A Study on Different Strategies for Chirp Rate Sampling in the Fan Chirp Transform

Isabela F. Apolinário and Luiz W. P. Biscainho

***Abstract*— This article is a study on different strategies for sampling the chirp rate parameter used in the Fan Chirp Transform. In particular, sampling based on the estimation of the parameter's PDF is proposed; some experiments are performed in order to get some insight on its potentialities, and the corresponding analysis is also presented. The main target is to reduce computational complexity while maintaining performance.**

***Keywords*— Fan Chirp Transform, Time-Frequency Domain, PDF, Kernel Estimation**

I. Introduction

Time-Frequency analysis is a much useful tool in many audio applications, such as automatic music transcription [1] and sound source separation [2]. It can be also an important preprocessing stage for other tools as sinusoidal analysis [3], which usually require good time-frequency resolution to provide a useful model.

The Fan Chirp Transform (FChT), for instance, enables one to represent an existing fundamental frequency and corresponding harmonics as a set of harmonically-related linear chirps within a time frame [4]. As a consequence, better time-frequency resolution is achieved, since the stationary assumption is dropped. This change may ease the detection of existing higher partials in a peak detection step, therefore improving analysis.

An important parameter for the FChT computation is the chirp rate α, which dictates the velocity of frequency variation within a frame. This parameter is, however, not known a priori and must be estimated, as is done in [5], by an exhaustive search. It is, then, of great importance to minimize the amount of α candidates, N_α, since the computational complexity of the FChT depends directly on this step. In the original implementation, a uniformly-spaced grid within a predetermined symmetrical range with a predetermined number of α values is defined and used as the set of viable candidates. This strategy, however, does not take into consideration which are the most common ranges of frequency variations in audio signals.

This article presents a statistical approach to determine a new sampling method based on the most frequent values of α found in audio signals. A probability density function (PDF) is estimated from a fundamental-frequency annotated database, and used as a starting point to determine new α candidates. Experiments using synthetic and real signals were carried out in order to evaluate the performance of the proposed sampling.

Isabela F. Apolinário and Luiz W. P. Biscainho are with the Program of Electrical Engineering, COPPE/Federal University of Rio de Janeiro, Rio de Janeiro-RJ, Brazil. E-mails: isabela.apolinario@smt.ufrj.br and wagner@smt.ufrj.br. This work was partially supported by CNPq and CAPES.

This paper is organized as follows. Section II briefly explains the target time-frequency representation, the FChT, along with some details on its implementation as defined in [5]. Section III explains the proposed PDF-based sampling. Section IV presents the performed experiments and respective results. Lastly, Section V draws some conclusions and future work.

II. The Fan Chirp Transform

This section briefly exposes the concepts of the FChT and clarifies each step in its computation.

A. Definition

The FChT is defined in [5] as

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

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

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

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

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

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

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

it is assumed that $x(t) = 0$ for $t \leq -1/\alpha$ to avoid aliasing [4].

From Equation (3), it is possible to notice that the FChT is actually the Fourier Transform of a time-warped version of the sinal $x(t)$, $x(\phi_\alpha^{-1}(t))$. Therefore, the FChT can profit from the fast implementation of the Discrete Fourier Transform, the FFT algorithm [5]. This means that the FChT of a given signal will take basically N_α times more to compute than its DFT.

B. Implementation

The FChT implementation is done in short consecutive time-frames, usually ranging from 20 ms to 100 ms, resulting in the so called Short-Time Fan Chirp Transform (STFChT). In each frame, the predominant fundamental is modeled as a linear chirp and, for this, a parameter α is estimated. As previously mentioned, this step is performed via an exhaustive search, in

which a predetermined set of values of α is tested for optimal sparsity.

From now on, each time-frame will be treated individually, since the same procedure to compute the corresponding FChT is applied to all of them. The first step in the computation of FChT is the time-warping forced by function $\phi_\alpha(t)$, as seen in Equation 3. Since the analysed signal is discrete in time, this step is performed by means of a nonlinear sampling. Since the only available samples are those acquired at time instances nT_s, where T_s is the sampling period, an interpolation must be carried out [5]. It is important to emphasize that, in order to perform this time warping operation, parameter α must have been previously chosen.

The estimation of the best chirp rate α should follow a criterion of maximum sparsity. For this purpose, a salience function based on harmonic-accumulation was adopted in this work, as in [5]. The procedure is as follows. Firstly many instances of the FChT for a set of A values of α are calculated, and then the salience function $\rho(f_0, \alpha_a)$ for each α_a, with $a = 1, 2, \ldots, A$, and a grid of frequency values f_0[1], are computed. As a result, a salience plane $\rho(f_0, \alpha)$ is obtained. The point (f_0^*, α^*) corresponding to the maximum value of $\rho(f_0, \alpha)$ is then chosen as the pair estimated fundamental frequency f_0^* / chirp rate α^*. Further details can be found in [5].

III. Estimation of Chirp Rate Parameter

As mentioned, a traditional way to choose the set of parameters α that will be input to the exhaustive search is to perform a uniform sampling within an empirical interval $[-\alpha_{\max}, \alpha_{\max}]$, in which $\alpha_{\max}$ is chosen accordingly to (known) a priori information on possibly found frequency variations. This analysis, however, ignores any previous knowledge on more or less probable ranges that parameter α can visit in music signals, and thus considers high values of α as so probable as values close to zero.

At a first sight, a more appropriate analysis would be to estimate more dense regions of α values, and assign a higher number of sampling points to such neighborhoods. A database from MIREX [6] containing excerpts of polyphonic audio for which the main melody's fundamental frequencies had been manually labeled was employed for this task. Each signal was divided into 50-ms frames, and the corresponding values of α were calculated by a polynomial fitting of the normalized fundamental frequencies for each frame. More details can be found in [7].

From the obtained values, a nonparametric approach to model the distribution of α is employed, namely a kernel density estimator (KDE) with Gaussian kernel function. This method was preferred over a parametric approach for its capability of generalization, and thus not being limited to a possibly poor choice of a density model [8]. Initially, samples of α higher than an empirical value of 20, based on the maximum obtained α of 21.45, were considered outliers and therefore discarded. Then, a Gaussian curve was placed over each resulting data point, and the contributions over the whole set were summed and averaged by the amount of points n in order to preserve normalization, resulting in a density model [8]. The standard deviation h_opt was chosen so as to optimize univariate-data Gaussian distributions for best smoothing [9], i.e.

$$h_\mathrm{opt} \approx 1.06 \hat{\sigma} n^{-1/5}, \tag{5}$$

where $\hat{\sigma}$ is the standard deviation of the samples[2].

From the estimated PDF, the estimated cumulative density function (CDF) is calculated by numerical integration, and therefrom the proposed sampling is performed as a projection in the $x-$axis (α values) of a linear sampling in the $y-$axis of the CDF. This way, a higher amount of sampling points will be located around higher slopes in the CDF, which represent more probable regions, thus justifying the procedure. This process is illustrated in Figure 1 (left column) for a number of α values $N_\alpha = 11$ and maximum α $\alpha_{\max} = 4$. The CDF is shown in blue and the sampling points, in red. The estimated PDF is shown (right column, upper line) in blue, along with sampling points in red. A comparison between this PDF-based sampling method and a standard uniform sampling is also shown (right column, lower line) by the points in red (PDF-based) and black (uniform).

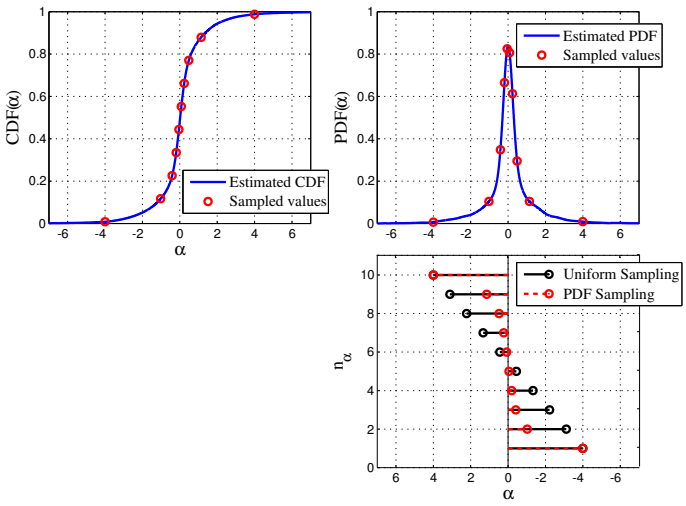

Fig. 1: Left column shows estimated CDF in blue and sampled points in red. Right column shows estimated PDF in blue along with sampled points in red (upper line) and a comparison between points obtained via a PDF-based sampling (red) and uniform-PDF-based sampling (black).

IV. Experiments and Results

In this section, experiments with the two sampling strategies are conducted and their results are analysed. For this purpose, four different audio signals, each with different characteristics, are chosen. Signal I is a synthetic signal that mimics both a glissando and a vibrato; Signal II is an excerpt of the song Michelle, sung by the Beatles, presenting a considerable

[1]The chosen frequency values are fundamental frequency candidates. Here, a grid of 192 geometrically-spaced values per octave was adopted.

[2]Here, $\hat{\sigma} \approx 1.35$ and $h_\mathrm{opt} \approx 0.21$.

amount of polyphony; Signal III is an opera excerpt, representing large frequency variations in time; and Signal IV is a violin solo with some subtle vibratos.

The experiments were conducted using the following input parameters.

- Uniform and PDF-based sampling strategies.
- Numbers of α values, N_α, equal to 5, 7, 9, and 11 for all signals. For the synthetic signal, further tests using 21, 31, and 51 values were conducted.
- Empirical values of $\alpha_{\max}$ equal to 1, 5, and 10 for the synthetic signal; equal to 5, 6, and 7 for signal II; equal to 1, 5, 8, and 10 for signal III; and equal to 1, 2, 5, and 10 for signal IV.
- Number of samples of the analysis window equal to 2048 and 4096, corresponding to approximately 46 and 93 ms, respectively.

For the synthetic signal, a more detailed analysis is performed in order to fully understand the implications of exchanging the original uniform sampling method for the proposed one. The signal consists of a fundamental frequency that increases according to a third order polynomial from 50 Hz to 1 kHz for 6 seconds with an intermediate 2-s vibrato of 5-Hz modulation frequency, plus 20 harmonic partials whose amplitudes decay quadratically with their indexes. The considered sampling frequency is $f_s = 44.1$ kHz. The evolution of the fundamental frequency with time along with α values estimated with an analysis window of $N = 4096$ samples are shown in Figure 2. It is important to point out that α is a relative measure of the derivative of the fundamental frequency, which explains the higher values at low frequencies.

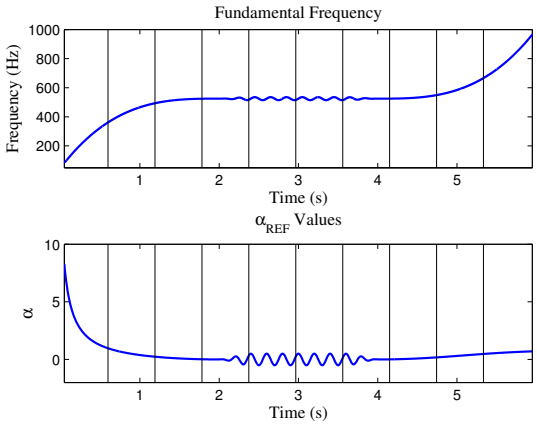

Fig. 2: Signal I: Upper figure shows the evolution of the fundamental frequency in time. Lower figure shows the reference values for α obtained from the former. Horizontal black lines divide the signal in 10 different regions.

The signal was divided in ten equal-length regions for the sake of a clearer analysis, marked in Figure 2 by vertical black lines. Region I presents the most prominent relative frequency variation; regions 5 and 6 are the vibrato region; regions 4 and 7 make the transitions from/to vibrato to/from a more stationary region; and regions 2, 3, 8, 9, and 10 present subtle relative frequency variations.

For each proposed experiment, estimated values of α were measured and then compared to their reference values. The mean squared error (MSE) for each case was then computed, and the best results for both sampling strategies, obtained with $N = 4096$, are shown in Figure 3 in logarithmic scale. MSE values are presented in red for uniform sampling and in blue for PDF-based sampling. The adopted $\alpha_{\max}$ was 10. As expected, one can see a significant improvement in the vibrato regions, 5 and 6, specially for lower values of N_α, when using a PDF-based sampling, since for this method the values of α around zero are more finely modeled. For region I, however, the uniform sampling method has a better performance, since more data points are available at higher values. As N_α is increased, the only persistent meaningful difference between both methods remains in region I.

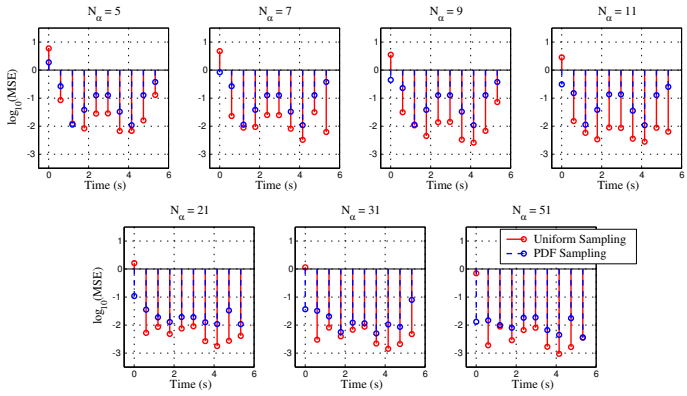

Fig. 3: MSE per region for values of N_α equal to 5, 7, 9, 11, 21, 31, and 51. Here, $N = 4096$ and $\alpha_{\max} = 10$.

Now, in order to evaluate the impact of different sampling strategies in the time-frequency representation, results for $N_\alpha = 11$ are presented in Figures 4 and 5. The Short-Time Fourier Transform (STFT) can be seen in the left column, the uniformly sampled STFChT in the middle column, and the PDF-sampled STFChT in the right column. Regions I, IV, and V were highlighted here. Figure 4 shows the glissando part (composed mainly by region I) of the constructed synthetic signal. It is possible to discern between more concentrated and blurry regions, specially for the $N = 4096$-case. For this type of high-frequency variation, the uniform sampling provides a better representation, since it contains more available points at higher values of α. Figure 5 shows the vibrato part (composed mainly by regions IV and V), which, on the other hand, require more available points at lower values of α. In this case, the PDF-Sampled STFChT provides a better resolution. Besides, when comparing the representations obtained by the different values of N, one can notice that, for high frequency variations, as in the glissando part of the signal, the results were better for $N = 4096$. For the vibrato part, the results were better for $N = 2048$, since the original assumption of linear fundamental frequency variation within a time-frame does not hold any longer.

When comparing both STFChTs with the STFT in Figure 4, one can notice a huge improvement from the STFT to either one of the STFChTs, justifying the use of a chirp model

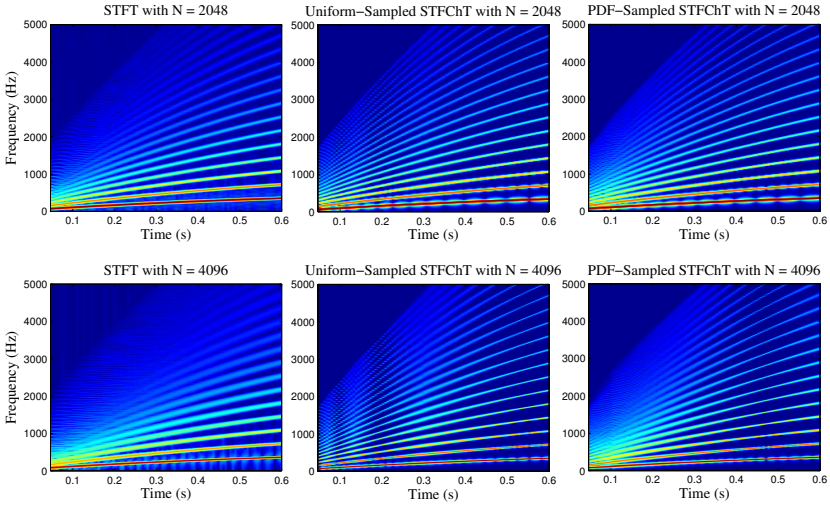

Fig. 4: Time-frequency representations: STFT (left column), STFChT with uniform sampling (middle column), and STFChT with PDf sampling (right column) using $N_\alpha = 11$ and $\alpha_{\max} = 10$ for region I of the synthetic signal. Values of $N = 2048$ (upper line) and $N = 4096$ (lower line) were used.

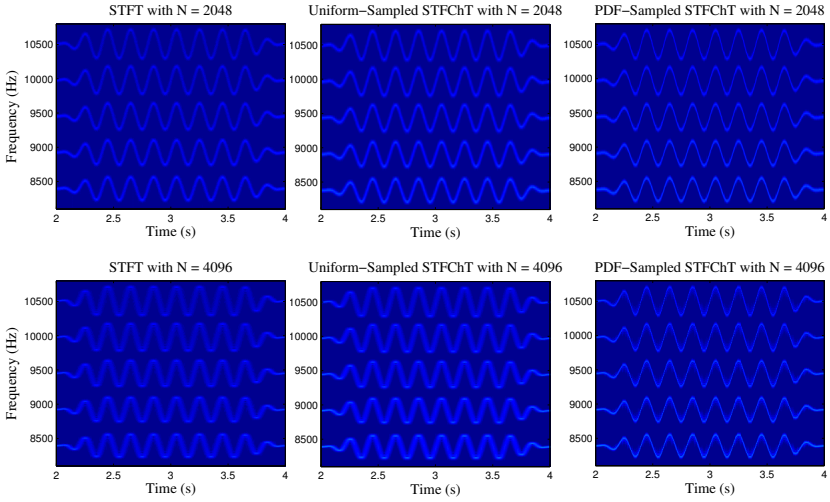

Fig. 5: Time-frequency representations: STFT (left column), STFChT with uniform sampling (middle column), and STFChT with PDf sampling (right column) using $N_\alpha = 11$ and $\alpha_{\max} = 10$ for regions IV and V of the synthetic signal. Values of $N = 2048$ (upper line) and $N = 4096$ (lower line) were used.

for this case. From Figure 5, however, one can see that the difference between the STFT and the uniform-sampled STFChT is not meaningful at all. The improvement would only be perceived at higher values of N_α, such as from 21 on, as can be deducted from Figure 3. With the increase of N_α, the visual difference between both representations becomes gradually less evident, since the chosen α values in both samplings become more similar. The obtained representation, at this point, is no longer sensible to small changes in sampling. These considerations indicate that for slowly varying signals lower values of N_α are required to achieve good time-frequency resolution when using the PDF-sampled STFChT if compared to the required N_α to achieve an equivalent time-frequency resolution when using the uniformly sampled STFChT. Therefore, the proposed strategy requires a lower computational complexity to achieve results comparable to those obtained with the original uniform sampling strategy. Since, from Section III, low frequency variations occur more often than high frequency variations, the results so far confirm the initial expectation.

It is also important to point out that visual changes in the time-frequency transform are more noticeable at higher frequency partials, where frequency resolution tends to be more degraded. This is why the presented results are usually shown within this range.

Other important remark regards the maximum value of α, $\alpha_{\max}$. When decreasing $\alpha_{\max}$, the representation at the vibrato regions improves significantly for the uniform-sampled STFChT, since more values of α are available to describe the small frequency variations. Results for different values of N_α and $\alpha_{\max}$ can be found at `www.smt.ufrj.br/~isabela.apolinario`.

From the simulated experiments with real signals, the Beatles' song was chosen as a significant example for its

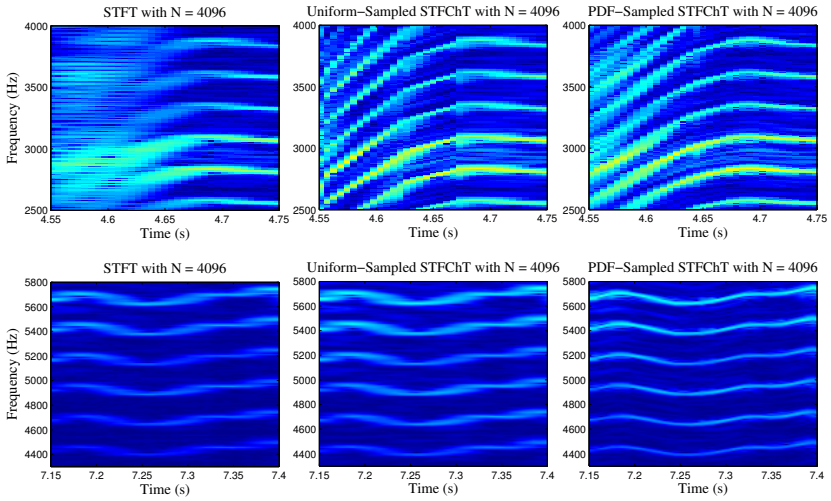

Fig. 6: Time-frequency representations: STFT (left column), STFChT with uniform sampling (middle column), and STFChT with PDF sampling (right column) using $N_\alpha = 11$ for different regions of the pop signal. Here, $N = 4096$ and $\alpha_{\max} = 7$.

content of both high and subtle frequency variations. This way, both aspects can be investigated in a more practical case. The time-frequency transforms for the chosen signal using $N = 4096$, $N_\alpha = 11$ and $\alpha_{\max} = 7$ can be seen in Figure 6. The Short-Time Fourier Transform (STFT) can be seen in the left column, the uniformly sampled STFChT in the middle column, and the PDF-sampled STFChT in the right column. Two different signal passages are shown: one representing a fast increase in frequency (upper line) and another representing small frequency variations around a somewhat fixed frequency (lower line), both performed by the main singer. An analysis similar to that presented for the synthetic signal applies here. For the high frequency variation, a better resolution was achieved by the uniformly sampled STFChT, while for subtle variations, the PDF-based sampling presented a more concentrated time-frequency transform.

A relevant remark is that the changes in the sparsity of the obtained representations were visually perceived only for an analysis window of $N = 4096$ samples, due to the higher frequency resolution obtained over the case of $N = 2048$ samples. A decrease in time resolution is, however, also noticed. In this case, important information about onsets and fast vibratos, such as the one presented in Figure 5 or operatic excerpts, for instance, could be lost. A better compromise could be reached by adapting the window length according to a priori information about the frequency variations found in the signal in order to optimize the results for each case.

Once more, the improvements from the STFT to both obtained STFChTs were expressive for fast varying frequencies, but the superiority of the PDF-sampled STFChT over its uniform counterpart is significant only for slower variations. For higher values of N_α, both samplings presented equivalent results in regions of slight frequency variations, but the uniform sampling continued to produce sparser representations whenever a higher fluctuation occurred. This way, an effort should be made in order to combine the sensitivity of the time-frequency transform to α parameters with the already considered probability of occurrence.

Results for all proposed real signals can be found at `www.smt.ufrj.br/~isabela.apolinario`.

V. Conclusions

This work presented some investigation on the exchange of a uniform for a PDF-based sampling of the FChT parameter α. The results show a promising improvement in regions of small frequency variations, specially when one employs a reduced number of α values, at the cost of some decrease in resolution for high-frequency variations. A more detailed study on the sensitivity of the time-frequency transform to this parameter should be performed as future work. Furthermore, experiments with different signal-to-noise ratios should be carried out in order to properly evaluate the effect of additive noise over the proposed sampling.

References

[1] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, New York, USA, 2006.

[2] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, v. 5, no. 4, pp. 1475–1487, May 2007.

[3] 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.

[4] L. Weruaga and M. Képesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, v. 87, n. 6, pp. 1504–1522, June 2007.

[5] P. Cancela, E. López, and M. Rocamora, "Fan-chirp transform for music representation," in *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010, pp. 1–8.

[6] J. Downie, "The Music Information Retrieval Evaluation Exchange (2005-2007): A Window Into Music Information Retrieval Research," *Acoustical Science and Technology*, v. 28, n. 4, pp. 247–255, September 2008.

[7] I. F. Apolinário, L. W. P. Biscainho, M. Rocamora, and P. Cancela, "Fan Chirp Transform with nonlinear time warping," *Anais do 13o Congresso Nacional da AES Brasil*, São Paulo, Brazil, May 2015.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, USA, 2006.

[9] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, England, 1986.