HIGH-DEFINITION TIME-FREQUENCY REPRESENTATION BASED ON ADAPTIVE COMBINATION OF FAN-CHIRP TRANSFORMS VIA STRUCTURE TENSOR

Maurício do Vale Madeira da Costa *

LTCI, Télécom Paris
Institut Polytechnique de Paris
Paris, France
SMT, PEE/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
mauricio.costa@smt.ufrj.br

Luiz Wagner Pereira Biscainho *

SMT, DEL/Poli & PEE/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
wagner@smt.ufrj.br

ABSTRACT

This paper presents a novel technique for producing high-definition time-frequency representations by combining different instances of short-time fan-chirp transforms. The proposed method uses directional information provided by an image processing technique named structure tensor, applied over a spectrogram of the input signal. This information indicates the best analysis window size and chirp parameter for each time-frequency bin, and feeds a simple interpolation procedure, which produces the final representation. The method allows the proper representation of more than one sound source simultaneously via fan-chirp transforms with different resolutions, and provides a precise reproduction of transient information. Experiments in both synthetic and real audio illustrate the performance of the proposed system.

1. INTRODUCTION

Time-Frequency Representations (TFRs) of audio signals have been used for decades in all sorts of audio processing tasks. Such representations allow to observe the temporal evolution of the frequency content present in a given signal by applying a time-to-frequency mapping, e.g. the Fourier transform, to time frames of the signal [1]. In this context, the availability of an appropriate time-frequency representation for a song recording can improve several tasks in Music Information Retrieval (MIR) [2, 3], e.g. automatic transcription, rhythmic analysis, identification of instruments, sound source separation, etc.

The main problem concerning TFRs is the incapability of representing both time and frequency content with arbitrarily high resolutions simultaneously, which is dictated by the uncertainty principle [1]. For instance, by increasing the length of the analysis window of a spectrogram, a higher frequency resolution is achieved, while the time resolution decreases. Therefore, one should choose the analysis window's length within a compromise.

Having high-definition TFRs is essencial for many MIR tasks [2, 3]. Although the uncertainty principle holds true, the TFR field has been receiving several contributions aiming at providing better time-frequency resolution (i.e. sparser representations) for audio tasks. A common approach is to perform some sort of combination of TFRs, taking advantage of characteristics of audio signals, in an attempt to locally optimize their representation [4, 5, 6, 7, 8, 9].

The Fan-Chirp Transform (FChT) [10, 11], in its discrete short-time version, can provide a sparse representation for signals containing harmonic content with fast fundamental frequency variation. This transform, presented here in Section 2, uses a basis composed of complex exponentials whose frequency varies linearly in time, hence assuming linearity within the excerpts (time frames) under analysis. When the transform properly models the variation present in the signal, its tonal frequency content can be sparsely described. This potentially overcomes the problem of energy smearing observed when a signal with fast frequency variations is represented by a spectrogram.

The main disadvantage of this transform is that it may only well represent one sound source at a time, since the whole exponential basis used follows a unique fundamental frequency variation rate. Unless the sources share the same slope parameter, an only source will be sparsely represented in detriment to the others. Therefore, the problem of dealing with multiple sources can only be tackled by combining multiple instances of Short-Time Fan-Chirp Transform (STFChTs), in a way that the best representations for each time-frequency bin remain in the final TFR.

To the authors' knowledge, unfortunately no TFR found in the literature seems capable of precisely representing polyphonic signals containing fast frequency variation and still preserve a natural and smooth representation of the magnitude of frequency lines. When the standard spectrogram is used for this matter, especially with large analysis windows, a blurred representation results whenever a frequency line presents a steep slope. For instance, this undesired effect is frequently observed in vocal signals, whose frequency content changes considerably fast. Having an appropriate representation is an obvious requirement of tasks involving signals with such characteristics.

In this work, we combine bins of different precomputed TFRs using the information provided by an image processing technique, namely, the structure tensor. This technique (presented in Section 3) allows to compute the predominant orientation angle of edges present in a given image [12, 13, 14, 15]. In the case of musical signals, the frequency lines found in a spectrogram can be interpreted as edges, whose direction can then be estimated. This procedure has two outputs, for each time-frequency bin (pixel): a priority angle, which indicates the direction of the edge; and the anisotropy measure, which informs how relevant is that direction

* The authors thank CAPES and CNPq Brazilian agencies for funding this work.

in terms of edginess. We use the information provided by the structure tensor to choose the best TFRs available (in terms of window size and slope parameter) and combine them bin-wise. This procedure allows one to attain a TFR with high resolution in frequency lines with any desired slope and in sharp transient information.

A flow-chart of the proposed method is depicted in Figure 1. Firstly, the audio signal x is processed to generate a set of different STFChTs using predetermined sets of chirp slopes and analysis window sizes, $\boldsymbol{\alpha}$ and $\mathbf{K}$, respectively. Then, all TFRs are interpolated[1] and assembled in a four-dimensional tensor $\underline{\mathbf{X}}$. From the structure tensor of the standard spectrogram $\mathbf{X}$, parameters $\mathbf{A}$, containing the preferable chirp rates (directional information), and $\mathbf{C}$, containing the preferable window sizes, are computed. Finally, a simple linear combination of the set of TFRs is performed according to $\mathbf{A}$ and $\mathbf{C}$ for each time-frequency bin (as described in Section 4), resulting in the combined TFR $\mathbf{X}^{\mathrm{Comb}}$.

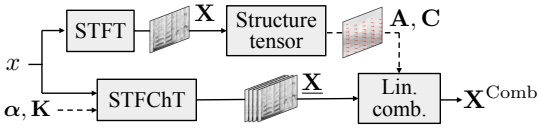


Figure 1: *Flow-chart of the proposed method.*

A set of experiments presented in Section 5 attest the method's performance, and conclusions are drawn in Section 6.

2. TIME-FREQUENCY REPRESENTATIONS

2.1. The Spectrogram

The most used time-frequency transform is the popular spectrogram, which is comprised of the magnitude[2] of the Short-Time Fourier Transform (STFT):

$$X(\tau, f) \triangleq \left| \int_{-\infty}^{\infty} x(t) w(t-\tau) \mathrm{e}^{-j2\pi f t} \mathrm{d}t \right|, \tag{1}$$

where $|\cdot|$ denotes the magnitude of its complex argument, w is a real-valued analysis window, and x is a real valued signal. The window w can be continuously shifted in time, providing a spectrum for each instant τ of the windowed version of $x(t)$.

The discrete version of the spectrogram, $\mathbf{X} \in \mathbb{R}^{K \times M}$, follows the same principle of using a window to focus on some part of the signal. Considering that the time support of the analysis window is limited to K samples, the discrete spectrogram can be described by

$$X_{k,m} \triangleq \left| \sum_{n=0}^{K-1} x_{n-hm} w_n \mathrm{e}^{-j\frac{2\pi}{K}kn} \right|, \tag{2}$$

where $k \in \mathcal{K} \triangleq \{0, 1, 2, ..., K/2\}$ is the frequency index, $m \in \mathcal{M} \triangleq \{1, 2, 3, ..., M\}$ is the time index of the STFT, w_n is the analysis window with K samples used for computation of the spectrogram, and $h \in \mathbb{N}$ is the analysis hop.

As mentioned, this method, as well as any other TFR, has the limitation stated by the uncertainty principle [1]: a signal cannot be represented with arbitrarily high time and frequency resolutions simultaneously. As the length of the analysis window grows, a greater frequency resolution is achieved, as longer excerpts of the signal are projected into the complex exponentials. For the same reason, the time resolution decreases. Parts of the signal with fast variations become blurred in the time-frequency map, for they are integrated with their neighborhood. Another issue is related to frequency variation within the period of the analysis window, which spreads the energy frequency-wise.

In order to address this issue, the fan-chirp transform can be used, allowing for representing harmonic signals whose fundamental frequency varies linearly in time. As long as the analysis window is short enough for the signal to fit this model, a sparse representation of fast frequency variations is attained.

2.2. The Fan-Chirp Transform

In the continuous time domain, the fan-chirp transform $X^{\mathrm{FChT}}(f, \alpha)$ of a given signal $x(t)$ is defined in [11] as

$$X^{\mathrm{FChT}}(f, \alpha) \triangleq \int_{-\infty}^{\infty} x(t) \phi'_\alpha(t) \mathrm{e}^{-\mathrm{j}2\pi f \phi_\alpha(t)} \mathrm{d}t, \tag{3}$$

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

$$\phi_\alpha(t) = \left(1 + \frac{1}{2}\alpha t\right) t, \tag{4}$$

and α is the chirp rate parameter.

This transform can be interpreted as a modification of the Fourier transform (which corresponds to $\alpha = 0$). By applying the variable change $\tau = \phi_\alpha(t)$ to Equation (3), the time domain itself can be warped:

$$X^{\mathrm{FChT}}(f, \alpha) = \int_{-1/\alpha}^{\infty} x(\phi_\alpha^{-1}(\tau)) \mathrm{e}^{-\mathrm{j}2\pi f \tau} \mathrm{d}\tau, \tag{5}$$

where

$$\phi_\alpha^{-1}(t) = -\frac{1}{\alpha} + \frac{\sqrt{1 + 2\alpha t}}{\alpha}. \tag{6}$$

In Equation (5), it is possible to observe that the FChT has the same formulation of the Fourier transform (Equation (1)), with the differences that the input signal $x(t)$ is pre-warped in time, and the inferior integration limit is changed—in order to avoid aliasing, $x(t) = 0$ for $t \leq -1/\alpha$ [10] should be assured. This constrains the usable values of α inside an analysis window with K samples to be within the interval

$$-2F_\mathrm{s}/K \leq \alpha \leq 2F_\mathrm{s}/K. \tag{7}$$

The Short-Time Fan-Chirp Transform (STFChT) $\mathbf{X}^{\mathrm{FChT}}$ is implemented by applying the same windowing procedure employed to compute the spectrogram, after resampling [11] the input signal x, and can be described as

$$X^{\mathrm{FChT}}_{k,m,\alpha} \triangleq \left| \sum_{n=0}^{K-1} \tilde{x}_{\alpha, n-hm} w_n \mathrm{e}^{-j\frac{2\pi}{K}kn} \right|, \tag{8}$$

where $\tilde{x}_n$ is the discrete version of the time warped signal $x(\phi_\alpha^{-1}(\tau))$, as long as the aliasing condition has been satisfied.

[1] The different TFRs must be interpolated not only to have the same dimensions, but also share the same time and frequency axes.

[2] The spectrogram can also be defined as the squared-magnitude of the STFT.

In practice, since $x(t)$ is not available, $\tilde{x}_n$ must be obtained by resampling x_n. With this formulation, the FChT can profit from a fast implementation of the Discrete Fourier Transform (DFT), i.e. an FFT algorithm [11].

By applying this procedure, the input signal is modeled as a sum of linear chirps for each time frame, and to this end, a value of α must be estimated. This step is originally performed via an exhaustive search, in which a predetermined set of values of α is tested, and the choice of the best one is made by searching for the α which provides maximum sparsity. Another reliable and significantly faster way to perform this estimation is proposed in [15], which consists in using the structure tensor [12, 13] technique to estimate priority directions. This procedure can also be used to estimate multiple simultaneous directions, which is useful for signals with more than one sound source [15]. In this paper, the structure tensor will also be used to estimate the target α's; however, differently from what is done in [15], the estimations are performed locally in the TFR, and are used to guide a combination procedure.

3. THE STRUCTURE TENSOR

The structure tensor technique allows the computation of angles of edges present in a given image [12, 13, 14, 15]. The idea is to interpret the spectrogram[3] as an image whose pixels are its time-frequency bins. In this work, the absolute value of the spectrogram compressed by the fourth-root $\hat{\mathbf{X}} = \mathbf{X}^{\frac{1}{4}}$ is used instead of the logarithm, which will be explained latter in this section.

3.1. Computation of the Structure Tensor

Initially, two derivative versions of $\hat{\mathbf{X}}$ are computed by the application of partial derivatives with respect to time index m and frequency index k:

$$\hat{\mathbf{X}}^{\mathrm{m}} = \mathbf{X} * \mathbf{D}, \tag{9}$$

$$\hat{\mathbf{X}}^{\mathrm{k}} = \mathbf{X} * \mathbf{D}^{\mathrm{T}}, \tag{10}$$

where $\mathbf{D}$ is a discrete differentiation operator, more specifically the Sobel-Operator

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \tag{11}$$

and $*$ denotes the 2-dimensional convolution.

Then, $\hat{\mathbf{X}}^{\mathrm{m}}$ and $\hat{\mathbf{X}}^{\mathrm{k}}$ are combined, producing 4 other matrices:

$$\mathbf{T}^{11} = [\hat{\mathbf{X}}^{\mathrm{m}} \odot \hat{\mathbf{X}}^{\mathrm{m}}] * \mathbf{G} \tag{12}$$

$$\mathbf{T}^{12} = \mathbf{T}^{21} = [\hat{\mathbf{X}}^{\mathrm{k}} \odot \hat{\mathbf{X}}^{\mathrm{m}}] * \mathbf{G} \tag{13}$$

$$\mathbf{T}^{22} = [\hat{\mathbf{X}}^{\mathrm{k}} \odot \hat{\mathbf{X}}^{\mathrm{k}}] * \mathbf{G}, \tag{14}$$

where operator $\odot$ denotes the Hadamard product (i.e., point-wise matrix multiplication), and matrix $\mathbf{G}$ is a 2-D Gaussian smoothing filter with standard deviations σ_{m} and σ_{k} in time- and frequency-index directions, respectively, intended to reduce noise interference. Matrix $\mathbf{T}^{11}$ contains information related to temporal (horizontal) variation in the image, $\mathbf{T}^{22}$ contains information about frequency (vertical) variation, and $\mathbf{T}^{12}$ and $\mathbf{T}^{21}$ convey both.

[3] In order to have more consistent results, the input signal is energy-normalized.

Now, each time frequency bin (k, m) has a group of four other values related to it: $T^{11}_{k,m}$, $T^{12}_{k,m}$, $T^{21}_{k,m}$, and $T^{22}_{k,m}$. Together, such bins form a structure tensor element $\mathbf{T}_{k,m}$, which is a 2×2 symmetric and positive semi-definite matrix:

$$\mathbf{T}_{k,m} = \begin{bmatrix} T^{11}_{k,m} & T^{12}_{k,m} \\ T^{21}_{k,m} & T^{22}_{k,m} \end{bmatrix}. \tag{15}$$

This matrix, whose values depend on the time-frequency bin under analysis of the given spectrogram, has interesting properties, since it carries information regarding amplitude variation in different directions. By computing its eigenvalues and eigenvectors, the direction of frequency lines near the analyzed time-frequency bin can be estimated, as well as the anisotropy measure, which indicates the degree of edginess of the given bin, as shown in the following section.

3.2. Computation of Angles and Anisotropy Measure

As mentioned, the information required to compute the angle and the anisotropy of a given time-frequency bin (k, m) is embedded in the eigenvalues and eigenvectors of the structure tensor element $\mathbf{T}_{k,m}$. Consider the eigenvalues $\lambda_{k,m}$ and $\mu_{k,m}$ of $\mathbf{T}_{k,m}$, with $\lambda_{k,m} \leq \mu_{k,m}$, and their respective eigenvectors $\mathbf{v}_{k,m}$ and $\mathbf{w}_{k,m}$. Since $\mathbf{v}_{k,m} = [v^1_{k,m}, v^2_{k,m}]^{\mathrm{T}}$ is related to the smallest eigenvalue, it is pointing in the direction of the smallest change, i.e. parallel to the direction of a frequency line near bin (k, m). Then, the angle of orientation $\theta_{k,m}$, in a horizontal perspective, is given by

$$\theta_{k,m} = \arctan\left(\frac{v^2_{k,m}}{v^1_{k,m}}\right) \in [-\pi/2, \pi/2], \tag{16}$$

with $v^1_{k,m}$ being the horizontal (temporal) component and $v^2_{k,m}$ being the vertical (frequency) component of $\mathbf{v}_{k,m}$.

The eigenvalues can also indicate the edginess of each bin (k, m) by informing how different from each other are the changes in the directions of the eigenvectors. This is called the anisotropy measure $C_{k,m} \in [0, 1]$, defined as

$$C_{k,m} = \begin{cases} \left(\dfrac{\mu_{k,m} - \lambda_{k,m}}{\mu_{k,m} + \lambda_{k,m}}\right)^2, & \mu_{k,m} + \lambda_{k,m} \geq \varepsilon \\ 0, & \text{else,} \end{cases}$$

where $\varepsilon \in \mathbb{R}^+$ is a threshold used to limit the range of what should be considered anisotropic [14], in order to increase robustness against background noise.

Bins within a more homogeneous neighborhood yield smaller values of $C_{k,m}$, while bins close to frequency lines in the spectrogram yield higher values of $C_{k,m}$. Here is where the use of the fourth-root compression is useful, since it presents much reduced dynamic range for amplitude variations at bins with small magnitude, when compared to the logarithm. With the fourth root compression, small magnitude values in the spectrogram, e.g. background noise, will lead to much smaller anisotropy values. A similar result could be obtained by applying an offset of 1 to the magnitude spectrogram before applying the logarithm.

3.3. Computation of α

Since the angles θ are related to the time-frequency bins of the given spectrogram, they live in the discrete time-frequency domain. Nevertheless, the fan-chirp transform is computed using α,

which is related to the analog time-frequency domain; therefore, a transformation must be performed in order to compute the set of α's from a set of θ's.

Let the angle ϑ be the continuous time-frequency domain version of the angle θ, and vector $\boldsymbol{\nu} = [\nu^1, \nu^2]^{\mathrm{T}}$ the continuous time-frequency domain version of vector $\mathbf{v} = [v^1, v^2]^{\mathrm{T}}$. This last conversion can be computed by $\nu^1 = v^1 h/F_{\mathrm{s}}$ and $\nu^2 = v^2 F_{\mathrm{s}}/K$, where F_{s} is the sampling rate, h is the hop-size of the STFT, and K is the number of samples used in the Fourier transform. Figure 2 depicts the geometrical relation between ϑ and α.

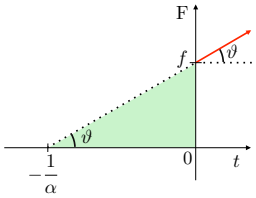


Figure 2: *Geometrical relation between the orientation angle ϑ and variable α, in the continuous time-frequency domain.*

By analyzing the triangle highlighted in green in Figure 2, one can verify that

$$\tan\vartheta = f\alpha, \tag{17}$$

which using the aforementioned conversions is written as

$$\tan\vartheta = \frac{\nu^2}{\nu^1} = \frac{v^2 F_{\mathrm{s}}/K}{v^1 h/F_{\mathrm{s}}} = \tan\theta\frac{F_{\mathrm{s}}^2}{Kh}. \tag{18}$$

Using the relation $f = kF_{\mathrm{s}}/K$,

$$\alpha_{k,m} = \tan\theta_{k,m}\frac{F_{\mathrm{s}}}{hk}. \tag{19}$$

Therefore, by performing this conversion one has a set of α parameters for each time-frequency bin (k, m) of the spectrogram.

4. COMBINATION METHOD

4.1. Principles of the Method

The structure tensor outputs, i.e. the set of angles θ and the set anisotropy measures C, comprise the information of direction and proximity to a straight edge of each time-frequency bin. Figure 3 depicts a small region of the spectrogram of an audio signal with blue arrows representing vectors pointing at direction θ, and having magnitude C. It is possible to observe that the arrows correctly follow the direction of frequency lines, and that the regions presenting only background noise, far from the frequency lines, exhibit no arrows ($C = 0$). In Figure 3, two different regions are highlighted: the arrows inside region 1 present smaller magnitude than the ones inside region 2. This occurs because the latter is surrounded by a much more linear frequency line excerpt than the former, and linear edges provide maximum difference between the eigenvalues. This effect depends on the dimensions of the smoothing filter $\mathbf{G}$ (Section 3.1): a small-dimension $\mathbf{G}$ induces smaller regions, which favor a linear model, and thus decreases the effect. Also, it is worth noting that the magnitudes vary smoothly over the whole time-frequency domain, which will assure smooth transitions between different TFRs in the combined result.

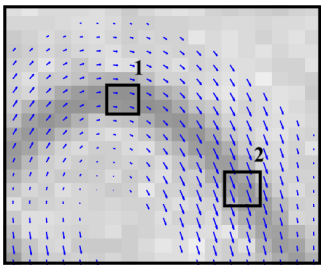


Figure 3: *Vectors in $\theta_{k,m}$ directions with magnitudes $C_{k,m}$.*

Note that the discrete fan-chirp transform models the input signal as a series of harmonically related linear frequency chirps, which means that the resulting TFR will present sparse results when the input signal matches this model within the analysis window period. As a result, using a larger analysis window allows the increase of the number of frequency chirp bins in the transform, providing minimum energy smearing only if the signal under analysis is indeed linearly varying with slope α for a longer period.

Such observations are key to the strategies used in the proposed combination procedure. The idea is to use the anisotropy measure as an indicator of the local linearity of frequency lines, and therefore an indicator of the analysis window length to be applied; and, in order to choose the best fan-chirp representation for each time-frequency bin, parameter α can then be inferred from the angle θ obtained for each bin. In the end, the method consists in performing a linear combination of time-frequency bins of the best candidates among a set of STFChTs with different α's and analysis window lengths K.

4.2. Computation of Tensor $\underline{\mathbf{X}}$

The tensor $\underline{\mathbf{X}}$ comprises TFRs which will be used in the combination. The objective is to span a broad variety of TFRs for audio signals. Three general situations can be observed in musical audio signals: (i) some sort of broadband noise produced, for instance, by blows, brushes in drums, fricative syllables in vocals, or just background noise; (ii) percussive information, as that contained in the attack of a note or a drum hit; and (iii) tonal information, possibly varying continually over time, as in the case of an instrument performing a vibrato. Figure 4 depicts the spectrogram of the onset of a harmonic pulse, zoomed in a region close to the attack. From left to right, it is possible to observe three distinct regions: background noise, the attack, and tonal information. Note that the angles computed by the structure tensor are very close to $\pi/2$ or $-\pi/2$ at the attack, indicating that the energy is distributed vertically.

Since the attacks are much better defined by transforms using short analysis windows, it is useful to define a maximum angle above which transient information should be considered predominant. This angular threshold $\vartheta^{\max}$ is then chosen in order to define two different regions: angles that represent attacks, for which STFTs with short windows will be used in the combination procedure, and angles that indicate the presence of tonal information, which will be represented by STFChTs with proper window length and parameter α. These two angular regions are indicated in Figure 5, similarly to what is done in [14].

For computation of the optimum α's distribution, an equally spaced distribution of angles ϑ is adopted, in order to minimize

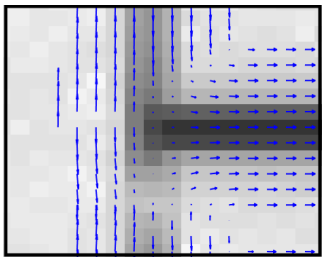


Figure 4: *Spectrogram: onset of a harmonic pulse. Vectors in $\theta_{k,m}$ directions with magnitudes $C_{k,m}$.*

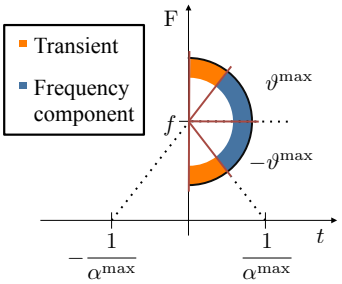


Figure 5: *Angular regions associated to transients and tonal information.*

the energy smearing in tonal regions. Consider that the angular region $[0, \vartheta^{\max}]$ will be divided in I parts. This maximum analog angle is related to an $\alpha^{\max}$ by the same relation described in Equation (17), which indicates that the analog angle ϑ is proportional to $\tan\alpha$. Since parameter α better describes the behavior of varying harmonic frequency content,[4] instead of setting a global maximum angle, it is better to consider a global $\alpha^{\max}$. This parameter can be set, for instance, considering Equation (7), since there is a range of α values that can be used given the analysis window size and the sampling frequency.

Now, angles $\vartheta_{k,m}$ that produce $\alpha_{k,m} > \alpha^{\max}$ will be considered transient information, while the others will be considered tonal information. Considering again the relation in Equation (17),

$$\tan(\vartheta^{\max}) = f\alpha^{\max}. \tag{20}$$

Considering a generic f, e.g. $f = 1$, and given $\alpha^{\max}$ and the number of α's I,

$$\vartheta^{\max} = \arctan(\alpha^{\max}), \tag{21}$$

and

$$\vartheta_i = i\frac{\vartheta^{\max}}{I} = i\frac{\arctan(\alpha^{\max})}{I}. \tag{22}$$

Finally, we can project a linear distribution of ϑ into α by computing α_i as

$$\alpha_i = \tan(\vartheta_i) = \tan(i\arctan(\alpha^{\max})/I), \tag{23}$$

and the set of α's that we shall use to compute the STFChT symmetrically spans this distribution with positive and negative values:

$$\boldsymbol{\alpha} = [-\bar{\alpha}_I, -\bar{\alpha}_{I-1}, \ldots, -\bar{\alpha}_1, \bar{\alpha}_0, \bar{\alpha}_1, \ldots, \bar{\alpha}_{I-1}, \bar{\alpha}_I]. \tag{24}$$

[4] Lower frequencies will have much smaller frequency variation than higher frequencies, for they follow a proportional relation.

For choosing the best distribution of $\mathbf{K} = [K_1, K_2, \ldots, K_J]$, since the FFT algorithm is used, having analysis window lengths of powers-of-two is desirable. This criterion is used to choose the elements of $\mathbf{K}$, optimizing this way the computational cost of this step. The parameters to be set are, then, $\mathbf{K}$ following the aforementioned criterion and the number of α values I. A set of TFRs is then composed of several instances of STFChTs using the combinations of $\mathbf{K}$ and $\boldsymbol{\alpha}$ and an STFT computed with K_1 (for the transients). Note that the sets of STFChTs also include spectrograms, since $\mathbf{X}^{\mathrm{FChT}}_{\alpha=0} = \mathbf{X}$.

Then, all the representations suffer two-dimensional linear interpolation in such a way that the highest time and frequency resolutions are preserved. All representations must have K_J frequency bins after the interpolation, and must be synchronized. In the present implementation, the same hop size is used for computing all TFRs, but this does not guarantee by itself the correct time alignment between analysis windows, reason why the time-wise interpolation (or a previous time shift in x) is also necessary. The set of parameters α and C computed via structure tensor procedure must also be interpolated, generating matrices $\mathbf{A}$ and $\mathbf{C}$, respectively. The best results are obtained when the conversion from θ to α is performed before the interpolation.

The last step is to equalize the energy of the TFRs and store them in a four-dimensional tensor $\underline{\mathbf{X}}$, with the element $X_{k,m;j,i}$ being related to the k-th frequency bin, at the m-th time frame, from a representation that has been computed with an analysis window of length K_j and a chirp rate parameter $\bar{\alpha}_i$. Since the transient information will be represented by a spectrogram computed with K_1, it is allocated at the first and at the last positions in dimension α, and therefore it will not be necessary to compute STFChTs using the first and the last values of α, i. e. $\bar{\alpha}_I$ and $-\bar{\alpha}_I$. Figure 6 depicts the tensor $\underline{\mathbf{X}}$, where groups of TFRs with different α's are illustrated clustered according to the original length of their analysis windows, K_j.

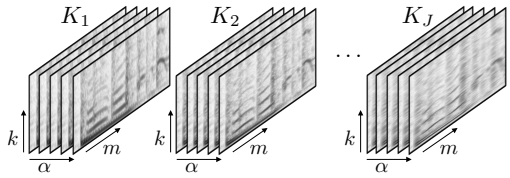


Figure 6: *Scheme of the four-dimensional TFR tensor $\underline{\mathbf{X}}$.*

4.3. Combination Procedure

The combination procedure is independently performed for each time-frequency bin (k, m), using $\alpha_{k,m}$, from $\mathbf{A}$, and $C_{k,m}$, from $\mathbf{C}$. A linear combination is performed using two different weights, one being related to $\alpha_{k,m}$, and another being related to $C_{k,m}$. The idea is to combine the representations that best suit these two parameters by using a simple linear interpolation, which can be represented as triangular complementary functions.

Figure 7 depicts an example of the weights related to the α parameters, λ^{α}, for $I = 2$. The weight λ^{α}_i is applied to the i-th layer of $\underline{\mathbf{X}}$, so the centered weight in the image, λ^{α}_0, in black, is related to the layers in $\underline{\mathbf{X}}$ which were computed with $\bar{\alpha}_0$, the others in blue, λ^{α}_1 and λ^{α}_{-1}, are related to the layers computed with $\bar{\alpha}_1$ and $\bar{\alpha}_{-1}$, and the last ones, λ^{α}_2 and λ^{α}_{-2}, in orange, are related to the STFT, which is used to represent the transients. For this reason, these last curves have a plateau in 1 for representing $\|\alpha\| \geq \bar{\alpha}_2$.

Analogously, λ^C (depicted in Figure 8) will be used for weighting the layers of $\underline{\mathbf{X}}$ along dimension j, which is related to K.

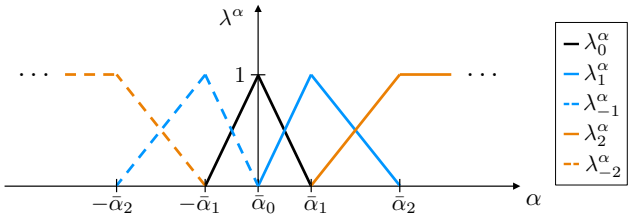


Figure 7: *Example of the weights used for combining TFRs with different α's ($I = 2$).*

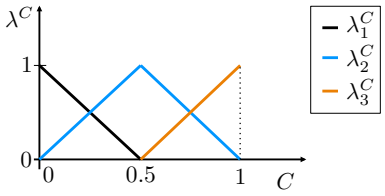


Figure 8: *Example of the weights used for combining TFRs with different K's ($J = 3$).*

The combined TFR $\mathbf{X}^{\text{Comb}}$ is then described by the following two-dimensional interpolation for each time-frequency bin (k, m):

$$X^{\text{Comb}}_{k,m} = \sum_{j=1}^{J} \sum_{i=-I}^{I} \lambda^C_{k,m;j} \lambda^\alpha_{k,m;i} X_{k,m;j,i}. \tag{25}$$

4.4. Practical Considerations

In practice, CPU processing can be saved by using only $\alpha = 0$ for small window sizes, e.g. $K = 1024$ (≈ 21 ms), achieving very similar results. Another practical consideration is regarding the storage of $\underline{\mathbf{X}}$ in memory. Since several TFRs may be stored, it is useful to process the combined TFRs in small excerpts of x, and then concatenate the results, giving a certain time margin to guarantee the proper computation of all TFRs and the structure tensor parameters. Once the combined TFR is processed for that given excerpt, its tensor $\underline{\mathbf{X}}$ is no longer needed, and therefore the memory space can be freed up. Also, in order to reduce backwards smearing of attacks, asymmetrical analysis windows having a longer tail on the left side can be used.

5. EXPERIMENTS

Experiments were conducted in order to assess the performance of the proposed method, using both synthetic and real-world audio signals. All input signals had sampling rate $F_s = 48000$ Hz. The system was set according to the following configuration. In the structure tensor procedure, the analysis windows of the spectrogram had length $K = 1024$ (21.3 ms); in the smoothing two-dimensional filter G, σ_k corresponded to 90 Hz and σ_m to 15 ms; and the threshold used in the anisotropy measure computation was $\varepsilon = 1$. The analysis window sizes for the STFChTs were chosen as $\mathbf{K} = [1024, 2048, 4096]$ (21.3, 42.6 and 85.3 ms); in order to reduce backwards energy smearing, asymmetric[5] analysis windows were used for the computation of STFChTs with K_2 and K_3; $\alpha^{\max} = 23.4$ resulted from the application of Equation (7); and all TFRs were computed with hop size $h = 256$ samples.

5.1. Proof of Concept

As a proof of concept, synthetic signals were selected to asses the method's performance in specific challenging scenarios with regards to time-frequency representations.

First, a pulse comprised of harmonically related sinusoids, with onset at 0.1 s and offset at 0.5 s, contaminated by additive white Gaussian noise (SNR = 50 dB), was used. Figures 9(a) and (b) depict the spectrograms obtained for this signal, using $K_1 = 1024$ and $K_3 = 4096$, respectively the shortest and longest window sizes; and Figures 9(c) and (d) depict the resulting TFRs using the proposed combination procedure, with $I = 1$ and $I = 5$ respectively. Red dashed lines indicate the onset and offset instants to facilitate the visualization. As can be clearly observed, the two TFRs computed with the proposed method yielded nearly identical results, combining the time precision provided by the first spectrogram with the frequency resolution of the second one. Since the frequency lines present in this signal are well represented by an STFChT with $\alpha = 0$ (i.e. a spectrogram), increasing the number of STFChTs available does not affect the result. This could be the case of representing signals of instruments with stable f_0, e.g. piano or harp.

The second example uses a harmonic series whose f_0 varies in a sinusoidal fashion with increasing amplitude, also contaminated by additive white Gaussian noise (SNR = 50 dB). This signal allows one to verify the capability of handling a wide variety of α's. The results are depicted in Figure 10, where it is possible to see the original spectrogram, and three resulting TFRs, computed with $I = 1$, $I = 3$ and $I = 5$. As expected, increasing I also increases the time-frequency resolution, yielding more concentrated and consistent frequency lines. For instance, the results obtained for $I = 3$ and $I = 5$ differ only in the steeper slopes, mainly on the right side of the pictures.

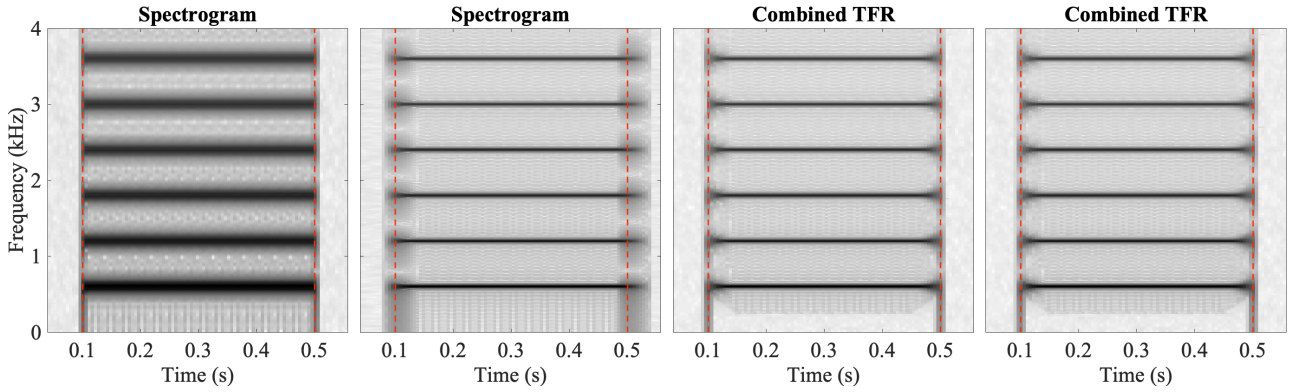
Finally, the last synthetic signal is a sum of two harmonic signals having different sinusoidal variations of f_0, with additive white Gaussian noise (SNR = 50 dB). Figure 11 depicts the spectrogram used for the computation of the structure tensor and the combined TFR, for which $I = 5$ was used. The resulting TFR represents the input signal with a much higher definition, and very smooth transitions can be observed. It is worth highlighting that even at places where more than one frequency line crosses the same bin, the signal is fairly well represented.

5.2. Real-World Signals

The experiments with real-world signals used the MedleyDB [16] dataset. From each track, an excerpt of 10 s containing part of the main melody was selected, so that each song contributed the same amount of data. These signals were divided into 1-s segments, totalling 1210 excerpts. In order to assess the method's performance, the Gini index[6] [17] was chosen as an objective figure-of-merit.

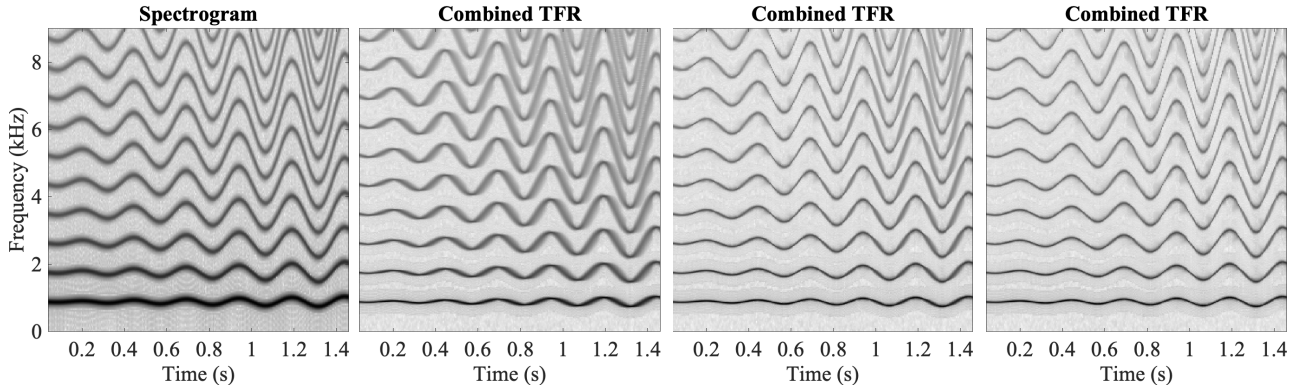
[5] The asymmetric windows are computed by concatenation of the first half of a Hanning window computed with K samples, and the second half of a Hanning window computed with $K/2$ samples.

[6] The Gini index is a measure of sparsity that indicates within the range $[0, 1]$ how concentrated is the energy in a given set of bins.



(a) $K = 1024$. (b) $K = 4096$. (c) $I = 1$. (d) $I = 5$.

Figure 9: *Spectrograms for different K values and combined TFRs with different I values computed for a pulse composed of harmonically related sinusoids. Onset and offset are indicated by the red dashed lines.*



(a) $K = 1024$. (b) $I = 1$. (c) $I = 3$. (d) $I = 5$.

Figure 10: *Varying vibrato: spectrogram and the combined TFRs with different I values.*

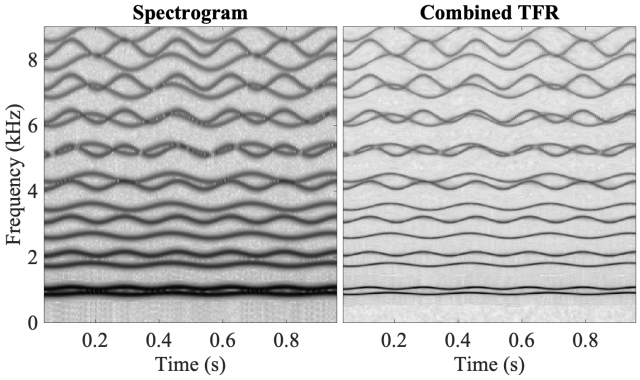


Figure 11: *Simultaneous vibratos: spectrogram ($K = 1024$) and combined TFR ($I = 5$).*

Each audio excerpt was then processed using the proposed method and standard spectrograms (using $K = 1024, K = 2048$ and $K = 4096$), and the Gini index was computed for each resulting TFR.

Figure 12 depicts the percentage of times each representation was ranked in each position according to the Gini index. The combined TFR is by far the most effective in terms of sparsity, ranking first in about 80% of the time. The STFT-1024 accounts for 50% of the second position, the STFT-2048 for 65% of the third, and the STFT-4096 for 50% of the fourth.

Finally, an excerpt from a piano and vocal recording was selected to illustrate how the TFR of a real-world audio signal can be improved by the proposed strategy. Figure 13 depicts its original spectrogram next to its combined TFR. It is possible to verify that in the spectrogram the piano is barely noticeable, while the TFR generated by the proposed method clearly represents both the piano and the singing vocal—which is performing a very fast melisma.

6. CONCLUSIONS

This paper presented a method for producing sparse TFRs by combining different STFChTs using the information provided by the structure tensor. We extract directional information from a spectrogram of the input signal, which guides an interpolation procedure. Experiments comprising synthetic and real recorded audio

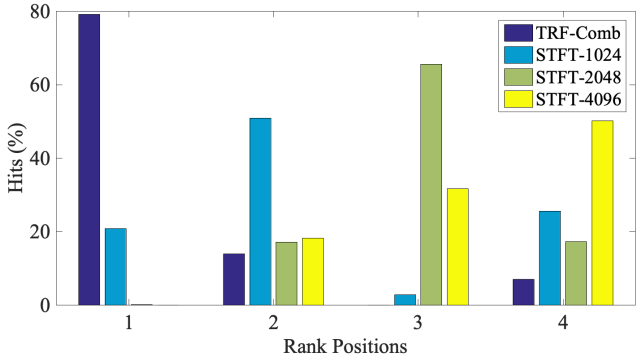


Figure 12: *Rank of different representations in terms of Gini Index for the MedleyDB.*

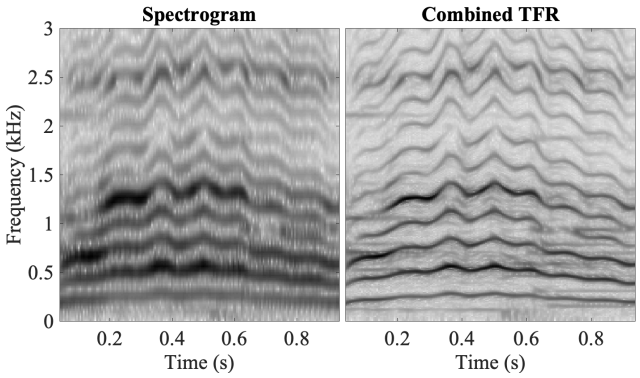


Figure 13: *Vocal and piano: spectrogram ($K = 1024$) and combined TFR ($I = 5$).*

signals suggest that the proposed method provides high-definition TFRs, improving the concentration of frequency lines with various slopes and the definition of transient information, when compared to standard TFRs, e.g. spectrograms. Given the method's ability to provide refined inputs for MIR tasks, such as main melody and multi-pitch extraction, the natural continuation of this research is the reformulation of state-of-the-art methods in MIR to take advantage of such representations in real application scenarios.

7. REFERENCES

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

[2] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri, "Automatic music transcription: Challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, December 2013.

[3] B. S. Gowrishankar and N. U. Bhajantri, "An exhaustive review of automatic music transcription techniques: Survey of music transcription techniques," in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, Paralakhemundi, India, June 2016, pp. 140–152.

[4] Maurício do V. M. da Costa and Luiz W. P. Biscainho, "Combining time-frequency representations for music information retrieval," in *15º Congresso de Engenharia de Áudio da AES-Brasil*, Florianópolis, Brazil, October 2017, pp. 12–18, AES.

[5] Rongping Lin, Chunhui Du, Shan Luo, and Qi Xu, "Performance on a combined representation for time-frequency analysis," in *2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, China, June 2017, pp. 858–862, IEEE.

[6] C. S. Detka, P. Loughlin, and A. El-Jaroudi, "On combining evolutionary spectral estimates," in *IEEE 7th Signal Processing Workshop on Statistical Signal and Array Processing*, Quebec, Canada, June 1994, pp. 243–246, IEEE.

[7] Patrick Loughlin, James Pitton, and Blake Hannaford, "Approximating time-frequency density functions via optimal combinations of spectrograms," *IEEE Signal Processing Letters*, vol. 1, no. 12, pp. 199–202, December 1994.

[8] Alexey Lukin and Jeremy Todd, "Adaptive time-frequency resolution for analysis and processing of audio," in *Audio Engineering Society 120th Convention*, Paris, France, May 2006, AES, Preprint 6717.

[9] François Auger and Patrick Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, May 1995.

[10] 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.

[11] Pablo Cancela, Ernesto López, and Martín Rocamora, "Fan chirp transformation for music representation," in *13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010, pp. 1–8.

[12] Josef Bigun and Gösta H. Granlund, "Optimal orientation detection of linear symmetry," in *IEEE First International Conference on Computer Vision*, London, UK, June 1987, pp. 433–438.

[13] Hans Knutsson, "Representing local structure using tensors," in *6th Scandinavian Conference on Image Analysis (SCIA)*, Oulu, Finland, June 1989, pp. 244–251.

[14] R. Füg, A. Niedermeier, J. Driedger, S. Disch, and M. Müller, "Harmonic-percussive-residual sound separation using the structure tensor on spectrograms," in *2016 IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 2016, pp. 445–449.

[15] I. F. Apolinário, Maurício do V. M. da Costa, and L. W. P. Biscainho, "Structure tensor applied to parameter estimation in the fan-chirp transform," in *2nd AES Latin American Congress of Audio Engineering*, Montevideo, Uruguay, September 2018.

[16] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive mir research," in *15th International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, October 2014.

[17] Niall Hurley and Scott Rickard, "Comparing measures of sparsity," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, October 2009.