Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 13º Congresso de Engenharia de Áudio
19ª Convenção Nacional da AES Brasil
25 a 28 de Maio de 2015, São Paulo, SP

Fan Chirp Transform with nonlinear time warping

Isabela F. Apolinário,[1] Luiz W. P. Biscainho,[1] Martín Rocamora,[2] and Pablo Cancela[2]

[1] Federal University of Rio de Janeiro, COPPE, Electrical Engineering Program
Rio de Janeiro, Brazil

[2] Universidad de la República, Instituto de Ingeniería Eléctrica, Montevideo, Uruguay

isabela.apolinario@smt.ufrj.br, wagner@smt.ufrj.br, rocamora@fing.edu.uy, pcancela@fing.edu.uy

ABSTRACT

This paper proposes an extension of a method for time-frequency analysis of nonstationary harmonic signals: the Fan Chirp Transform (FChT). In its original form, the FChT considers that each fundamental frequency (along with the higher harmonics) of the signal may vary linearly in a period of time. This model, however, may be considered poor for some types of signals, especially those whose fundamental frequencies vary rapidly with time. By allowing quadratic frequency variation, this article presents a solution to this problem, which may be considered the next step of the FChT. The proposed technique is assessed in the context of music signals.

0 INTRODUCTION

Analyzing the frequency content of a signal is one of the most essential operations in Signal Processing. To do that, one usually considers the signal under analysis to be time-invariant, and computes its Fourier Transform. However, this is not appropriate when the signal has rapid fluctuations in frequency, such as the ones produced by pitch variations in speech signals.

Many methods have been proposed to analyse those type of signals, such as the Short-Time Fourier Transform (STFT), the Wigner Distribution, the Wavelet Transform, and so on [1], [2], [3]. Among them, the Fan Chirp Transform (FChT) was introduced in [4], originally devised for speech processing. The goal of this transform is to provide a representation as concentrated as possible of the energy of a harmonic linear chirp in the time-frequency plane.

When it comes to music, it makes no sense to analyze the signal as a whole. Its frequency content is changing with each musical note, which in turn depends on how it is played by the musician. It is important to consider such frequency fluctuations in time in order to have a more precise description of the time-frequency content of a music signal. Because of that, music-

oriented time-frequency transformations, such as sinusoidal modeling [5], [6] and the Constant-Q Transform [7] have become powerful tools. In [8], the FChT was applied to the analysis of music signals by means of the Short-Time Fan Chirp Transform (STFChT).

For some applications, there is a certain interest in exploring the sparsity provided by the FChT, especially in higher harmonics. Works as [9] and [10] take this advantage into account when computing the sinusoidal modeling and extracting the main melody of a music signal, respectively.

The present work still focuses on audio analysis, but allows that the fundamental frequency of the signal varies quadratically in time instead of only linearly as in previous works. This new approach is expected to yield better resolution for signals with rapid frequency fluctuations such as vibratos.

The next section details the FChT and discusses its FFT-based implementation. In Section 2, the FChT with nonlinear warping is introduced. Some experimental results are shown in Section 3 while conclusions and future work are presented in Section 4.

1 THE FAN CHIRP TRANSFORM

This section defines the FChT and briefly discusses its implementation and use in real signals.

1.1 Definition

As said before, the FChT provides an acute representation of harmonically related linear chirp signals. It is described in [8] as

$$X(f,\alpha) \triangleq \int_{-\infty}^{\infty} x(t)\phi'_\alpha(t)\mathrm{e}^{-\mathrm{j}2\pi f\phi_\alpha(t)}dt, \qquad (1)$$

where $\phi_\alpha(t)$ is a linear time warping function given by

$$\phi_\alpha(t) = \left(1+\frac{1}{2}\alpha t\right)t. \qquad (2)$$

Applying the variable change $\tau = \phi_\alpha(t)$ to Equation (1), one obtains

$$X(f,\alpha) = \int_{-1/\alpha}^{\infty} x(\phi_\alpha^{-1}(\tau))\mathrm{e}^{-\mathrm{j}2\pi f\tau}d\tau, \qquad (3)$$

where α is the chirp rate parameter, $\phi_\alpha^{-1}(t)$ is given by

$$\phi_\alpha^{-1}(t) = -\frac{1}{\alpha} + \frac{\sqrt{1+2\alpha t}}{\alpha}, \qquad (4)$$

and one assumes that $x(t) = 0$ for $t \leq -1/\alpha$ to avoid aliasing [4].

From Equation (3), it is possible to see that the FChT is the Fourier Transform of a time-warped version of signal $x(t)$. Therefore, the FChT can be calculated by taking advantage of the fast implementation of the Fourier transform, the FFT algorithm [8].

1.2 Implementation

This work considers the analysis of audio signals as the main application. The fan geometry of the FChT seems appropriate to represent these types of signal, as long as they are essentially composed by tones, each of them consisting of a fundamental frequency and higher harmonics. Nevertheless, this fundamental frequency can be well approximated by a linear chirp only for a short period of time, which forces the FChT to be calculated in consecutive short-time signal frames. This is called the STFChT and can be seen as a generalization of the spectrogram [8].

The first step of an implementation of the FChT is the time warping of each frame of the discrete signal $x[n]$. This step is performed via a nonuniform resampling. Since one only has access to its samples at time instants nT_s, where T_s is the sampling period, an interpolation is carried out [8]. Next, the FFT of the time-warped signal is calculated.

The main difficulty here is to find the appropriate value of α. For doing so, an exhaustive search is performed, where all the admissible values of fundamental frequencies f_0 and chirp rates α are considered. An auxiliary function $\rho(f_0,\alpha)$, called salience plane, is created to help with this task. This function consists on an harmonic accumulation performed for every (f_0,α) pair. If f_0 actually represents an existing fundamental, then the energy at its partials is significant and, therefore, a higher value of $\rho(f_0,\alpha)$ is expected. Likewise, if the correct α value is applied, the sparsity of $\rho(f_0,\alpha)$ is maximum, i.e., the peak value obtained is the higher possible one. For any other α value, the energy of the peak (corresponding to the existing fundamental) would be spread among adjacent bins.

The procedure [8] is briefly described below:

- Many instances of the FChT are calculated for different pre-determined values of α.
- A fundamental frequency grid is defined and, for each f_0 and FChT instance, the sum of the harmonics' log-magnitudes is calculated as [11]

$$\rho(f_0,\alpha) = \frac{1}{n_\mathrm{H}}\sum_{i=1}^{n_\mathrm{H}}\log|S(if_0,\alpha)|, \qquad (5)$$

  where $S(f,\alpha)$ is the FChT and n_H is a predetermined number of harmonics.
- Now, one has a dense plane $\rho(f_0,\alpha)$ that concentrates energy in some points (f_0,α), each of which represents an audio source with fundamental frequency f_0 increasing (or decreasing) at the rate of $|\alpha|$ Hz.
- The highest value of the salience plane $\rho(f_0,\alpha)$ for each f_0 is chosen, yielding a salience function $\overline{\rho}(f_0)$. The peaks of $\overline{\rho}(f_0)$ represent, as said before, audio sources; from them, the chirp rate parameter α for each source is obtained.

In practice, there is no such correct α value, since the proposed model is a first order approximation of the fundamental frequency. Additionally, because of complexity purposes, only a finite number of parameter values is tested. These assumptions introduce small errors to the estimation of α.

It is also important to emphasize that it is only possible to choose one α value to compute the FChT for each time frame. This way, when the signal involves various simultaneous sound sources, it can properly represent one of them at a time, while giving a poor representation for the remaining ones [8].

The desired time-frequency signal representation is provided by the STFChT, which is built as the concatenation of all previously frame-wise computed FChTs. By concatenating the salience functions $\overline{\rho}(f_0)$, one generates a "summarizing" time-frequency representation known as F0gram, which shows the temporal evolution of pitch for all harmonic tones in a music signal [8], and can give some insights about the estimated α values, as will be done in the following sections.

2 THE FCHT WITH NONLINEAR TIME WARPING

The fundamental frequency of a music signal can sometimes present rapid fluctuations in a short period of time. In this case, its approximation by a linear function would be poor, whereas choosing a set capable of representing higher variations in frequency could result in a sparser transformation. In this work, a second-order polynomial is chosen to approximate the fundamental.

Figure 1 (up) shows an example of the ground-truth melody (fundamental frequencies) of an opera excerpt along time, where zero values represent note absences in the foreground melody. It can be noticed that the analyzed signal presents rapid fluctuations in frequency and, as mentioned before, the fundamental frequency may not be well approximated by a linear function when considering short periods of time. By adding a third term to Equation (2) one expects to improve the representation of the fundamental frequency. One has

$$\phi_{\alpha,\beta}(t) = \left(1 + \frac{1}{2}\alpha t + \frac{1}{3}\beta t^2\right) t, \qquad (6)$$

where β is called the curvature parameter.

Note that, by this definition, the instantaneous frequency is given by

$$\nu(t) = f\frac{\mathrm{d}}{\mathrm{d}t}\,\phi_{\alpha,\beta}(t) = (1 + \alpha t + \beta t^2)f. \qquad (7)$$

This shows that the fundamental frequency will now be approximated by a second-order polynomial instead of first-order as before.

This section aims at verifying whether this representation is viable and, then, searching for the best choice of values for parameters α and β. The next step is to add the nonlinear warping to the current FChT implementation and see whether or when it brings an improvement to the time-frequency representation.

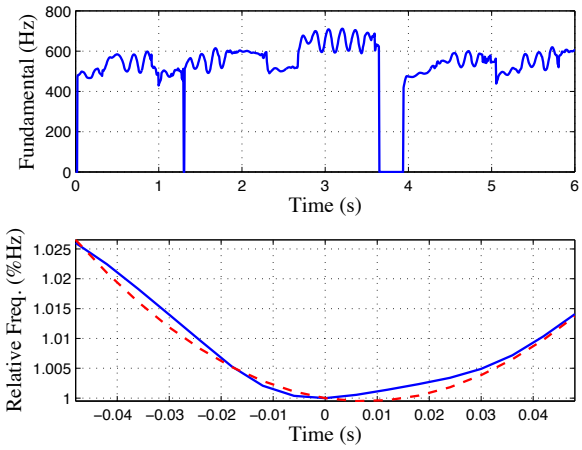

Figure 1: Fundamental frequency of an opera signal (upper graph) and quadratic approximation in dashed red for 100 ms of an audio signal in blue (lower graph).

2.1 Parameter Sampling

A 100-ms frame extracted from the frequency track shown in Figure 1 (up) is zoomed in blue in the bottom plot, and compared with its second-order polynomial approximation drawn in dashed red. Both curves are normalized w.r.t. their frequencies in the middle of the frame ($t = 0$ s).

As mentioned before, for this example a second order approximation seems clearly more suitable. For most signals, though, the fundamental frequency variation may be subtle enough to be approximated by a linear function. The main goal here is, however, to analyze such music signals that in fact exhibit this rapid fluctuations in pitch, like the opera excerpt.

The first step toward the nonlinear warping is to determine which are the possible values for parameters α and β. A database from MIREX [12] containing excerpts of polyphonic audio for which the main melodies' fundamental frequency had been manually labeled (one of which depicted in Figure 1) was employed to aid in this task. Each signal was divided into 100-ms frames[1], and the values of α and β that yielded the best fitting were calculated for each frame. It is important to point out that the fundamental frequency frame should be normalized as shown in Figure 1 (bottom) prior to parameter computation.

The obtained set of pairs (α, β) was then used to construct a histogram. Parameter ranges were set to $[-4, 4]$ for α and $[-50, 50]$ for β, partitioned into 22 bins each. The result can be seen in Figure 2. We see

[1] Since we are interested in following typical pitch variations of audio signals, a time frame way larger than the 20-ms standard was chosen in order to bring forth significant values of β.

that the majority of α and β values concentrate in point (0,0), but there is still a considerable amount of energy around it. From the preferential values for pairs (α, β) depicted in the histogram, different samplings can be proposed, of which three examples are shown in Figure 2. In the first case, the sampling consists of 23 points representing the two main directions of the (α, β) plane: $\alpha = 0$ and $\beta = 0$. In the second case, the sampling consists of the first one added to 12 additional points around the origin $(0, 0)$. In the third case, the sampling consists of the 175-point ellipse around the origin with 90% of the values of α and β. Since an exhaustive search is performed during the computation of the FChT, the number of sampling points is directly related to the computational cost. This fact determined the adoption of restricted ranges for both parameters.

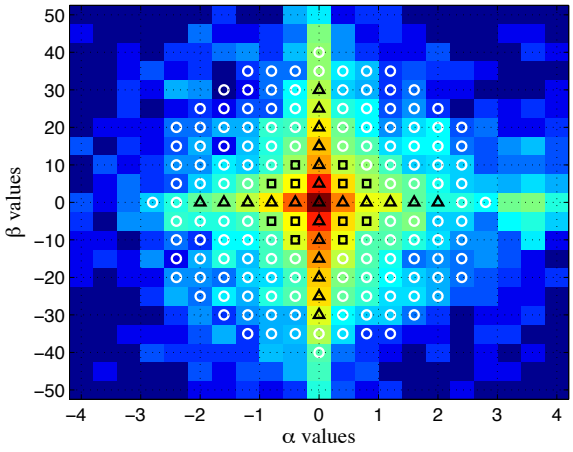

Figure 2: Examples of sampling strategies: 1) the cross denoted by black triangles; 2) the points of the cross plus the ones denoted by black squares; 3) the first two samples plus the points denoted by white circles. The image shows the logarithm of the results.

After choosing an adequate sampling, it is now possible to proceed to the implementation.

2.2 Implementation

The implementation of the FChT with nonlinear time warping also requires an exhaustive search. This time, however, one performs the search over a set of pairs (α, β) instead of just α values. Such pairs can be arranged as an N-element vector, where N is the number of sampling points chosen. If each position γ in this vector represents a pair (α, β), the same method explained in Section 1 can be employed: compute a dense plane $\rho(f_0, \gamma)$ to find the peaks corresponding to audio sources.

3 EXPERIMENTS AND RESULTS

In order to illustrate the effect of the proposed change in the representation, two signals were considered. The first one is a synthetic harmonic signal frequency-modulated by a sinusoid. Its fundamental frequency f_0 is given by the following expression:

$$f_0 = f_1(1 - 2^{1/12}) \sin(2\pi f_2 t) + f_1,$$

where f_1 is the central frequency and f_2, the modulation frequency. The values were chosen to mimic a typical vibrato, as found in singing voice performances, namely, 500 Hz and 6 Hz, respectively. The second signal is the opera excerpt introduced in Section 2, which presents rapid pitch fluctuations with time. For both signals, the sampling frequency is 44100 Hz.

Figure 3 shows the magnitude of the STFChT of the synthetic signal for both linear (second column) and nonlinear (third column) warpings. The STFT for the same signal is also shown (first column). Three different window sizes were chosen for comparison: 1024 samples (first row), 2048 samples (second row), and 4096 samples (third row). This is done in order to enable a fair comparison of the methods.

It is possible to notice the effect of the chosen window size in the resolution of high and low frequencies in the STFT. Increasing the window length yields higher frequency resolution, which provides a better representation for slow varying harmonics, as found in low frequencies. On the other hand, a shorter window is needed to improve the temporal resolution of rapid varying harmonics, which can be seen in the high frequency range. This is a classical issue when dealing with the STFT whose effects are mitigated, as shown, by the STFChT. As a matter of fact, it is desirable to have the largest possible window when using the STFChT, since this means a smaller spread in frequency due to the windowing process. This upper bound is restricted by the chosen warping type. The window length can be increased as long as the frequency variations within the frame can be properly approximated by the warping. For instance, when the window length is increased to 4096 samples, the linear warping is no longer capable of modeling the evolution of the harmonics, and a resolution decrease arises especially in the regions with high curvature. When using the nonlinear warping, one has one more degree of freedom to model the fundamental frequency variations and can, therefore, allow the analysis window to have a larger number of samples. It is important to notice, however, that the computation of the STFChT is dependent on the parameter α (and β, for the nonlinear warping case). They have to be correctly estimated in order to allow a sparse representation.

The analysis obtained with the nonlinear warping, shown in Figure 3, exhibits a sparser representation compared to the linear warping, especially around peaks and valleys of the partial contours. Not surprisingly, the higher order model can approximate with more detail the actual frequency variations. Some artifacts can be observed in this representation due to the estimation of α and β values from a discrete set.

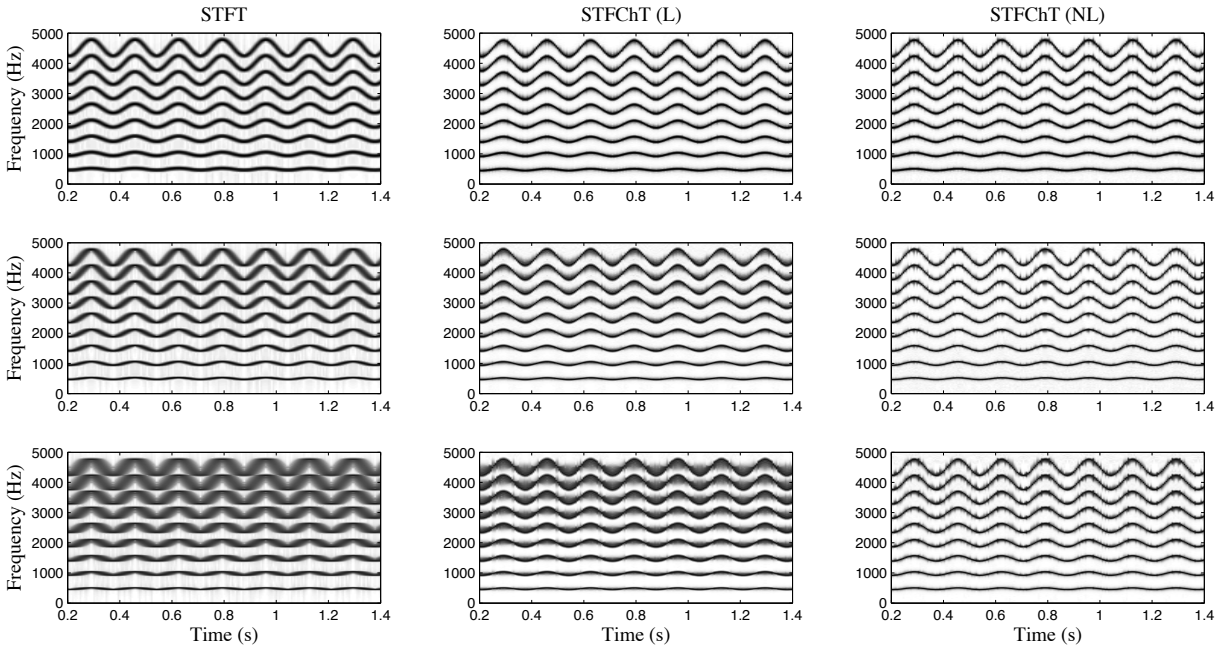

Figure 3: STFT (first column), STFChTs with linear (second column, L) and nonlinear (third column, NL) warpings of a synthetic signal. The following window sizes were used: 1024 (first row), 2048 (second row), and 4096 (third row) samples.

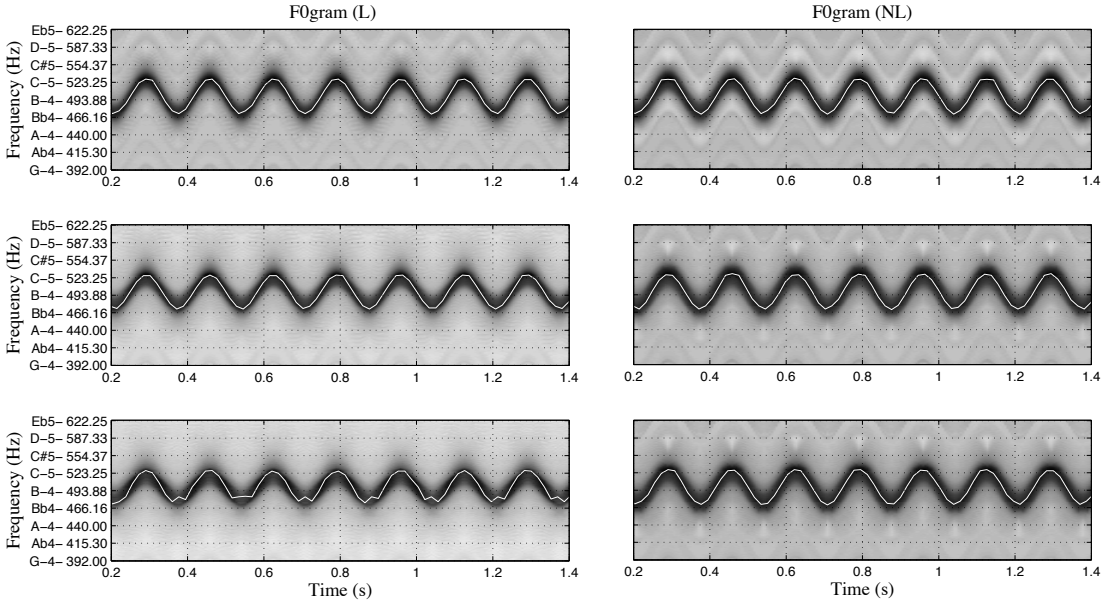

Figure 4: F0gram graphics for the STFChTs with linear (first column, L) and nonlinear (second column, NL) warpings of a synthetic signal. The following window sizes were used: 1024 (first row), 2048 (second row), and 4096 (third row) samples.

Such behavior can also be outlined through the corresponding F0gram graphics, shown in Figure 4. As before, the linear (left column) and nonlinear (right column) warping cases are presented with window sizes of 1024 samples (first row), 2048 samples (second row), and 4096 samples (third row). The fundamental frequency, estimated as the maximum value for each time instant, is shown in white on top of the F0gram. The time-frequency resolution of the representation has a noticeable impact on the accuracy of the fundamental frequency estimation. In particular, the linear warping for a 4096 samples window fails to properly follow the actual fundamental frequency contour.

Figure 5 shows the magnitude of the STFChT of the chosen real music signal for both linear (middle) and nonlinear warpings (bottom). Again, the magnitude

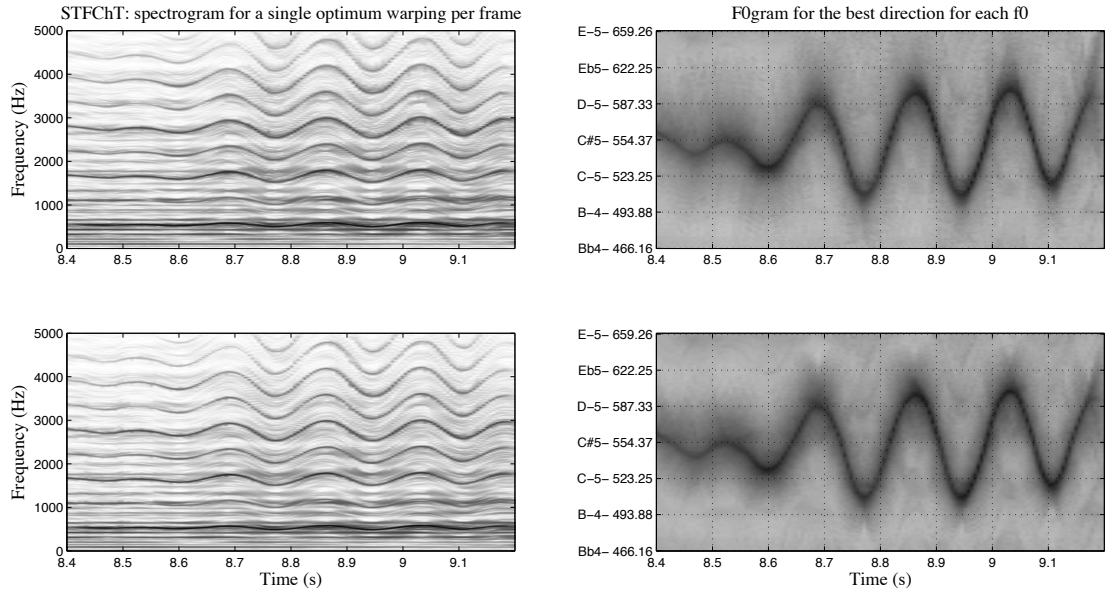

Figure 6: STFChT with linear (higher row) and nonlinear (lower row) warpings of an opera signal.

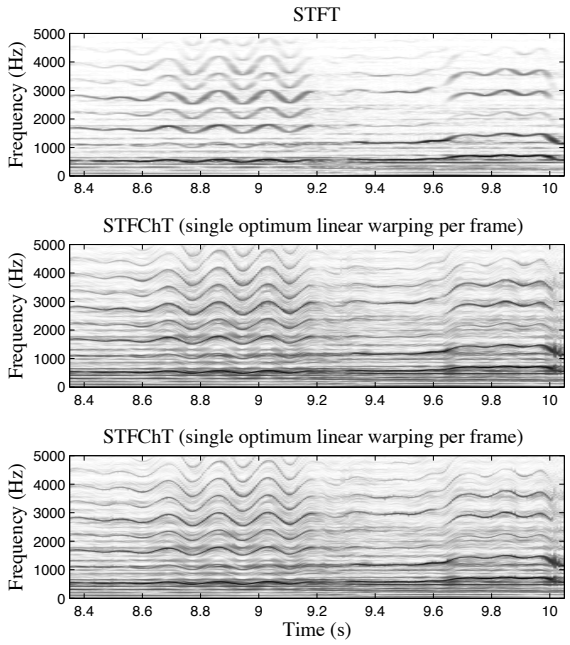

Figure 5: STFT (upper graph) and STFChT with linear (middle graph) and nonlinear (lower graph) warpings of an opera signal.

of the STFT (up) is shown for comparison. The window size used for the these images was 2048 samples, which, considering a sampling frequency of 44100 Hz, correspond to approximately 46.4 ms. Since this signal presents rapid enough pitch fluctuations, a time frame of around half the duration proposed in Section 2.1 was applicable to highlight the main differences between the three representations. As previously stated, it is possible to see that the representation shown in the lower image (STFChT with nonlinear warping) exhibits higher resolution.

Figure 6 shows the F0grams for both the STFChT with linear (up) and nonlinear warpings (bottom). Both methods seem to correctly estimate the fundamental frequency of the signal (not represented here), but it is possible to notice that, for the nonlinear case, the F0gram graphic is slightly sparser.

Another important aspect to consider is the processing time of the nonlinear method. For the nonuniform case and the third type of sampling, for example, since the search is done over a larger amount of samples, the time needed to calculate the FChT is approximately 5 times higher. A further analysis of the complexity is still in progress. This time difference should be taken into account in order to balance this issue and the benefits the nonlinear warping can bring to the representations of signals with fast pitch variations.

4 CONCLUSION

The presented work is an expansion of the formerly proposed FChT that improves the time-frequency resolution of signals that present rapid pitch fluctuations with time.

Among other applications, the FChT can be used in systems for melody detection, denoising, and sound source separation [8]. The proposed modification can improve the performance of the FChT as a result of a finer time-frequency resolution.

It is important to point out that, for complexity reasons, a relatively sparse grid should be employed with the modified STFChT, which can eventually lead to slightly worse resolution in linear segments than the conventional transform with finer α grid would yield. Moreover, the parameter ranges must be blindly chosen, when there is no *a priori* information on the signals. A possible solution to this issue could be guiding

the obtained parameters in an adaptive way, in order to take into consideration only currently plausible values of α and β in the exhaustive search.

It should also be mentioned that the required processing time is higher for the nonlinear warping and thus its application should be restricted to those situations in which fast frequency variations call for better tracking. Future work must include a strong effort to alleviate the computational requirements inherent to the proposed method.

REFERENCES

[1] L. Cohen, *Time-frequency Analysis*, Prentice Hall, Englewood Cliffs, USA, 1995.

[2] P. Flandrin, *Time-frequency/Time-Scale Analysis*, Academic Press, France, 1999.

[3] C. S. Burrus, R. A. Gopinath, , and H. Guo, *Introduction to Wavelets and Wavelet Transforms*, Prentice Hall, Upper Saddle River, USA, 1998.

[4] L. Weruaga and M. Képesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, vol. 87, no. 6, pp. 1504–1522, June 2007.

[5] P. R. Cook, *Real Sound Synthesis for Interactive Applications*, A K Peters, Natick, USA, 2002.

[6] P. A. A. Esquef and L. W. P. Biscainho, "Spectral-based analysis and synthesis of audio signals," in *Advances in Audio and Speech Signal Processing: Technologies and Applications*, Hector Perez-Meana, Ed. February 2007, pp. 56–92, Hershey: IGI Global.

[7] J. C. Brown, "Calculation of a constant q spectral transform," *Journal of the Acoustical Society of America*, vol. 80, no. 1, pp. 425–434, January 1991.

[8] P. Cancela, E. López, and M. Rocamora, "Fan-chirp transform for music representation," in *11th Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010.

[9] M. Bartkowiak, "Application of the fan-chirp transform to hybrid sinusoidal+noise modeling of polyphonic audio," in *16th European Signal Processing Conference (EUSIPCO 2008)*, Lausanne, Switzerland, August 2008, pp. 161–166.

[10] Z. Tang and D. Black, "Melody extraction from polyphonic audio of western opera: A method based on detection of the singer's formant," in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, October 2014.

[11] M. Képesi and L. Weruaga, "Adaptive chirp-based time-frequency analysis of speech signals," *Speech Communication*, vol. 48, no. 5, pp. 474–492, May 2006.

[12] J. Downie, "The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 28, no. 4, pp. 247–255, September 2008.