



# A Real-Time Variable-Q Non-Stationary Gabor Transform for Pitch Shifting

Dong-Yan Huang, Minghui Dong and Haizhou Li

Human Language Technology Department, Institute for Infocomm Research/A\*STAR  
 #21-01 Connexis (South Tower), Singapore 138632

{huang, mdong, hli}@i2r.a-star.edu.sg

## Abstract

This paper proposes a real-time variable-Q non-stationary Gabor transform (VQ-NSGT) system for speech pitch shifting. The system allows for time-frequency representations of speech on variable-Q (VQ) with perfect reconstruction and computational efficiency. The proposed VQ-NSGT phase vocoder can be used for pitch shifting by simple frequency translation (transposing partials along the frequency axis) instead of spectral stretching in frequency domain by the Fourier transform. In order to retain natural sounding pitch shifted speech, a hybrid of smoothly varying Q scheme is used to retain the formant structure of the original signal at both low and high frequencies. Moreover, the preservation of transients of speech are improved due to the high time resolution of VQ-NSGT at high frequencies. A sliced VQ-NSGT is used to retain inter-partial phase coherence by synchronized overlap-add method. Therefore, the proposed system lends itself to real-time processing while retaining the formant structure of the original signal and inter-partial phase coherence. The simulation results showed that the proposed approach is suitable for pitch shifting of both speech and music signals.

**Index Terms:** Time-frequency representation, perfect reconstruction, constant-Q transform, variable-Q transform, non-stationary Gabor transform, real-time pitch shifting

## 1. Introduction

Pitch shifting is one of the most popular digital audio effects that shift the pitch of sound without changing its duration. It means that all frequencies are raised or reduced by a constant factor. The pitch shifting technology can be found applications such as voice transformation, pitch correction in an audio recording or performance, and transposing songs to desired keys [1, 2, 3, 4].

The pitch-scale modification of speech and music signals can be achieved by a two-stage process: the time scaling and sampling rate conversion. The standard time scaling can be operated either in the time domain [5] or in the time-frequency domain [6]. An alternative approach is to shift the pitch of sound directly without a time scaling stage. Most of algorithms are based on the phase vocoder [7] and on synchronous overlap-add (SOLA) [8, 9], where the pitch shifting in time domain is usually efficient. The phase-vocoder algorithms can achieve higher quality for both speech and music signals, however suffering from the artifacts. The problems stem from the loss of "horizontal" phase coherence between frames and loss of "vertical" phase coherence within frames [10]. The solution to solve these problems is to estimate the instantaneous frequency and phase un-wrapping to establish phase coherence between frames and a phase-locking scheme to maintain the phase coherence within frame. In spite of the above challenges, there are several advantages of the pitch shifting approach against the two-stage

time-scaling/resampling process: the computational complexity is independent of the scaling factor and sinusoidal components can be shifted independently.

The constant-Q transform (CQT) can provide a solution to all of the inconveniences of the STFT-based pitch shifting approach. Constant-Q transform aims at decomposing an input signal into the time-frequency domain so that the center frequencies of the frequency bins are geometrically spaced in a constant Q factor in logarithmic scale. Recently a new time-frequency (TF) transform called constant-Q non-stationary Gabor transform (CQ-NSGT) provides auditory TF resolution for audio representations [11, 12, 13]. The CQ-NSGT is developed based on frame theory and can be understood as a non-uniform filterbank. It is able to construct analysis-synthesis systems with desirable properties such as invertibility, computational efficiency, and adaptable redundancy. However, the CQT suffers from low time resolution for lower frequencies. A smoothly varying Q scheme is proposed for improving the time resolution while keeping the formant structure of the original signal. A sliced CQT with 75 % overlap is proposed to reduce amplitude modulation and retain inter-phase coherence.

In this paper, we present a pitch shifting algorithm based on the variable-Q NSGT representation of speech signals. In Section 2, we give a brief presentation on the concept of phase-vocoder. In Section 3, we present CQ-NSGT phase vocoder and some crucial aspects for CQT-based pitch-shifting algorithm and drawbacks. In Section 4, we propose a sliced variable-Q non-stationary Gabor transform. In Section 5, we present how to maintain inter- and intra-phase coherence (fractional and integer shifting) in CQT for pitch shifting. In Section 6, we evaluate the audio samples through subjective listening tests to show the achieved quality of the pitch shifting voices. Finally, conclusion will be given in Section 7.

## 2. Phase Vocoder

The essential idea of phase vocoder is to assume that a signal  $f(n)$ , sampled at frequency  $\xi_s$ , is expressed as a sum of  $N$  sinusoids, called partials [14]

$$f(n) = \sum_{k \in N} a_k \cos\left(\frac{n}{\xi_s} \omega_k + \phi_k\right) \quad (1)$$

each is described by its own angular frequency  $\omega_k$ , amplitude  $a_k$ , and  $\phi_k$ . Assuming these three parameters vary relatively so slowly, quasi-stationary and pseudo-periodic of the signal (e.g., speech and music) are maintained. The idea of the STFT-based SOLA (synchronized overlap add) is to slice the signal into overlapping frames and shift the frames to reduce frequency and amplitude modulations in the output signal [7]. As the STFT frequency bins have a fixed resolution over the time-frequency (TF) plane, the TF resolutions are not suitable for broadband

10.21437/Interspeech.2015-578

audio signals because they do not fit to that of auditory system. Due to based on sinusoidal models, the STFT-based phase vocoder does not provide satisfying results for non-sinusoidal signals [15, 16]. We seek new tools for this issue.

### 3. The Constant-Q Transform (CQT) for Pitch Shifting and Its Limitations

A pitch shifting using CQT has been proposed [17]. A "rasterized" CQT representation is used. It means that all CQT coefficients are temporally aligned, enabling the CQT coefficient shifts along the frequency dimension without changing their position in time. In order to retain vertical phase coherence, window functions are modified to support arbitrary window lengths and sampling of the window center for all atoms by setting odd length of window. The vertical phase coherence can be retained by setting the phase of all CQT coefficients within the region of influence of a peak to the peak phase. A simple phase update approach is used to retain horizontal phase coherence of CQT coefficients between frame-to-frame.

Although this CQT can achieve a satisfactory quality reconstruction (around 55 dB SNR) of a signal from its transform coefficients with an efficient computation of CQT coefficients, the CQT phase vocoder for pitch shifting has the following limitations: 1) annoying artifacts are introduced due to lack of vertical phase coherence among partials, especially for speech and singing voice; 2) as formants are independent of the fundamental frequencies, a natural-sounding pitch-shifted should preserve the formant structure of the original as formants. There is not any formant preservation technique in the actual CQT-based pitch shifter.

### 4. Sliced Variable-Q Non-stationary Gabor Transform

To address the issues of lack of vertical phase coherence and formant preservation in the CQT-based pitch shifter [17], we propose a sliced variable-Q non-stationary Gabor transform to shift pitch of sound.

#### 4.1. Sliced Constant-Q non-stationary Gabor transform

In this paper, we consider real-valued signals of length  $L$ . We denote the inner product of two discrete signals  $f, g$  is  $\langle f, g \rangle = \sum_{l=0}^L f(l)g(l)$  and the energy of a signal is defined as  $\|f\|^2 = \langle f, f \rangle$ . The Fourier transform of  $f$  is denoted by  $\mathcal{F} : f \mapsto F$ .

##### 4.1.1. Constant-Q non-stationary Gabor transform

The constant-Q transform (CQT), originally introduced by Brown [18], is characterized by the geometrically spaced and equal Q-factors of the center frequencies of bins for the time-frequency representation of a signal. Recently, an NGS system can construct a non-uniform filterbank, which resolution evolves across frequency with a set of different windows [11]

$$\mathcal{G}(\mathbf{g}, \mathbf{D}) = \{g_{n,k}[l]\} = (g_k[l - nD_k]) \quad (2)$$

where indexes  $n, k \in \mathcal{Z}$  are related to time position and frequency bin, respectively.  $g_k$  is a set of frequency-dependent filters with down-sampling factors  $D_k$ . The NSGT is developed based on frame theory. A collection  $\{g_{n,k}\}$  is a Gabor frame for  $L^2(R)$  if there exist two positive constants  $A$  and  $B$  such that

$$A\|f\|^2 \leq \sum_{n,k \in \mathcal{Z}} |\langle f, g_{n,k} \rangle|^2 \leq B\|f\|^2 \quad (3)$$

for all  $f \in L^2(R)$ . The constants  $A$  and  $B$  are called lower and upper frame bounds, respectively. The analysis coefficients  $c_{n,k} = \langle f, g_{n,k} \rangle$  are representations of the signal  $f$  and the synthesis is given by  $\hat{f} = \sum_{n,k \in \mathcal{Z}} c_{n,k} g_{n,k}$  and the frame operator  $S$  given by  $S = \sum_{n,k \in \mathcal{Z}} \langle f, g_{n,k} \rangle g_{n,k}$ . If the frame operator is invertible, the reconstruction  $f$  can be expressed with the canonical dual frame sequence  $\tilde{\mathcal{G}}(\mathbf{g}, \mathbf{D}) \{\tilde{g}_{n,k} = \mathbf{S}^{-1} g_{n,k}\}$ ,

$$f = SS^{-1}f = \sum_{n,k \in \mathcal{Z}} \langle f, g_{n,k} \rangle \tilde{g}_{n,k} \quad (4)$$

The lower and upper frame bounds of the dual frames are  $1/B$  and  $1/A$ , respectively. We are interested in Gabor frames whose windows are supported on some compact interval, where  $\tilde{g}_k$  is non-zero, denoted by  $\max \text{supp}(\tilde{g}_k) = L_k$ . If the samples in each channel satisfies the condition  $\lceil L/D_k \rceil \geq 2L_k$ , then the operator

$$\hat{S} := \mathcal{F}S\mathcal{F}^{-1} \quad (5)$$

is diagonal and invertible. This defines the painless case, where the dual window  $\tilde{g}_{n,k}$  can be easily calculated as follows

$$\tilde{g}_k = \mathbf{S}^{-1}g_k = \frac{g_k}{b^{-1} \sum_{n \in \mathcal{Z}} |g(l - nD_k)|^2} \quad (6)$$

The Eq. 2 shows the condition for the atoms under which the  $f$  can be reconstructed by just shifting the time-frequency the window  $g_k$  according to the lattice. The Eq. 6 gives the formula of calculation of  $\tilde{g}_k$  in the painless case. The analysis and synthesis can be implemented with fast FFT methods.

In practice, real-time CQ-NSGT is required for applications with bounded delay in processing inputs and low complexity. A SLICQ has been developed by judicious selection of both the slicing window  $q_m$  and the analysis windows  $g_k$  for CQ-NSGT [19]. The conditions for  $g_k$  and  $q_m$  are detailed in the following

**Theorem 1** Assuming that  $G(g, a)$  and  $G(\tilde{g}, a)$  are dual NSG systems for  $C^{2N}$ . Further let  $q_0, \tilde{q}_0 \in C^K$  satisfy

$$\sum_{m=0}^{L/N-1} (q_{m,N} \overline{\tilde{q}_{m,N}}) \equiv 1 \quad (7)$$

If  $s$  is the output of  $\text{sliCQL,N}(f, q_0, g, a)$ , then the output  $\tilde{f}$  of  $\text{isliCQL,N}(s, \tilde{q}_0, \tilde{g}, a)$  equals  $f$ , i.e.,  $\tilde{f} = f$ .

#### 4.2. Variable-Q

The CQT can be used to analog to auditory filters in the human auditory system. These filters are described by the equivalent rectangular bandwidth (ERB). The ERB (in Hz) of the auditory filter centered at frequency  $\xi_k$  is [20]

$$\text{ERB}(\xi_k) = 24.7 + \frac{\xi_k}{9.265} \quad (8)$$

Eq.(8) shows that auditory frequency resolution in ERBs are approximately constant-Q only for frequencies above 500 Hz. The full range of audible frequencies is from 2 Hz to 20kHz. The ERBs range from 27 Hz to 10 kHz [20]. The ERBlet transform has been proposed to address this issue [13], where the bin bandwidths and center frequencies correspond to the equivalent rectangular bandwidths (ERB) [20] and their corresponding frequency distribution, respectively. In order to increase the time resolution at lower frequencies, we adopt an approach for

smoothly decreasing the  $Q$ -factors of the bins towards low frequencies in [21]. The bandwidth  $\Omega_k$  of filter channel  $k$  is defined as

$$\Omega_k = \alpha \xi_k + \gamma \quad (9)$$

where  $\alpha = 1/Q = 2^{1/B} - 2^{-1/B}$  is determined by the number of bins per octave,  $b$ .  $\gamma = 0$  and  $\gamma = \Gamma$  are two special cases, constant- $Q$  and the bandwidths equal to constant fraction of the ERB and bandwidth [21]. Here

$$\Gamma = \frac{24.7}{0.108} \alpha, \Omega = \frac{\alpha}{0.108} \text{ERB} \quad (10)$$

## 5. Pitch Shifting

Considering a constant-frequency sinusoidal signal, phase coherence can be achieved if each STFT coefficient in the region of influence (peak's phase) is simply multiplied by the complex

$$Z_u = e^{j \frac{2\pi R}{\xi_s} \delta \xi_{m,u}} \quad (11)$$

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

$$Z_{u+1} = Z_u e^{j \frac{2\pi R}{\xi_s} \delta \xi_{m,u}} \quad (12)$$

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.

### 5.1. Vertical Phase Coherence

In order to retain vertical (within) phase coherence, the phase locking scheme is based on the assumption that the phase relationships between a peak bin and its neighbours are invariant under a frequency shift. For a constant-frequency sinusoid, this assumption holds for the STFT representation in the absence of interfering signal components. To establish the same property for the CQT, it is suggested to ensure that all CQT atoms corresponding to the same time instance (atom stack) exhibit equal group delays [17]. Hence, the CQT atoms need to meet two constraints: First, the (symmetric) continuous window function  $g_{k,n}$  has to be sampled so that there exists a sample  $N_k$  that is located exactly at the window center. For supporting fractional window lengths and exact window-center placement for any  $N_k$ , an implementation of a discrete-time window function thus modified is given by

$$g_{k,n} = W(E_N(n)) \quad (13)$$

where  $N \in R^+$  is the window length,  $n$  is an integer and  $0 \leq n \leq 2 \lfloor \frac{N}{2} \rfloor$ .  $E_N(n)$  is a function that defines where the continuous window function is sampled and

$$E_N(n) = \frac{N}{2} - \lfloor \frac{N}{2} \rfloor + n \quad (14)$$

$g_{k,n}$  is always defined for  $2 \lfloor \frac{N}{2} \rfloor + 1$  samples. Second, the phases of the CQT transform basis functions (atoms)  $c_{k,n}$  have to satisfy

$$L_k(N_k) \stackrel{\Delta}{=} \text{const} \quad (15)$$

for all supported  $k$ . All atoms  $g_{k,n}$  is thus implemented, neighbouring CQT bins excited by the same sinusoid (within their main-lobes) will exhibit equal phase values and vertical phase coherence can be retained by phase-locking the translated CQT coefficients by setting the phases of all CQT coefficients within the region of influence of a peak to the peak phase.

### 5.2. Horizontal Phase Coherence

We review CQT-based approach involving the frame-to-frame phase update process. For an input signal with only one constant-frequency sinusoid of frequency  $\xi_1$ , the center frequency of the corresponding peak bin is assumed as  $\hat{\xi}_1$ . The phase difference  $\Delta \phi_1$  between two transform coefficients of consecutive time frames  $u-1$  and  $u$  is given by  $\Delta \phi_1 = 2\pi R \frac{\xi_1}{\xi_s}$ .

If the entire input signal up by  $r$  CQT bins, the frequency of the sinusoid after the shift is  $\xi_2 = \xi_1 2^{r/B}$  and the center frequency of the corresponding peak bin is  $\hat{\xi}_2 = \hat{\xi}_1 2^{r/B}$ . The phase difference  $\Delta \phi_2$  between two consecutive time frames after the shift is given by  $\Delta \phi_2 = 2\pi R \frac{\xi_2}{\xi_s}$ . A phase value  $\Phi_{cqt}$  need to be accumulate-added to each coefficient, where

$$\begin{aligned} \Phi_{CQT} &= \Delta \phi_2 - \Delta \phi_1 = \frac{2\pi R}{\xi_s} (\xi_2 - \xi_1) \\ &= \frac{2\pi R}{\xi_s} \Delta \xi \end{aligned} \quad (16)$$

In order to correctly update the phase values in horizontal direction, the instantaneous frequency  $\xi_1$  need to be estimated [6], that is

$$\phi_{CQT} \approx \frac{2\pi R}{\xi_s} \hat{\xi}_1 (2^{r/B} - 1) \quad (17)$$

This approximation introduces slight frequency and amplitude modulations in the output signal. We shall use an 75% overlap-add processing for a frame-synchronous time-domain amplitude modulation instead of 50% overlap-add processing for the sliced-NSGT proposed in [19].

## 6. Implementation

The goal of this paper is to explore the basic tools based Gabor frame theory for a complete scheme for the analysis, transformation, and re-synthesis of a sound [19, 21].

### 6.1. Choice of Sliced Window Length

In this paper, we propose the following structure of the sliCQ transform for pitch shifting. In the analysis, the signal  $f$  is sliced into overlapping slices  $f_m$  of length  $2N$  by multiplication with uniform translates of a slicing window  $q_0$ , centered at 0; then the coefficients  $c_m$  are obtained for each sliced  $f_m$  by applying CQ-NSGT $_{2N}(f, g, a)$  (Eq (4)); The sliced coefficients  $c_m$  are re-arranged into 2-layer array relating two consecutive slices because of the overlap of the slicing window. To exactly mimic time-domain subsampling in the frequency domain, all non-zero spectral components in the range between  $-\xi_s/2$  and  $\xi_s/2$  have to be mapped to the frequency range  $]-\xi_s^k/2, \xi_s^k/2]$  with the mapping function

$$M(\xi, \xi_s^k) = \xi - \lfloor \frac{\xi}{\xi_s^k} \rfloor \xi_s^k \quad (18)$$

where  $\xi$  is the original frequency,  $M(\xi, \xi_s^k)$  is the image frequency after subsampling and  $\lfloor \cdot \rfloor$  denotes rounding towards negative infinity. The mapping function  $M(\xi, \xi_s^k)$  generates a circularly shifted spectrum where the shift is given by  $M(\xi, \xi_s^k)$ . In the synthesis, the coefficients  $c_m$  are retrieved by partitioning; Then compute the dual frame  $G(\bar{g}, a)$  for  $G(g, a)$

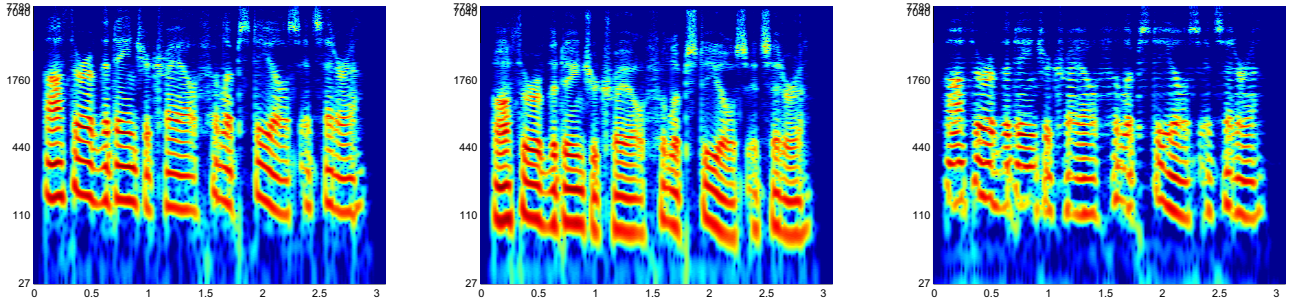


Figure 1: Spectrograms of pitch shifted signal by the proposed method (left), the original signal (middle) and pitch shifted signal by the CQT (right).

for all  $m$ ,  $\tilde{f}_m = \text{iCQ-NSGT}_{2N}(c_m, \tilde{g}, a)$ ; The signal  $f$  is recovered by overlap-add. The Turkey window is chosen as sliced window. The window length and overlap length are shown in Table 1 for a sentence of 2 sec. We select the length of the window for trade off the quality and running time. The length of Turkey window is chosen as 16384 for real-time processing.

Table 1: Relative Error (DB) vs Window & Overlapping Length

SL	2048	4096	4096	16384
Tr	SL/4	SL/4	SL/8	SL/4
Relative Err (DB)	-58.8	-358.4	-358.8	-355.4
Time (s)	14.89	7.61	7.69	2.46

## 6.2. Choice of Window Function

For the CQ-NSGT, it is important to determine which time window function  $g$  to use to calculate the constant Q coefficients. The strategy is to keep the leakage side as small as possible. Figure 2 shows the original time windows and their frequency response. The original window functions used are Hanning, Blackman, Nuttall, and Black Harris windows [22, 23]. We observe that the Nuttall window function and its frequency response can give us a slightly better performance than other windows. It should be note that the pitch shifting of CQT coef-

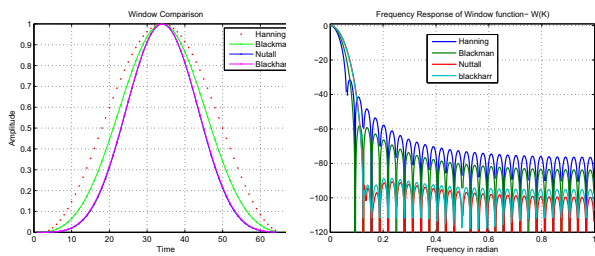


Figure 2: Temporal variation and frequency response for selected window functions: original windows (left) and frequency response of windows (right).

ficients depends not only on the placement of sampling points in frequency domain, but also on their placement in time domain. The minimal redundancy of the CQT representation is still invertible, but they can not used for pitch shifting. We use a rasterized CQT representation where the hop sizes for all center

frequencies are set to the hop size with a reasonable redundancy in the representation [17].

## 6.3. Simulation Results

The CMU ARCTIC corpora are used to evaluate the performance of the pitch shifter. We compare the voice quality of the pitch shifted speech for pitch-shifting in the range of  $\pm 1$  octave for speech. The first 10 sound samples are used from each of the 4 US English (2 female (stl, clb) voices and 2 male (bdl, rms) voices). In the experiment, given isolated sentence generated by CQT and VQT pitch shifters by shifting CQT or VQT bin to 10 or 30, respectively, twelve people including 3 staff and 9 students are asked to label it on a Mean Opinion Score (MOS) scale in terms of naturalness. The average MOSs are shown in Table 2. The results of VQT are much better than those of CQT. From Figure 1, the performance of proposed algorithm shows better than that of CQT-based pitch shifter. In speech,

Table 2: Subjective Listening Test Results

Data Sets	CQT	VQT	CQT	VQT
Shifting Bin	10	10	30	30
Female	3.4	3.8	3.2	3.6
Male	3.3	3.5	3.0	3.3

the aperiodic signal components have very different properties [10, 24]. Results showed that 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.

## 7. Conclusions

A real-time variable-Q non-stationary Gabor transform (VQ-NSGT) system is proposed for speech pitch shifting by simple frequency translation. A hybrid of smoothly varying Q scheme is used to attempt to retain the formant structure of the original signal at both low and high frequencies. Moreover, the preservation of transients of speech are improved due to the high time resolution of VQ-NSGT at high frequencies. A sliced VQ-NSGT is used to retain inter-partial phase coherence by synchronized overlap-add method. The simulation results showed that the proposed approach is suitable for pitch shifting of both speech and music signals. We will develop adequate methods to manage modified analysis coefficients to preserve or even improve the existing speech transformation techniques.

## 8. References

- [1] M. Dong, S. W. Lee, H. Li, P. Chan, X. Peng, J. W. Ehnes, and D.-Y. Huang, "I2r speech2singing perfects everyone's singing," in *Proceedings of (Show and Tell) INTERSPEECH*, Sept 2014.
- [2] D.-Y. Huang, S. Rahardja, and E. Ong, "High level emotional speech morphing using straight," in *Proceedings of 7th ISCA Speech Synthesis Workshop (SSW7)*, 2010, pp. 345 – 350.
- [3] —, "Lombard effect mimicking," in *Proceedings of 7th ISCA Speech Synthesis Workshop (SSW7)*, 2010, p. 258–263.
- [4] D.-Y. Huang, S. Rahardja, E. Ong, M. Dong, and H. Li, "Transformation of vocal characteristics: A review of literature," in *Proceedings of World Academy of Science, Engineering and Technology*, vol. 60, 2009, pp. 340–349.
- [5] E. Coyle, D. Dorran, and R. Lawlor, "A comparison of time-domain time-scale modification algorithms," in *Proceedings of the 120th Convention of the Audio Engineering Society*, convention paper 6674, May 2006.
- [6] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 3, p. 323332, 1999.
- [7] —, "New phase-vocoder techniques are real-time pitch shifting, chorusing, harmonizing, and other exotic audio modifications," *J. Audio Eng. Soc.*, vol. 47, no. 11, pp. 928–936, 1999. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=12086>
- [8] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5, p. 453467, 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)*, 1993, p. 131134.
- [10] A. Röbel, "A shape-invariant phase vocoder for speech transformation," in *Proc. Int. Conf. on Digital Audio Effects*, Sept 2010.
- [11] P. Balazs, M. Dörfler, F. Jaillet, N. Holighaus, and G. A. Velasco, "Theory, implementation and applications of nonstationary gabor frames," *Journal of Computational and Applied Mathematics*, vol. 236, no. 6, pp. 1481 – 1496, 2011.
- [12] G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an invertible constant-q transform with non-stationary gabor frames," in *Proc. Int. Conf. on Digital Audio Effects*, Sept 2011.
- [13] T. Necciari, P. Balazs, N. Holighaus, and P. Sondergaard, "The ERBlet transform: An auditory-based time-frequency representation with perfect reconstruction," in *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*. IEEE, 2013, pp. 498–502.
- [14] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Speech and Audio Processing*, vol. 34, pp. 744–754, 1986.
- [15] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, no. 5, p. 926940, 2011.
- [16] W.-H. Liao, A. Röbel, and A. W. Y. Su, "On stretching gaussian noises with the phase vocoder," in *Proc. of the 15th Int. Conf. on Digital Audio Effects*, Sept 2012.
- [17] C. Schörkhuber, A. Klapuri, and A. Sontacchi, "Audio pitch shifting of signals using the constant-q transform," *J. Audio Eng. Soc.*, vol. 61, no. 7/8, p. 562572, 2013.
- [18] J. Brown, "Calculation of a constant q spectral transform," *Journal of Acous. Soc. Amer.*, vol. 89, no. 1, p. 425434, 1991.
- [19] N. Holighaus, M. Dörfler, G. A. M. Velasco, and G. Thomas, "A framework for invertible, real-time constant-q transforms," *IEEE Trans. Speech and Audio Processing*, vol. 21, no. 4, pp. 775–785, 2013.
- [20] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.
- [21] C. Schörkhuber, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society (53rd Conference on Semantic Audio)*, G. Fazekas, Ed., AES (Vereinigete Staaten (USA)), 1 2014, procedure: peer-reviewed.
- [22] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.
- [23] A. H. Nuttall, "Some windows with very good sidelobe behavior," *Acoustics, Speech and Signal Processing*, vol. 29, no. 1, pp. 84–91, 1981.
- [24] G. Richard and C. d' Alessandro, "Analysis/synthesis and modification of the speech aperiodic component," *Speech Communication*, vol. 19, no. 3, pp. 221–244, 1996.