyond Bases: The Advent of Frames (Part i)," *Signal Processing Magazine, IEEE*, vol. 24, no. 4, pp. 86–104 (2007).

[15] P. Balazs, M. Dörfler, F. Jaillet, N. Holighaus, and G. Velasco, "Theory, Implementation and Applications of Nonstationary Gabor Frames," *J. Computational and Applied Mathematics*, pp. 236:1481–1496 (2011).

[16] I. W. Selesnick, "Wavelet Transform with Tunable Q-Factor," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3560 (2011).

[17] C. Schörkhuber, A. Klapuri, and A. Sontacchi, "Pitch Shifting of Audio Signals Using the Constant-Q Transform," in *Proc. Digital Audio Effects (DAFx-12)* (2012).

[18] C. Schörkhuber, "Pitch Shifting Using the CQT: Audio Examples," Available at http://www.iem.at/ schoerkhuber/cqt/, accessed July 09, 2012.

[19] B. Boashash, *Time Frequency Signal Analysis and Processing: A Comprehensive Reference* (Elsevier, 2003).

[20] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models, 2nd Edition* (Springer Heidelberg, 1999).

[21] D. Dorran, E. Coyle, and R. Lawlor, "An Efficient Phasiness Reduction Technique for Moderate Audio Time-Scale Modification," in *Proc. Digital Audio Effects (DAFx-04)* (2004), pp. 83–88.

[22] A. Röbel, "A New Approach to Transient Processing in the Phase Vocoder," in *Proc. Digital Audio Effects (DAFx-03)*(2003).

[23] S. Sofianos, A. Ariyaeeinia, R. Polfreman, and R. Sotudeh, "H-Semantics: A Hybrid Approach to Singing Voice Separation," *J. Audio Eng. Soc*, vol. 60, pp. 831–841 (2012 Oct.).

[24] D. Gunawan and D. Sen, "Separation of Harmonic Musical Instrument Notes Using Spectro-Temporal Modeling of Harmonic Magnitudes and Spectrogram Inversion with Phase Optimization," *J. Audio Eng. Soc*, vol. 60, pp. 1004–1014 (2012 Dec.).

[25] D. Fitzgerald, "Harmonic/Percussive Separation Using Median Filtering," in *Proc. Digital Audio Effects (DAFx-10)*(2010).

[26] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," in *Proc. EUSIPCO* (2008).

[27] J. E. Markel and A. H. Gray, *Linear Prediction of Speech* (Springer-Verlag New York, Inc., 1982).

[28] A. Röbel and X. Rodet, "Efficient Spectral Envelope Estimation and its Application to Pitch Shifting and Envelope preservation," in *Proc. Digital Audio Effects (DAFx-05)* (2005).

THE AUTHORS

Christian Schörkhuber

Anssi Klapuri

Alois Sontacchi

Christian Schörkhuber received his Dipl. Ing. degree from Graz University of Technology and University of Music and Performing Arts, Graz, Austria, in 2012. He visited as an undergraduate researcher at the Centre for Digital Music at Queen Mary University of London, London, U.K., in 2010. In 2012 he joined the Institute of Electronic Music and Acoustics (IEM) at University of Music and Performing Arts, Graz, Austria, as a research fellow and was a visiting researcher at Tampere University of Technology, Tampere, Finland in 2013. Currently he is persuing his PhD at IEM in spatial audio, music information retrieval and audio signal processing.

●

Anssi Klapuri (M'06) received his Ph.D. degree from the Tampere University of Technology (TUT), Tampere, Finland, in 2004. He visited as a post-doc researcher at Ecole Centrale de Lille, France, and Cambridge University, UK, in 2005 and 2006, respectively. He worked as a lecturer at the Centre for Digital Music at Queen Mary University of London, London, UK, in 2010–2011. He is currently CTO at Ovelin, Finland, and Associate Professor at TUT. His research interests include audio signal processing, auditory modeling, and machine learning.

●

Alois Sontacchi (Dipl. Ing. '99) received his Ph.D. from Graz University of Technology, Austria, in 2003. From 2003 to 2009 he was a research assistant at the Institute of Electronic Music and Acoustics (IEM) at University of Music and Performing Arts, Graz, Austria. Since 2009 he is a senior scientist at IEM and head of the institute since 2010. His research interests include psychoacoustical perceptual evaluation and test design, information retrieval, as well as spatial and auditory-motivated audio signal processing.