FAN CHIRP TRANSFORM FOR MUSIC REPRESENTATION

Pablo Cancela *Ernesto López* *Martín Rocamora*

Instituto de Ingeniería Eléctrica,
Universidad de la República,
Montevideo, Uruguay
{pcancela,elopez,rocamora}@fing.edu.uy

ABSTRACT

In this work the Fan Chirp Transform (FChT), which provides an acute representation of harmonically related linear chirp signals, is applied to the analysis of pitch content in polyphonic music. The implementation introduced was devised to be computationally manageable and enables the generalization of the FChT for the analysis of non-linear chirps. The combination with the Constant Q Transform is explored to build a multi-resolution FChT. An existing method to compute pitch salience from the FChT is improved and adapted to handle polyphonic music. In this way a useful melodic content visualization tool is obtained. The results of a frame based melody detection evaluation indicate that the introduced technique is very promising as a front-end for music analysis.

1. INTRODUCTION

Most real signals (for instance, music signals) are non-stationary by nature. Moreover, usually an important part of the information of interest has to do with the non stationarity (beginning and end of events, modulations, drifts, etc). For this reason, the development of time-frequency representations for the analysis of signals whose spectral content varies in time is an active field of research in signal processing [1]. The representation is commonly adapted to the signal in order to enhance significant events so as to facilitate the detection, estimation or classification. An alternative goal is to obtain a sparse representation for compression or denoising. In some cases the elements of the sparse representation become associated with salient features of the signal thus also providing feature extraction [2].

The Short Time Fourier Transform (STFT) [3] is the standard method for time-frequency analysis. This representation is appropriate under the assumption that the signal is stationary within the analysis frame. In addition, time-frequency resolution is constant in the STFT. However, for the analysis of music signals a non uniform tiling of the time-frequency plane is highly desired. Higher frequency resolution is needed in the low and mid frequencies where there is a higher density of harmonics. On the contrary, frequency modulation (typical of the singing voice rapid pitch fluctuations) calls for improved time resolution in higher frequencies. Different multi-resolution time-frequency alternatives to the STFT have been proposed such as the Constant-Q Transform (CQT) [4].

Precisely representing frequency modulated signals, like singing voice, is a challenging problem in signal processing. Many time-frequency transforms can be applied for this purpose. The most popular quadratic time-frequency representation is the Wigner-Ville Distribution (WVD), which offers good time-frequency localization but suffers from interfering cross-terms. Several alternatives were proposed to attenuate the interferences such as the Smoothed Pseudo WVD and other Cohen class distributions [3], but with the side effect of resolution loss due to the smoothing. A different approach to perform the analysis is considering the projection over frequency modulated sinusoids (chirps), in order to obtain a non-Cartesian tiling of the time-frequency plane that closely matches the pitch change rate. Among the chirp-based transforms, the Chirplet Transform [5] and the Fractional Fourier Transform [6] involve the scalar product between the signal and linear chirps (linear FM), and can reach optimal resolution for a single component linear chirp. However, many sounds present in music (e.g. voice) have an harmonic structure, and these transforms are not able to offer optimal resolution simultaneously for all the partials of a harmonic chirp (harmonically related chirps). In the case of harmonic signals, the Fan Chirp Transform (FChT) [7] is better suited as it provides optimal time-frequency localization in a “fan” geometry. The FChT can be considered as a time warping followed by a Fourier Transform, which enables an efficient implementation using the FFT. Although many of these techniques were applied to speech [8], the use of time-frequency representations other than the STFT for music analysis remains rather scarce [2, 9] and in particular the FChT to the best of our knowledge has almost not been explored for this purpose, except a very few exceptions [10, 11]. See figures 1 and 2 for a comparison of different time-frequency representations applied to a music audio excerpt.

In this work the FChT is applied to the analysis of pitch content in polyphonic music. Besides, it is combined with the CQT to provide time-frequency multi-resolution in the fan geometry. The formulation and implementation of the FChT differ from the proposed in [7]. The goal of the formulation is to obtain a more acute representation of linear chirps. A positive byproduct is that it also enables the application of arbitrary warpings in order to analyze non-linear chirps straightforwardly. In addition, the implementation is designed with an emphasis on computational cost. Among the various existing approaches for pitch salience computation, the technique adopted in this work is based on gathering harmonically related peaks of the FChT as proposed in [8], which is improved and adapted to deal with polyphonic music. Finally the application of the FChT to melodic content visualization and automatic melody detection is illustrated.

2. FAN CHIRP TRANSFORM

2.1. Formulation

In this work, the definition of the FChT adopted is,

$$X(f,\alpha) \triangleq \int_{-\infty}^{\infty} x(t)\ \phi'_\alpha(t)\ e^{-j2\pi f \phi_\alpha(t)} dt, \qquad (1)$$

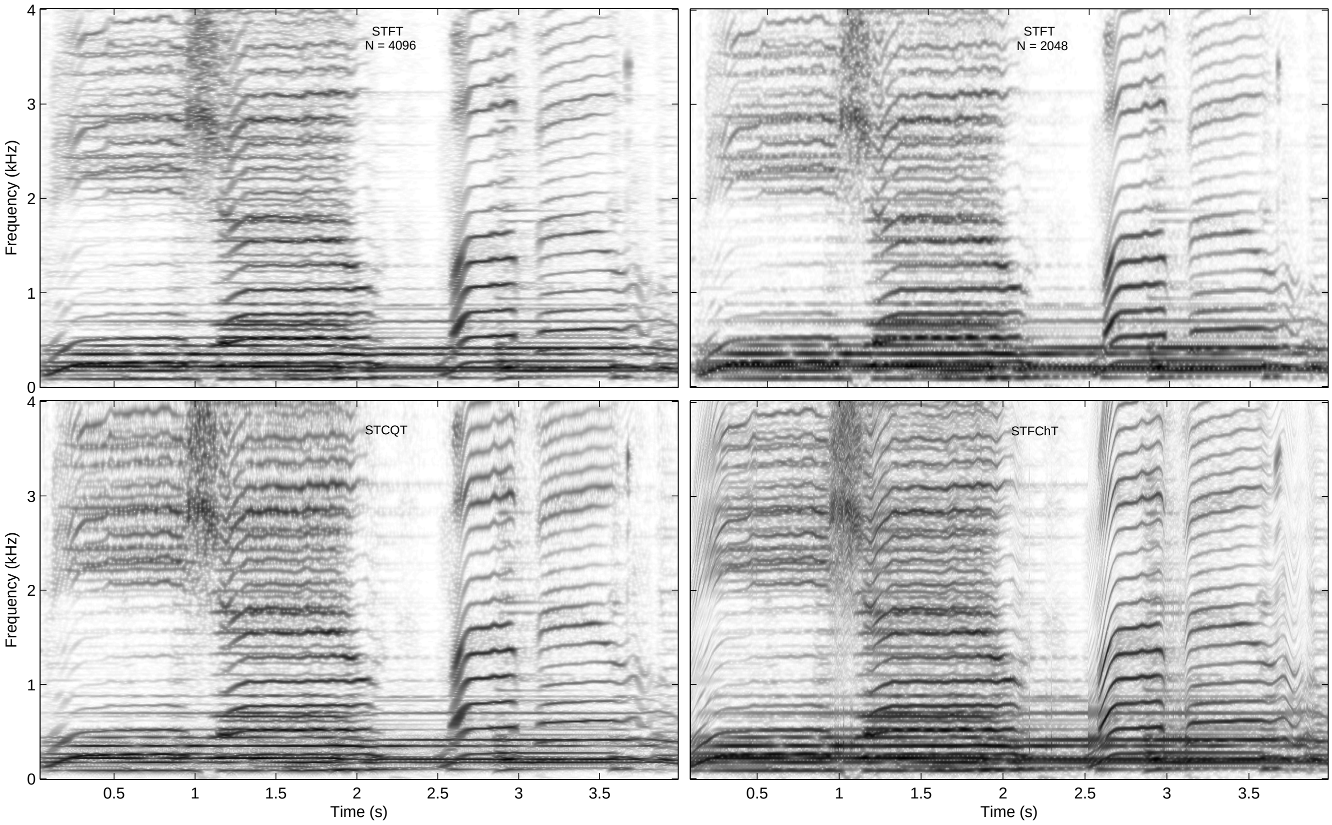

Figure 1: *Time-frequency representations comparison for an audio excerpt (which is used throughout the article) of the music file "pop1.wav" from the MIREX [12] melody extraction test set. It consist of three simultaneous prominent singing voices in the first part followed by a single voice in the second part, and a rather soft accompaniment whithout percussion. The representations depicted are: Spectrograms for window length of* 4096 *and* 2048 *samples at* $f_s = 44100$ *Hz, a Short Time CQT and a Short Time FChT using the herein proposed method. Note the improved time-frequency resolution for the most prominent singing voice in the latter representation.*

where $\phi_\alpha(t) = (1 + \frac{1}{2}\alpha t)\, t$, is a time warping function. This was formulated independently from the original work [7], so the properties are slightly different as will be indicated later. Notice that by the variable change $\tau = \phi_\alpha(t)$, the formulation becomes,

$$X(f,\alpha) = \int_{-\infty}^{\infty} x(\phi_\alpha^{-1}(\tau))\, e^{-j2\pi f \tau} d\tau, \qquad (2)$$

which can be regarded as the Fourier Transform of a time warped version of the signal $x(t)$, and enables an efficient implementation based on the FFT. The goal pursued is to obtain an acute representation of linear chirp signals of the form $x_c(t,f) = e^{j2\pi f \phi_\alpha(t)}$. Considering a limited analysis time support, the analysis basis is $\Gamma = \{\gamma_k\}_{k \in \mathbb{Z}}, \gamma_k = \phi'_\alpha(t)\, e^{j2\pi \frac{k}{\breve{T}} \phi_\alpha(t)}, t \in \left[\phi_\alpha^{-1}(-\frac{\breve{T}}{2}), \phi_\alpha^{-1}(\frac{\breve{T}}{2})\right]$. The inner product of the chirp and a basis element results in,

$$\begin{aligned}\left\langle x_{ch}(t, 2\pi \tfrac{l}{\breve{T}}), \gamma_k \right\rangle &= \frac{1}{\breve{T}} \int_{\phi_\alpha^{-1}(-\frac{\breve{T}}{2})}^{\phi_\alpha^{-1}(\frac{\breve{T}}{2})} \phi'_\alpha(t)\, e^{j2\pi \frac{l-k}{\breve{T}} \phi_\alpha(t)} dt \\ &= \frac{1}{\breve{T}} \int_{-\frac{\breve{T}}{2}}^{\frac{\breve{T}}{2}} e^{j2\pi \frac{l-k}{\breve{T}} \tau} d\tau = \delta[l-k], \qquad (3)\end{aligned}$$

which denotes that only one element of the basis represents the chirp. Note that the limits of integration include an integer number of cycles of the chirp, in the warped and the original time interval.

In [7] the basis are designed to be orthonormal, in order to obtain perfect reconstruction directly from the analysis basis. However, its response to a chirp of constant amplitude is not represented by a single element. It is important to note that when the signal is windowed the orthogonality disappears so as the perfect reconstruction. In a similar way, result given by equation 3 does not hold anymore. To that end, it is worth defining a more appropriate goal, that is what kind of response would be desirable for a time limited chirp. The solution proposed in this work permits to achieve a delta convolved with the Fourier Transform of a well-behaved analysis window. This motivates the above definition of the analysis basis Γ and the application of the analysis window to the time warped signal (which also differs from [7]). Then, the proposed FChT for a time limited support is,

$$X_w(f,\alpha) = \int_{-\infty}^{\infty} x(t)\; w(\phi_\alpha(t))\; \phi'_\alpha(t)\; e^{-j2\pi f \phi_\alpha(t)} dt \qquad (4)$$

where $w(t)$ stands for a time limited window, such as Hann.

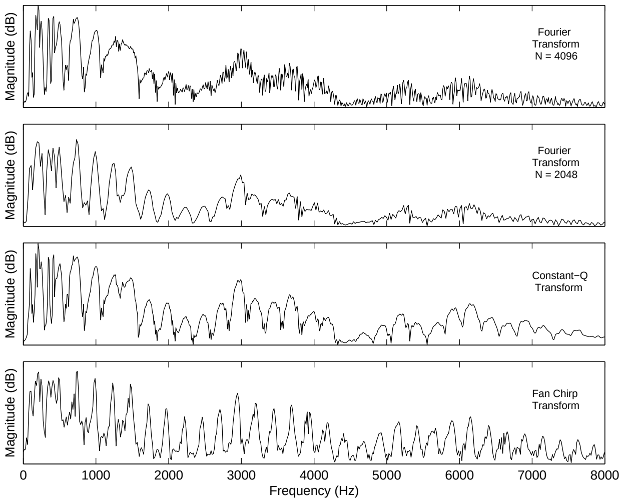

Figure 2: *Time-frequency representations comparison for a frame of the audio excerpt at time instant $t = 2.66s$. The prominent singing voice has a very high pitch change rate at this instant. This produces a blurry representation of the strongly non-stationary higher harmonics in the Fourier Transform. The representation of these harmonics is improved with the CQT because of the use of shorter time windows in high frequency. The FChT exhibits a more clear harmonic peak structure for the fan chirp rate that best represents the pitch change rate of the singing voice.*

Consider the case of a signal composed of L harmonically related linear chirps, $x_{hc}(t, f_0, L) = \sum_{k=1}^{L} e^{j2\pi k f_0 \phi_\alpha}$. All components share the same fan chirp rate α, so applying the appropriate warping ϕ_α delivers constant frequency harmonically related sinusoidal components. The FChT representation therefore shows a sharp harmonic structure as it is composed of deltas convolved with the Fourier Transform of the window.

2.2. Discrete time implementation

As stated before, the FChT of a signal $x(t)$ can be computed by the Fourier Transform of the time warped signal $\breve{x}(t) = x\left(\phi_\alpha^{-1}(t)\right)$, where

$$\phi_\alpha^{-1}(t) = -\frac{1}{\alpha} + \frac{\sqrt{1+2\alpha t}}{\alpha}. \tag{5}$$

This warping function transforms linear chirps of instantaneous frequency $\nu(t) = (1+\alpha t)\, f$ into sinusoids of frequency $\breve{\nu}(t) = f$. In practice, the original signal is processed in short time frames. In order to properly represent it with its warped counterpart, temporal warping is implemented by adopting the following criteria. After the time warping, the frequency of the resulting sinusoid is the frequency of the linear chirp at the centre of the analysis window. Besides, the amplitude value of the warped signal remains unchanged in the central instant of the window. Note this implies that the duration of the original signal and the warped signal may be different and is not imposed in [7].

Let $x[n] = x((n-(N-1)/2)/f_s - t_a)$ be a finite length signal frame at central time t_a, where $n = 0, \ldots, N-1$. When working with discrete signals, the temporal warping is implemented by non-uniform resampling of $x[n]$. If the sampling frequency is f_s, the frame duration is $T = (N-1)/f_s$. The time instant corresponding to the n-th sample of $x[n]$ is defined as $t_n = (n-(N-1)/2)/f_s$. Thus, the time domain of the signal is $\mathcal{D}_x = [-T/2, T/2]$. Similarly, let $\breve{x}_\alpha[m]$ be the time warped signal, with $m = 0, \ldots, M-1$, and sampling frequency $\breve{f}_s$. Its duration is then $\breve{T} = (M-1)/\breve{f}_s$ and the time instant of the m-th sample is defined as $\breve{t}_m = (m-(M-1)/2)/\breve{f}_s$. The warped time domain is $\mathcal{D}_{\breve{x}} = [-\breve{T}/2, \breve{T}/2]$. To compute the sample corresponding to time instant $\breve{t}_m$ of the warped signal, it is necessary to evaluate $x[n]$ at time instant $t_m = \phi_\alpha^{-1}(\breve{t}_m)$. As this instant may not coincide with a sampling time, the evaluation must be done using some interpolation technique. Time warping process is illustrated in figure 3. The last step of the FChT is to apply an analysis window to the time warped signal and compute the DFT.

The transform parameters are the number of samples M and the sampling rate $\breve{f}_s$. They should be selected in order to avoid aliasing in the resampling process of the original signal $x[n]$. Given a sampling rate f_s, suppose that the signal is band limited to $f_{\max}$ (with $f_{\max} \leq f_s/2$). To set M and $\breve{f}_s$ to avoid aliasing, the maximum resamplig period $T_s^{\max}$ must fulfill $T_s^{\max} \leq 1/(2f_{\max})$. If $\alpha \geq 0$, the previous condition is met if,

$$(\phi_\alpha^{-1})'\left(-\frac{\breve{T}}{2}\right) \leq \frac{T_s^{\max}}{\breve{T}_s} = \frac{\breve{f}_s}{2f_{\max}},$$

where $\breve{T}_s = 1/\breve{f}_s$ is the sampling period of $\breve{x}_\alpha[m]$. Using equation 5 and considering that $\breve{T} = (M-1)/\breve{f}_s$, the condition becomes,

$$\breve{f}_s \geq \frac{2f_{\max}}{\sqrt{1-|\alpha|\frac{M-1}{\breve{f}_s}}}.$$

The length N of the analysis window must be large enough to be able to interpolate $x[n]$ at every time instant t_m. Specifically, it must fulfill $T/2 \geq \max_m |t_m|$. If $\alpha \geq 0$, $\max_m |t_m| = |t_0| = |\phi_\alpha^{-1}(-\breve{T}/2)|$ and thus $T/2 \geq |\phi_\alpha^{-1}(-\breve{T}/2)|$, which leads to

$$N > 2f_s \frac{1-\sqrt{1-|\alpha|\frac{M-1}{\breve{f}_s}}}{|\alpha|}.$$

In the current implementation, the discrete signal, originally sampled at 44100 Hz, is first lowpass filtered to limit the spectral content up to $f_{\max} = 10000$ Hz, and then upsampled to double the sampling rate, so $f_s = 88200$ Hz. The computation of the warped samples is done using linear interpolation. The above mentioned upsampling is performed to obtain a more accurate and efficient interpolation. The maximum absolute value of the fan chirp rate employed is $\alpha_{\max} = 6$. With this parameters, values of $M = 2048$, $\breve{f}_s = 30000$ Hz and $N = 10000$ meet the above conditions. Note that proposed implementation permits to choose M as a small power of two to take advantage of the FFT efficiency.

Another consideration regarding the implementation, is that time warping design is performed numerically based on relative instantaneous frequency functions. More precisely, the design begins with the selection of the warping instantaneous frequency $f_r[n]$ for each sample. Then, the function $\phi[n]$ is obtained by numerical integration of $f_r[n]$. Finally the function $\phi^{-1}[n]$, needed to compute the resampling times, is obtained by numerical inversion. This allows the implementation of arbitrary warpings functions instead of only linear warpings.

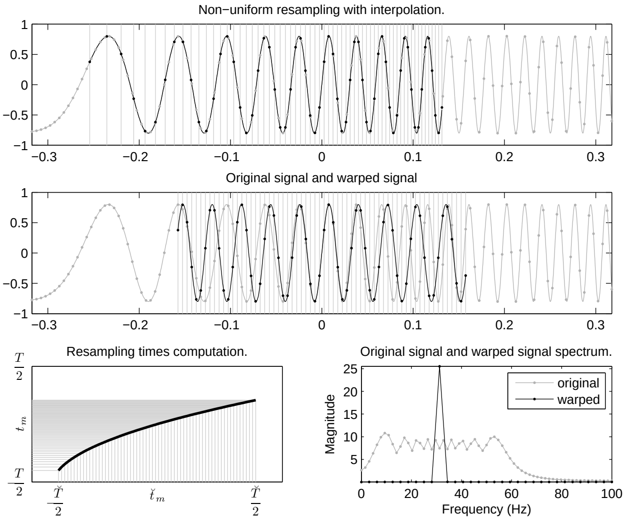

Figure 3: *Warping process illustration. A sinusoid is obtained by appropriately warping a linear chirp. Note that the central time instant remains the same and the time supports are different. The FChT of the linear chirp shows a sharp high peak.*

3. FAN CHIRP TRANSFORM FOR MUSIC REPRESENTATION

In a practical situation, real signals such as speech or music sounds can be assimilated to harmonically related linear chirps only within short time intervals, where the evolution of frequency components can be approximated by a first order model. This suggests the application of the FChT to consecutive short time signal frames, so as to build a time-frequency representation as a generalization of the spectrogram [7]. In the monophonic case, a single value of the fan chirp rate α that best matches the signal pitch variation rate should be determined for each frame. This is the key factor to obtain a detailed representation using the FChT. Different approaches could be followed, such as predicting the pitch evolution and estimating α as the relative derivative of the pitch [7].

In the polyphonic case, there is no single appropriate value of α, because the multiple harmonic sounds present are likely to change their fundamental frequency (f_0) differently within the analysis frame. For this reason, a multi-dimensional representation for each frame seems better suited in this case, consisting in several FChT instances with different α values. A given FChT is tuned to represent one of the harmonic sounds with reduced spectrum spread, whereas poorly representing the remaining ones. The selection of a reduced set of α values for each frame that produce the better representation of each sound present, can be tackled by means of sinusoidal modeling techniques as in [10]. In this work a straightforward exhaustive approach is adopted, that consists in computing a dense (f_0, α) plane and selecting the best chirp rates based on pitch salience. In addition, the pitch salience computation from the FChT produces itself a detailed representation of the melodic content of the signal, that can be useful in several applications. This is described in detail in the following section.

4. PITCH SALIENCE COMPUTATION

The aim of pitch salience computation is to build a continuous function that gives a prominence value for each fundamental frequency in a certain range of interest. Ideally it shows pronounced peaks at the positions corresponding to the true pitches present in the signal frame. This detection function typically suffers from the presence of spurious peaks at multiples and submultiples of the true pitches, so some sort of refinement is required to reduce this ambiguity. A common approach for pitch salience calculation is to define a fundamental frequency grid, and compute for each frequency value a weighted sum of the partial amplitudes in a whitened spectrum. A method of this kind was used in [13] for melody extraction, which is formulated in the following according to the log-spectrum gathering proposed in [8].

4.1. Gathered log-spectrum (GlogS)

The salience of a given fundamental frequency candidate f_0 can be obtained by gathering the log-spectrum at the positions of the corresponding harmonics as [8],

$$\rho_0(f_0) = \frac{1}{n_H} \sum_{i=1}^{n_H} log|S(if_0)|, \tag{6}$$

where $|S(f)|$ is the power spectrum and n_H is the number of harmonics that are supposed to lie within the analysis bandwidth. Linear interpolation from the discrete log-spectrum is applied to estimate the values at arbitrary frequency positions. The logarithm provides better results compared to the gathering of the linear spectrum. This makes sense, because the logarithm function can be regarded as a kind of whitening in order to make the pitch salience computation more robust against formant structure and noise. With this respect, it is interesting to note that a p-norm with $0 < p < 1$ is also appropriate and shows similar results. Note that this seems coherent with the use of more robust norms which is advocated in sparsity research. Therefore, the actual implementation is $log(\gamma|S(if_0)| + 1)$ which adds the flexibility to custom the norm applied by means of the γ parameter[1].

4.2. Postprocessing of the gathered log-spectrum

The harmonic accumulation shows peaks not only at the position of the true pitches, but also at multiples and submultiples (see figure 4). To handle the ambiguity produced by multiples, the following simple non-linear processing is proposed in [8],

$$\rho_1(f_0) = \rho_0(f_0) - \max_{q \in \mathbb{N}} \rho_0(f_0/q). \tag{7}$$

This is quite effective in removing pitch candidates multiples of the actual one (as can be seen in figure 4). When dealing with monophonic signals this suppression is enough. If pitch estimation is obtained as the position of the maximum of $\rho_1(f_0)$, $\hat{f}_0 =$ arg max $\rho_1(f_0)$, submultiple spurious peaks do not affect the estimation because their amplitude is necessarily lower than for the true pitch. However, in the polyphonic case, submultiple peaks should also be removed. For this reason, the detection function is further processed to remove the (k-1)-th submultiple according to,

$$\rho_2(f_0) = \rho_1(f_0) - a_k \, \rho_1(kf_0) \tag{8}$$

[1] Higher values tend to a 0-norm while lower values tend to a 1-norm. All the results reported correspond to $\gamma = 10$.

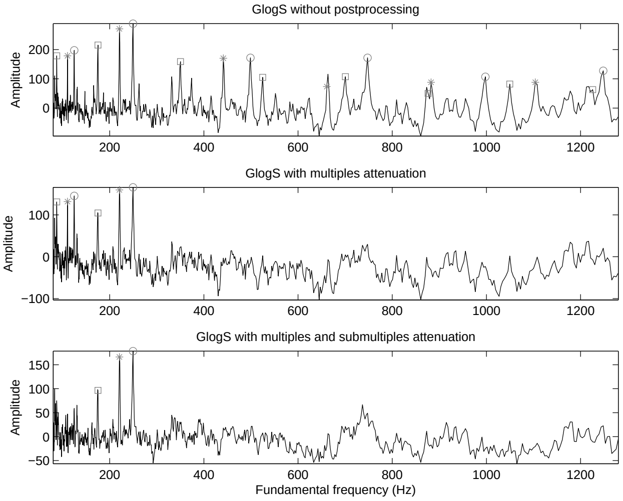

Figure 4: *Normalized gathered log spectrum and the postprocessing stages for a frame of the audio excerpt at $t = 0.36s$, with three prominent simultaneous singing voices and low accompaniment. Positions of each corresponding f_0, multiples and first submultiple are also depicted.*

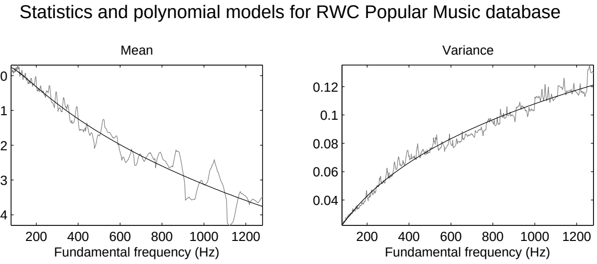

Figure 5: *Gathered log spectrum normalization model.*

where a is an attenuation factor. From the simulations conducted it turned out that removing only the first submultiple ($k = 2$) is commonly sufficient for melodic content visualization and melody detection (see figure 4). For a single ideal harmonic sound, following a similar reasoning to that of [8] appendix B, it can be shown that the attenuation factor is $a_2 = 1/2$. However, it can also be shown that the variance of $\rho_0(f_0)$ is proportional to fundamental frequency (see also [8] appendix B). In practice a true pitch peak can be unnecessarily attenuated due to the large variance at its multiple, so a more conservative attenuation factor is preferred. Slightly better results were obtained over polyphonic music for $a_2 = 1/3$, and this is the value used for the reported results.

4.3. Normalization of the gathered log-spectrum

The variance increase with f_0 is an undesired feature. When applied to melodic content visualization different frequency regions are unbalanced and it leads to incorrect detections when pursuing melody extraction. For this reason, the last step in pitch salience computation is to normalize $\rho_2(f_0)$ to zero mean and unit variance. To do this, the mean and the variance of $\rho_2(f_0)$ are collected at each f_0 for every frame from a music collection (the complete RWC Popular Music Database [14] was used for this purpose). Each one of these statistics are then approximated by a second order polynomial, as illustrated in figure 5. The polynomials evaluated at each f_0 are the model used to obtain a normalized gathered log-spectrum $\overline{\rho}_2(f_0)$. The fundamental frequency grid used is logarithmically spaced with 192 points per octave.

4.4. Fan chirp rate selection using pitch salience

As early mentioned, the α values that best represent the different harmonic sounds in a signal frame are selected by means of pitch salience. Several FChT instances are computed for each frame using different α values. For each FChT a gathered log spectrum is calculated as described above, so as to build a dense pitch salience plane $\overline{\rho}_2(f_0, \alpha)$. See figure 6 for an example of this dense pitch salience plane. Given a sound source of fundamental frequency $\hat{f}_0$, the energy of its harmonics is more concentrated at the FChT instance corresponding to the best matching α value $\hat{\alpha}$. Therefore, the value of $\overline{\rho}_2(\hat{f}_0, \hat{\alpha})$ is the highest among the different available α values. For this reason, a different α value is selected for each f_0 in the grid, giving a single pitch salience value for each f_0 (see figure 6). The reduced set of α values can be selected according to their corresponding pitch salience.

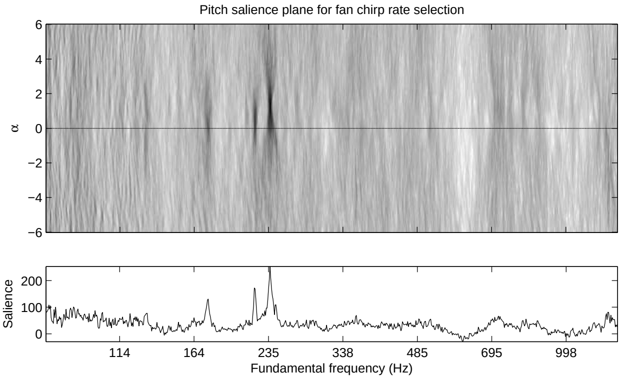

Figure 6: *Pitch salience plane $\overline{\rho}_2(f_0, \alpha)$ for a frame of the audio excerpt at $t = 0.27s$. Prominent salience peaks (darker regions) can be distinguished corresponding to the three singing voices. Note that two of them are located approximately at $\alpha = 0$ and one at $\alpha = 1.3$. This indicates that two of the voices are quite stationary within the frame while the other is increasing its pitch. The maximum pitch salience value for each f_0 is also depicted.*

5. MULTI-RESOLUTION FAN CHIRP TRANSFORM

5.1. Constant Q Transform

The CQT provides a variable time-frequency resolution in such a way that good frequency resolution in the low frequencies and good time resolution in the high frequencies is obtained simultaneously. It is calculated using a different time window length for each frequency bin. The window length is chosen so as to contain the same effective number of oscillations of the corresponding frequency. Among the various existing implementations of the CQT, in this article the IIR CQT [15] is selected, which gives a good compromise between computational cost and design flexibility.

5.2. Combination with the Fan Chirp Transform

As mentioned before, the benefits from using a CQT transform instead of a fixed time window rely on the non stationarity of the signal, such as voice. The higher partials of a harmonic chirp are particularly non stationary. Using the classical STFT their representation is blurry. The use of a CQT with an adequate value of Q makes it possible to achieve a good time-frequency resolution compromise for non stationary signals along all the spectrum.

Although the FChT using linear warpings is devised to analyse non stationary signals, the analysis may be further improved because of two main reasons. The first is that the range of fan chirp rates α used in the analysis is discretized. If the signal chirp rate does not closely matches any of the available α values, higher partials behave non stationarely after the time warping. Using the CQT, with a relatively high Q value, alleviates this problem. As a result, the number of analysis α values can be diminished, reducing the computational cost at no significant performance loss. The second reason is that considering a linear evolution of the instantaneous fundamental frequency could be a crude approximation for a signal frame with a non linear pitch evolution. In this case, the warping once again outputs a non stationary signal and the CQT is beneficial, specially for the higher partials (see figure 7). However, the addition of the CQT produces some degradation in the analysis of a signal with a linear pitch evolution, so a relatively high Q value should be chosen in order to obtain a good performance for a wider set of signals.

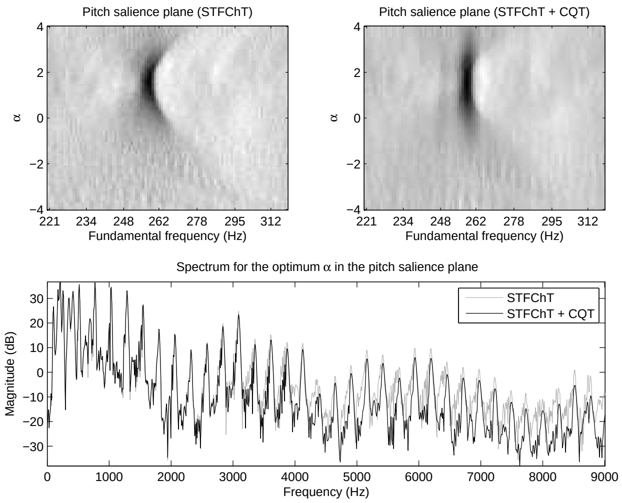

Figure 7: *Example of the FChT analysis including the CQT for a frame of the audio excerpt at $t = 2.68s$. The singing voice has a high pitch curvature at this instant. Adding the CQT enhances the representation of the higher partials. Note also that the peak region in the pitch salience plane is much concentrated in frequency and the range of α values with significant salience is widen which allows a more sparse α discretization.*

6. APPLICATIONS

6.1. Pitch visualization

The pitch salience function of each frame obtained as described in subsection 4.4 is used to build a F0gram that shows the temporal evolution of pitch for all the harmonic sounds of a musical signal, as can be seen in figure 8. Note that even in the case that two sources coincide in time and frequency they can be correctly represented if their pitch change rate is different, which is observable at time $t = 0.23$ s. It can also be seen the precise pitch contour evolution obtained, even for severe pitch fluctuations. Additionally, the gathered log spectrum normalization provides a balanced contrast of the F0gram, without spurious noticeable peaks when no harmonic sound is present. A drawback of the selected pitch salience function is that it tends to underestimate low frequency harmonic sounds with a small number of prominent partials. This is the case for the accompaniment in the selected example, that only appears when no singing voice is present.

This kind of visualization tool can be useful itself for analyzing performance expressive features such as glissando, vibrato, and pitch slides, that turn out clearly distinguishable.

6.2. Melody detection

A frame based melody detection evaluation was conducted to asses the usefulness of the proposed FChT-based method for music analysis. To do this, two different labeled databases were considered, namely the 2004-2005 MIREX [12] melody extraction test set (only vocal files) and the RWC Popular Music database [14]. The former comprises 21 music excerpts while the latter contains 100 complete songs, for a total duration of 8 minutes and 6 hours respectively. The RWC is a more difficult dataset due to higher polyphony (including prominent percussion) and dynamic compression, whereas the MIREX although much smaller is publicly available and more diverse (e.g. includes opera). For each frame the most prominent F0gram peaks were selected and their corresponding fundamental frequencies were considered as main melody pitch candidates. Only those frames for which the melody was present according to the labels were taken into account to compute the evaluation measure according to,

$$\text{score}(f_0) = \min\{1, \max\{0, (\text{tol}_{\max} - \Delta f_0)/(\text{tol}_{\max} - \text{tol}_{\min})\}\}$$

where $\Delta f_0 = 100|f_0 - f_0^{gt}|/f_0^{gt}$ is the relative error between a candidate and the ground truth, and the tolerances $\text{tol}_{\max}$ and $\text{tol}_{\min}$ correspond to 3% and 1% respectively. This represents a strict soft thresholding of the estimation performance[2].

Considering that the pitch of a main melody is not equiprobable in the f_0 selected range, it is reasonable to include this a priori information in the selection of candidates. To do this, salience is weighted by a gaussian centered at MIDI note 60 (C4) and with a standard deviation of an octave and a half. This values were selected considering the main melody pitch distribution of the databases (see figure 9), but setting the model parameters to favour generalization (in particular tripling the standard deviation). In figure 10 an example of the melody detection result is depicted. Note that the first candidate correctly follows the main melody

[2] Note that this performance measure is stricter and more discriminative than the binary 3% threshold used in MIREX. This was devised in order to better distinguish the performance of the different methods evaluated and considering that the ground truth labels are not perfectly accurate.

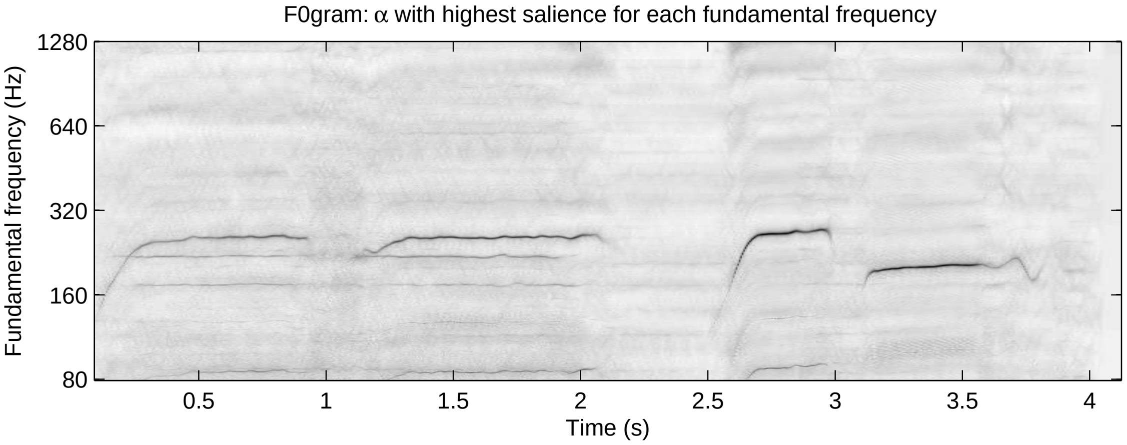

Figure 8: *Example of melodic content visualization for the selected audio excerpt. The pitch contour of the three simultaneous singing voices followed by a single voice can be clearly appreciated. Note that the 2nd submultiple is noticeable as it has not been attenuated.*

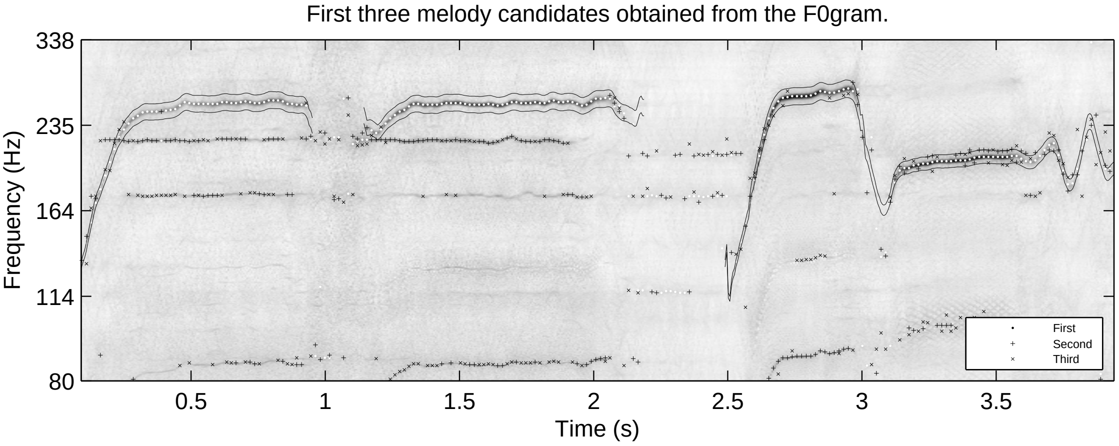

Figure 10: *Example of melody detection. The three first candidates and a* ±3% *band centered at the label are displayed. First candidate tends to correctly match the melody while the remianing ones usually chose other secondary voice. Note that not attenuated submultiples mislead the detection of these secondary sources.*

when it is the most prominent. The fan chirp rate estimated for the first candidate matches the actual value extracted from labels, as shown in figure 11. This information can be further exploited, for example when performing the temporal tracking.

Table 1 shows the scores obtained for the different methods when applied to each database. The optimal parameter values were grid searched for every method. It turned out that for the FChT-based methods similar results were obtained for parameters around the values specified in section 2.2 and different number of α values. Results reported correspond to 15 fan chirp rates. In this case, running time in a desktop computer reaches real time for a Matlab and C code implementation of the FChT [3].

The results indicate that the proposed improvements to the pitch salience computation (submultiple attenuation, normalization and a priori model) contribute to a significant performance increase (compare the two STFT results). Further performance improvement is achieved by the use of the FChT. Although the combination of the FChT and the CQT leads to a better representation of some frames, its impact into the melody detection results is marginal. This is related to the fact that higher partials are less influential in the pitch salience value given their lower amplitude. Additionally, the number of frames with highly non-linear pitch evolution represent a small amount of the total.

7. CONCLUSIONS AND FURTHER WORK

In this work, a non classical time-frequency representation, namely the FChT, was applied to polyphonic music analysis. The formulation presented provides an acute representation of harmonically related linear chirp signals. The implementation introduced was devised to be computationally manageable and enables the use of non linear warpings. Both the formulation and the implementation

[3] The implementation is available at `http://iie.fing.edu.uy/~pcancela/fcht`.

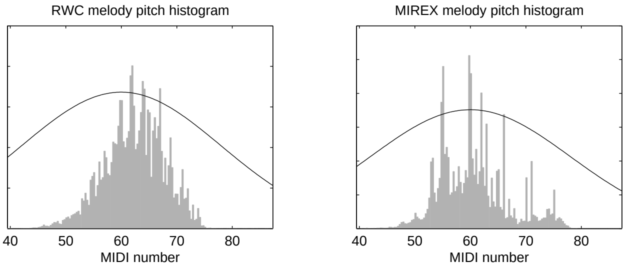

Figure 9: *Pitch preference function (mean = 60, stdev = 18) and melody pitch histogram for RWC Popular and MIREX data.*

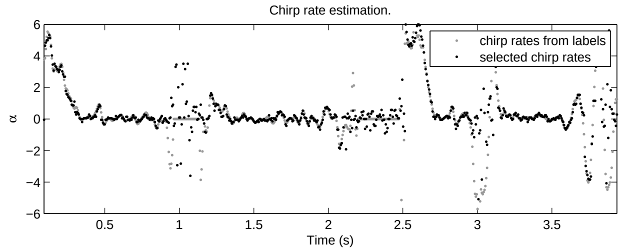

Figure 11: *Estimated fan chirp rate for the first candidate and actual value extracted from labels.*

differ from early proposals [7]. The combination with the CQT was explored to build a multi-resolution FChT which improves the representation of harmonic signals with non linear pitch variations. In order to precisely represent the different pitched sources in a signal using the FChT an existing method based on pitch salience was adopted, which was improved and adapted to handle polyphonic music. The pitch salience computation from the FChT provides itself a useful melodic content visualization tool. Results obtained for a frame based melody detection evaluation indicate that the introduced F0gram is very promising as a front-end for music analysis. A similar method to the herein described (adding temporal tracking) was submitted to MIREX 2008 Audio Melody Extraction Contest [13], performing best on Overall Accuracy.

Several applications will be tackled in future work (e.g. multiple f_0 detection, sound source separation). Additionally, the usefulness of non linear warpings will be explored.

Table 1: *Melody detection scores (%) from 1st to 5th candidate. Except for the first method where only multiple attenuation is applied, all the others include the proposed improvements: submultiple attenuation, $\rho_2(f_0)$ normalization and f_0 preference function.*

MIREX	**STFT plain**	**STFT**	**CQT**	**FChT**	**FChT +CQT**
1	71.32	75.72	75.50	81.92	82.09
1-2	78.90	82.82	81.51	87.32	87.41
1-3	82.63	85.95	84.35	89.67	89.82
1-4	85.14	87.71	86.07	91.09	91.24
1-5	86.80	88.92	87.31	92.11	92.19
RWC	**STFT plain**	**STFT**	**CQT**	**FChT**	**FChT +CQT**
1	48.60	63.19	-	68.21	68.52
1-2	59.50	72.96	-	76.85	77.50
1-3	65.62	77.53	-	80.81	81.55
1-4	69.86	80.33	-	83.27	84.07
1-5	73.05	82.32	-	85.05	85.84

8. REFERENCES

[1] P. Flandrin, *Time-Frequency/Time-scale Analysis*, Wavelet Analysis and Its Applications. Academic Press, 1999.

[2] S. A. Abdallah and M. D. Plumbley, “Unsupervised analysis of polyphonic music by sparse coding,” *IEEE Trans. Neural Networks*, vol. 17, no. 1, pp. 179–196, 2006.

[3] L. Cohen, “Time-frequency distributions: a review,” *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941 – 981, 1989.

[4] J. C. Brown, “Calculation of a constant Q spectral transform,” *JASA*, vol. 89, no. 1, pp. 425–434, 1991.

[5] S. Mann and S. Haykin, “The chirplet transform: physical considerations,” *IEEE Transactions on Signal Processing*, vol. 41, no. 11, pp. 2745–2761, 1991.

[6] L. B. Almeida, “The fractional fourier transform and time-frequency representations,” *IEEE Transactions on Signal Processing*, vol. 42, no. 11, pp. 3084 – 3091, 1994.

[7] L. Weruaga and M. Képesi, “The fan-chirp transform for nonstationary harmonic signals,” *Signal Processing*, vol. 87, no. 6, pp. 1504–1522, 2007.

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

[9] C. Kereliuk and P. Depalle, “Improved hidden markov model partial tracking through time-frequency analysis,” in *Proc. of the 11th Conference on Digital Audio Effects (DAFx-08), Espoo, Finland*, September 2008.

[10] M. Bartkowiak, “Application of the fan-chirp transform to hybrid sinusoidal+noise modeling of polyphonic audio,” in *16th European Signal Processing Conference*, 2008.

[11] P. Zhao, Z. Zhang, and X. Wu, “Monaural speech separation based on multi-scale fan-chirp transform,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.*, 2008, pp. 161–164.

[12] J. S. Downie, “The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research,” *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[13] P. Cancela, “Tracking melody in polyphonic audio,” in *Music Information Retrieval Evaluation eXchange*, 2008.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “Rwc music database: Popular, classical, and jazz music databases,” in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, October 2002, pp. 287–288.

[15] P. Cancela, M. Rocamora, and E. López, “An efficient multi-resolution spectral transform for music analysis,” in *International Society for Music Information Retrieval Conference, 10th. ISMIR 2009. Kobe, Japan*, 2009, pp. 309–314.