
# NOVEL TIME-FREQUENCY REPRESENTATIONS FOR MUSIC INFORMATION RETRIEVAL

Maurício do Vale Madeira da Costa

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Luiz Wagner Pereira Biscainho

Rio de Janeiro
Abril de 2020

NOVEL TIME-FREQUENCY REPRESENTATIONS FOR MUSIC INFORMATION RETRIEVAL

Maurício do Vale Madeira da Costa

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Orientador: Luiz Wagner Pereira Biscainho

Aprovada por: Prof. Luiz Wagner Pereira Biscainho
Prof. Diego Barreto Haddad
Prof. Marcello Luiz Rodrigues de Campos
Prof. Markus Vinicius Santos Lima
Prof. Martín Rocamora

RIO DE JANEIRO, RJ – BRASIL
ABRIL DE 2020

da Costa, Maurício do Vale Madeira

Novel Time-Frequency Representations for Music Information Retrieval/Maurício do Vale Madeira da Costa. – Rio de Janeiro: UFRJ/COPPE, 2020.

XVI, 121 p.: il.; 29, 7cm.

Orientador: Luiz Wagner Pereira Biscainho

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2020.

Referências Bibliográficas: p. 107 – 121.

1. Digital Signal Processing. 2. Time-frequency Representation. 3. Time-frequency Analysis. 4. Fan-chirp Transform. 5. MIR. 6. Dominant Melody. 7. Main Melody. 8. Brazilian music. 9. Samba. I. Biscainho, Luiz Wagner Pereira. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

*To my family.*

# Agradecimentos

Como forma de melhor expressar minha gratidão, tomo a liberdade de escrever os agradecimentos em português, minha língua nativa, uma vez que assim também alcançarei de forma mais efetiva muitos daqueles a quem desejo aqui me dirigir. Em primeiro lugar, sou grato a Deus por todas as oportunidades de progresso que me tem permitido experienciar. Agradeço à minha família, cujo apoio incondicional me permitiu seguir estudando e investindo em minha formação acadêmica, e sem o qual penso que isso não seria realizado. Agradeço também à Maitê, minha parceira de caminhada, que também se tornou parte importante da família, e cujos amor, apoio e companhia também foram essenciais durante essa trajetória. Agradeço ao amigo Luiz Wagner, que me orientou durante quase 11 anos, pela dedicação e atenção inigualáveis que me prestou, além da companhia sempre divertida e musical. Agradeço aos colegas, funcionários e professores do laboratório SMT, e em especial aos amigos mais próximos: Lucas Simões, amigo e companheiro de experiências internacionais, Lucas Arrabal, Vitor, Victor, Allan, Isabela, cuja parceria nos rendeu colaborações em nossos trabalhos, Luís Felipe, Wallace, Luiz Gustavo, Igor, Pedro e Iandra. Agradeço também aos amigos e colegas que me fizeram companhia durante minha estadia na universidade Télécom Paris, em particular ao Lucas Simões, Magdalena e Robert, cuja amizade e companhia tornaram nossos dias em Paris muito melhores. Agradeço aos membros das bancas examinadoras do meu exame de qualificação e dessa tese, que aceitaram avaliar o trabalho e também contribuir para que fosse o melhor possível. Por fim, mas não menos importante, agradeço às agências de fomento CNPq e CAPES, pela concessão das bolsas de estudo para a realização dessa pesquisa. Sabendo que não é possível escrever os agradecimentos de qualquer trabalho sem cometer alguma injustiça por não citar nominalmente alguém, peço desculpas aos tantos outros demais amigos, colegas ou profissionais cujo concurso contribuiu de forma mais ou menos importante para o desenvolvimento deste trabalho.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

# NOVAS REPRESENTAÇÕES TEMPO-FREQUENCIAIS PARA EXTRAÇÃO DE INFORMAÇÃO MUSICAL

Maurício do Vale Madeira da Costa

Abril/2020

Orientador: Luiz Wagner Pereira Biscainho

Programa: Engenharia Elétrica

Representações tempo-frequenciais (RTFs) são uma das ferramentas mais valiosas em processamento digital de áudio, sendo utilizadas em diversas aplicações. RTFs podem ser calculadas tendo diferentes resoluções em tempo e em frequência e podem, inclusive, representar determinadas variações em frequência, como no caso do uso da transformada de *fan-chirp*. A maior deficiência de RTFs é o espalhamento de energia relacionado à não-estacionariedade do sinal na região da janela de análise. Esse tipo de artefato geralmente resulta em prejuízo de desempenho da aplicação que utilize tal RTF; portanto, ter RTFs que representem precisamente os sinais de interesse é essencial para melhorar o desempenho de tais sistemas.

Uma forma de se calcular RTFs de alta resolução é combinar RTFs de diferentes resoluções de forma a preservar os melhores aspectos de cada uma. Essa é a ideia geral que embasa todos os métodos propostos nessa tese, da qual o principal objetivo é possibilitar a representação precisa de sinais de melodia principal em contextos polifônicos. Os métodos são classificados como: combinações ponto-a-ponto, combinações baseadas em informação local, e combinações baseadas em análise de imagem. Seus desempenhos são medidos por meio de diversos experimentos, em que são utilizados sinais sintéticos controlados e sinais reais, e os resultados apontam o método proposto de interpolação de *fan-chirps* em multirresolução como o melhor em termos de largura de banda de frequência, definição de *onset* e faixa dinâmica.

Ademais, um método para anotação automática foi desenvolvido com a finalidade de facilitar o trabalho de transcrição de padrões rítmicos. Esse método utiliza RTFs com baixa resolução frequencial e um procedimento de agrupamento para classificar os tipos de toque. Estima-se uma acurácia de cerca de 75% a 80% em termos de classificação inicial.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

NOVEL TIME-FREQUENCY REPRESENTATIONS FOR MUSIC INFORMATION RETRIEVAL

Maurício do Vale Madeira da Costa

April/2020

Advisor: Luiz Wagner Pereira Biscainho

Department: Electrical Engineering

Time-frequency representations (TFR) are one of the most valuable tools in digital audio processing, being used in many applications. TFRs can be computed having different time or frequency resolutions and can even represent a certain frequency variation over time, e.g. when using the fan-chirp transform. The main shortcoming of TFRs is the energy smearing related to non-stationarity of the signal within the analysis windows used. This kind of artifact usually results in performance degradation of applications that make use of TFRs, hence providing TFRs that precisely represent the signals of interest is crucial to enhance the performance of such systems.

A way to compute a high-resolution TFR is to combine TFRs having different resolutions in such a way that preserves the best aspects of each representation. This is the general idea behind all methods proposed in this thesis, of which the main goal is to allow for a sharp representation of main melody signals in polyphonic contexts. The methods are classified as: bin-wise combinations, combinations based on local information, and methods based on image analysis. Their performance are assessed by means of several experiments using both synthetic and real-world signals, and the results indicate the proposed multi-resolution fan-chirp interpolation method as the best in terms of frequency bandwidth, onset definition and dynamic range.

Also, an automatic annotation scheme was devised to diminish the human effort in the transcription of rhythm patterns. This method utilizes TFRs with coarse frequency resolution and a clustering procedure to classify the types of hit. The estimated accuracy in terms of classification is around 75% to 80%.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Presentation

## 1.1 Introduction

Music is one of the most universal components in culture, and it has been produced since the earliest ages of the mankind, being one of the first forms of expression and communication and playing an important role in rituals, celebrations and ceremonies. Since the invention of the phonograph in 1877 by Thomas Edison [1], with which for the first time musical performances could be registered and reproduced, the recording technology has evolved, and many different types of media have been used for this purpose. In the last decades, with the advent of digital audio, music has became more and more accessible, especially nowadays with the low cost associated with digital devices and music streaming services.

The digital revolution has contributed for a democratization of the access to music consumption and production by considerably decreasing the cost associated with players, storage devices, recording equipment and software for music production. Furthermore, not only the cost of equivalent tools that existed before has dropped in the digital era, but also the development of new techniques increased the power for analyzing audio signals [2–27].

Digital signal processing has brought a whole new universe of possibilities for manipulating and analyzing signals, being one of the most important set of tools and techniques in the information era, along with machine learning. Tasks such as automatic music transcription [28–47] and sound source separation [11, 46, 48–57] can now be performed by digital processors and provide new ways of human interaction with music.

In this context, being able to identify the energy concentration in frequency over time is crucial. The mathematical tools that allow signals to be represented in both time and frequency are the so-called time-frequency representations (TFRs) [58]. In audio signal processing in general, several state-of-the-art techniques depend directly

on the TFR of the signal under analysis [29–34, 36, 37, 40, 41, 59–67], usually being the only source of information available. Also, in many applications in the area of music information retrieval (MIR), TFRs are often used as input for their calculations. Therefore, providing a sparse[1] TFR, in which the different components are well defined and as separated as possible from each other, may be a good way to improve the performance of such systems.

In practice, when using common TFRs, e.g. the short-time Fourier transform (STFT), one can improve frequency resolution by using larger analysis windows, to the detriment of time resolution, and also the other way around, but not both simultaneously—as dictated by the uncertainty principle [58]; one must cope with this compromise by choosing an optimum size for the analysis window, which will depend on the application. For instance, in the study of microtiming[2] [62, 68, 69], the identification of onsets (e.g. when a drum is hit or a guitar string is plucked) requires fine time resolution, thus calling for short analysis windows. On the other hand, for pitch related tasks [29, 30, 34–36, 43, 70, 71], the requirement is a frequency resolution that allows one to discriminate between different musical notes, hence the need for long analysis windows. Moreover, the analysis of expression includes the careful dynamic following of pitch, which can vary continually with time, e.g. when a musician plays or sings using vibrato or glissando. Such fast pitch variations call for a sufficiently fine resolution of frequency over time.

Although, in essence, the statement of the uncertainty principle is unbreakable, depending on the nature of the signal to be analyzed and the mathematical properties that one may not be interested in preserving, some techniques can enhance the overall time-frequency resolution [58, 72–91]. Each method has its own limitations and is better suited for signals with certain characteristics.

## 1.2 Motivation and Scope of this Thesis

In many genres of music, the main melody represents a fundamental element of the songs, usually being the most important and memorable part of it. For this reason, dominant melody analysis is one of the most important tasks in the area of MIR, and has actively been studied by its research community [28, 40–47, 49–53, 55–57, 66, 71, 92–99] for many years. The majority of the state-of-the-art methods in MIR make use of machine learning techniques, such as neural networks [14, 31, 40, 41, 43, 49, 50, 53, 53, 56, 59, 62, 92, 93, 100–104], and of TFRs of the audio signal as input. Therefore, this and many other applications in the context of MIR may

[1] Here, the term 'sparsity' is used in the sense of energy concentration in a small number of coefficients.

[2] Microtiming can be considered the small deviations from regular time a performer applies when playing a musical piece, such as what is identified as swing in jazz.

profit from TFRs with higher resolution. Especially in the case of melody analysis, TFRs which are able to represent the typical dynamic frequency variation observed in such signals seem to be a valuable target.

Although the area of TFRs has already a solid and well stablished mathematical foundation [58, 81], new methods for generating high-resolution representations are still being developed [72–91, 105]. This thesis aims at studying and designing methods for this purpose, especially for representing polyphonic signals containing fast frequency variations. The main idea that pervades all the approaches present in this work is to combine different TFRs so that the final TFR gathers the best characteristics of each one. Since the main goal is to provide representations for analysis purposes, this work is limited to generating only the magnitude spectrograms, without phase information or reconstruction to time domain, which can be addressed in future work.

Also in the context of MIR, this thesis develops some work over the Brazilian Rhythmic Instruments Dataset (BRID) [106], a dataset comprised of solo- and multiple-instrument recordings, spanning 10 different percussion instrument classes and 5 different Brazilian traditional rhythm classes. This dataset is copyright-free, and available for research within the MIR community, helping to fill a gap concerning this genre in the literature. Annotations of beat and down-beat have been made, and the ones related to type of articulation are in progress. An automatic scheme is proposed for helping on this last annotation procedure. This dataset will allow one to perform various rhythmic analyses, e.g. microtiming and pattern recognition [106].

Part of this thesis is related to the StaReL[3] project, which have counted with collaboration of researchers from the Universidade Federal do Rio de Janeiro (UFRJ), in Brazil, the universities Télécom Paris and CentraleSupélec, in France, and the Universidad de la República, in Uruguay. The Brazilian researchers were financed by the federal Brazilian agency CAPES. In this project, tools for the computational analysis of rhythm and expressiveness in Afro-rooted Latin-American music (samba and candombe) were developed, along with the production of datasets and some of the methods proposed in this thesis. The research related to this project took place at the Télécom Paris university, from January to September of 2019.

This thesis is aligned with open and reproducible research guidelines, so all the ideas, models and codes produced within the context of this research will be shared with the community under open licenses whenever possible.

[3]For details, visit `www.smt.ufrj.br/~starel`.

## 1.3 Thesis Overview

The text is divided into two parts: Part I, which presents a background and the proposed methods for combining time-frequency representations; and Part II, which contains experiments, applications and tools for music information retrieval. Chapter 2, which opens Part I, comprises an introduction to time-frequency representations, where some of the principal representations used for musical signal analysis are presented; Chapter 3 presents bin-wise combinations of spectrograms; Chapter 4 presents methods for combination of spectrograms based on local information; and Chapter 5 presents combinations based on image analysis. In Part II, Chapter 6 presents experiments focused on main melody analysis, performed with some of the methods studied; Chapter 7 presents a method for automatic onset classification, along with the BRID dataset; and in Chapter 8 the document ends with the main conclusions of this work, along with directions for future research.

## 1.4 Publications Summary

In Chapter 3, it is proposed a novel method for bin-wise combination of spectrograms, namely, the sample weighted geometric mean (SWGM), which performs a weighted geometric mean whose weights are given by a function of the samples to be combined themselves.

- DA COSTA, M. V. M., BISCAINHO, L. W. P. "Combining Time-Frequency Representations for Music Information Retrieval". In: *Anais do 15o Congresso de Engenharia de Áudio da AES-Brasil*, Florianópolis, Brazil, October 2017.

In Chapter 4, two combination methods based on local information that use the Gini index as a measure of sparsity are proposed: the local sparsity (LS) and the smoothed local sparsity (SLS) methods.

- DA COSTA, M. V. M., BISCAINHO, L. W. P. "Combining Time-Frequency Representations via Local Sparsity Criterion". In: *Proceedings of the 2nd AES Latin American Congress of Audio Engineering*, Montevideo, Uruguay, September 2018.

- DA COSTA, M. V. M., APOLINÁRIO, I. F., BISCAINHO, L. W. P. "Sparse Time-Frequency Representations for Polyphonic Audio Based on Combined Efficient Fan-Chirp Transforms", *Journal of the Audio Engineering Society*, v. 67, n. 11, pp. 894–905, November 2019.

In Chapter 5, it is presented a new strategy for estimating and combining different instances of the fan-chirp transform, whose frequency slope parameters are

computed via a fast implementation based on the structure tensor: the frame-based method for estimation of main directions (FEMD).

- APOLINÁRIO, I. F., DA COSTA, M. V. M., BISCAINHO, L. W. P. "Structure Tensor Applied to Parameter Estimation in the Fan-Chirp Transform". In: *Proceedings of the 2nd AES Latin American Congress of Audio Engineering*, Montevideo, Uruguay, September 2018.

- DA COSTA, M. V. M., APOLINÁRIO, I. F., BISCAINHO, L. W. P. "Sparse Time-Frequency Representations for Polyphonic Audio Based on Combined Efficient Fan-Chirp Transforms", *Journal of the Audio Engineering Society*, v. 67, n. 11, pp. 894–905, November 2019.

Also in Chapter 5, the structure tensor is used in another novel method that combines samples of a multi-resolution dictionary of fan-chirp representations: the multi-resolution fan-chirp interpolation method (MRFCI).

- DA COSTA, M. V. M., BISCAINHO, L. W. P. "High-Definition Time- Frequency Representation Based on Adaptive Combination of Fan-Chirp Transforms via Structure Tensor". In: *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx)*, Birmingham, United Kingdom, September 2019.

In Chapter 7, it is described a novel dataset comprising a wide range of recordings and annotations of Brazilian traditional rhythmic instruments.

- MAIA, L. S., DE TOMAZ JÚNIOR, P. D., FUENTES, M., et al. "A Novel Dataset of Brazilian Rhythmic Instruments and Some Experiments in Computational Rhythm Analysis". In: *Proceedings of the 2nd AES Latin American Congress of Audio Engineering*, Montevideo, Uruguay, September 2018.

# Part I

# Time-Frequency Representations

# Chapter 2

# Introduction to Time-Frequency Representations

When dealing with audio signals, knowing its frequency content can be very useful. The most common tools used to this end are transforms which convert signals in time domain into signals in frequency domain. If the audio signal under scrutiny is stationary during the period in which it is being observed, time to frequency transforms, e.g. the Fourier transform, can provide sparse results, i.e. a signal in the frequency domain containing very pronounced peaks. Nevertheless, considering audio signals to be perfectly stationary is unreasonable, since their frequency content varies in time. Therefore, in order to focus the analysis around a given time $\tau$, during which the signal could be approximately stationary, it is necessary to use a time windowing function. Such a function, e.g. Gaussian or Hamming [107], is used to emphasize parts of this signal near $\tau$, while suppressing the remaining (unwanted) parts of it. Shifting such a function in time and multiplying it by the signal is then what allows one to analyze the evolution of the spectrum of the given signal over time. This way, the corresponding spectrum (i.e. the signal in the frequency domain) will inform the frequency content of the signal around that specific instant. By performing this procedure, one represents the signal in a time-frequency domain.

Time-frequency analysis methods are among the most important tools in the audio signal processing area, being a mature field with a solid mathematical foundation [58]. Due to the great utility of TFRs in different contexts, a wide variety of TFRs and methods for enhancing such representations have been developed over the years. In this chapter, some of the principal time-frequency representations used for musical signal analysis will be briefly studied.

## 2.1 The Spectrogram

The most used time-to-frequency transform is the Fourier transform, being adopted in many procedures [29, 30, 32–34, 36, 37, 59–62, 64, 65] due to its straightforward interpretation and low computational cost. The Fourier transform [107] of a real-valued signal $x(t)$, denoted as $\mathcal{F}(x(t)) = X(f)$, is given by its projection on a basis comprised of complex exponentials with different frequencies, and can be defined as

$$X(f) \triangleq \int_{-\infty}^{\infty} x(t)\mathrm{e}^{-j2\pi ft}\mathrm{d}t, \tag{2.1}$$

where $t$ denotes time, $f$ denotes frequency and $j^2 = -1$.

A real-valued analysis window $w$, used for analyzing a specific excerpt of the signal, can be continuously shifted in time, providing a spectrum for each instant $\tau$ of the windowed version of $x(t)$. This way, the spectrogram, or energy density spectrum, $X(\tau, f)$ is defined as the squared[1] absolute value of the short-time Fourier transform (STFT):

$$X(\tau, f) \triangleq \left\| \int_{-\infty}^{\infty} x(t)w(t-\tau)\mathrm{e}^{-j2\pi ft}\mathrm{d}t \right\|^2, \tag{2.2}$$

where $\|\cdot\|$ denotes the magnitude of its complex argument.

In the discrete time domain $n$, the Fourier transform (DTFT) of a given signal $x$ can be computed as

$$X(e^{j\Omega}) \triangleq \sum_{n=-\infty}^{\infty} x_n \mathrm{e}^{-j\Omega n}, \tag{2.3}$$

where $X(e^{j\Omega})$ is continuous and periodical, with period $2\pi$.

The spectrum of a given signal $x$ with finite duration of $N$ samples is continuous and it can be completely summarized by $N$ equally spaced samples, computed by the so called discrete Fourier transform (DFT) as

$$X_k \triangleq \sum_{n=0}^{N-1} x_n \mathrm{e}^{-j\frac{2\pi}{N}kn}, \tag{2.4}$$

where $k$ is the frequency index. In practice, more efficient algorithms, namely, Fast Fourier Transform (FFT) algorithms [107], are used to compute the DFT.

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 in $N$ samples, the discrete spectrogram

[1] The spectrogram can also be defined simply as the magnitude of the STFT.

can be described by

$$X_{k,m} \triangleq \left\| \sum_{n=0}^{N-1} x_{n-hm} w_n \mathrm{e}^{-j\frac{2\pi}{N}kn} \right\|^2, \tag{2.5}$$

where $k \in \mathcal{K} \triangleq \{0, 1, 2, ..., K-1\}$ 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 $N$ samples used for computation of the spectrogram, and $h \in \mathbb{N}$ is the analysis hop size in samples. The time-frequency bins can then be represented in a matrix with the following form:

$$\mathbf{X} = \begin{bmatrix} X_{K-1,1} & X_{K-1,2} & \dots & X_{K-1,M-1} & X_{K-1,M} \\ X_{K-2,1} & X_{K-2,2} & \dots & X_{K-2,M-1} & X_{K-2,M} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{1,1} & X_{1,2} & \dots & X_{1,M-1} & X_{1,M} \\ X_{0,1} & X_{0,2} & \dots & X_{0,M-1} & X_{0,M} \end{bmatrix}. \tag{2.6}$$

This matrix can then be interpreted as an image having a vertical frequency axis and a horizontal time axis. Note that this matrix has the vertical axis inverted in comparison with the line indexes, so that it will display the frequency axis in ascending order from bottom to top, providing a more intuitive visualization. Also, since audio signals are real, the second half of the magnitude spectrum will be just a mirrored version of the first half; hence, assuming that $N$ is even, all information can be fully represented by the first half of the spectrum, i.e. $K = 1 + N/2$.

Although the spectrogram is not a true joint distribution of time and frequency, since it does not satisfy the fundamental requirements of marginals due to the windowing process [58], i.e.

$$\sum_k X_{k,m} \neq \|x_m\|^2 \tag{2.7}$$

$$\sum_m X_{k,m} \neq \|X_k\|^2, \tag{2.8}$$

it is a time-frequency distribution that gives the energy density spectrum over discrete time. It has the limitation dictated by the uncertainty principle [58]: a signal cannot be represented with arbitrarily high time and frequency resolutions simultaneously. As the length of the analysis window gets longer, a greater frequency resolution can be achieved, as longer excerpts of the signal are projected into the complex exponentials. On the other hand, it provides a poorer time resolution, for the same reason. As a consequence, non-stationary parts of the signal, such as attack of notes or fast frequency variations, become blurred on the representation, for they are integrated with their neighborhood. Choosing a generally good analysis window length is then a typical compromise in all tasks in which the spectrogram

is used.

As an example of this principle, two different spectrograms of a piano recording, one using an analysis window of 21.3 ms and another using a window of 85.3 ms, are depicted in Figures 2.1, where the first six harmonics of a note are highlighted, and 2.2[2], respectively. Note that the harmonics from a single note are represented in a regular interval in frequency, due to the linear scale of the frequency axis. As expected, the frequency lines, which are disposed horizontally, are much better defined in the spectrogram using the longer analysis window, while the onsets, i.e. the instant of the beginning of notes, are better defined in the spectrogram using the shorter analysis window.

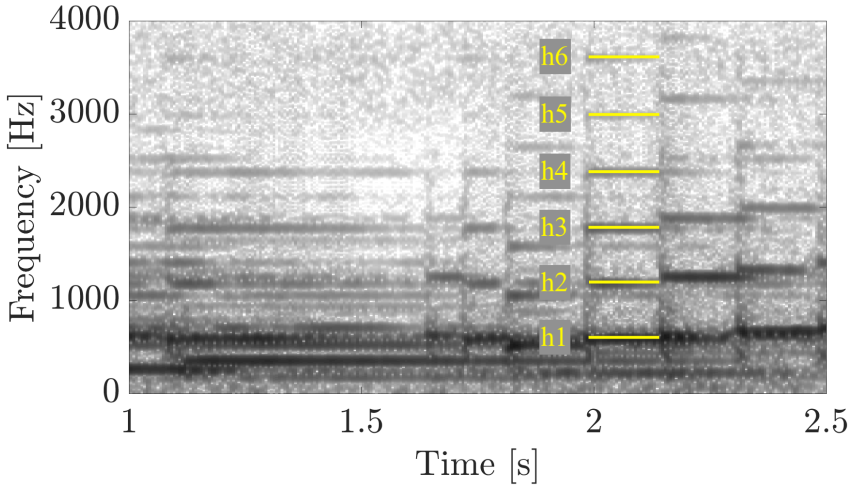



Figure 2.1: Spectrogram (21.3 ms) of a piano signal.

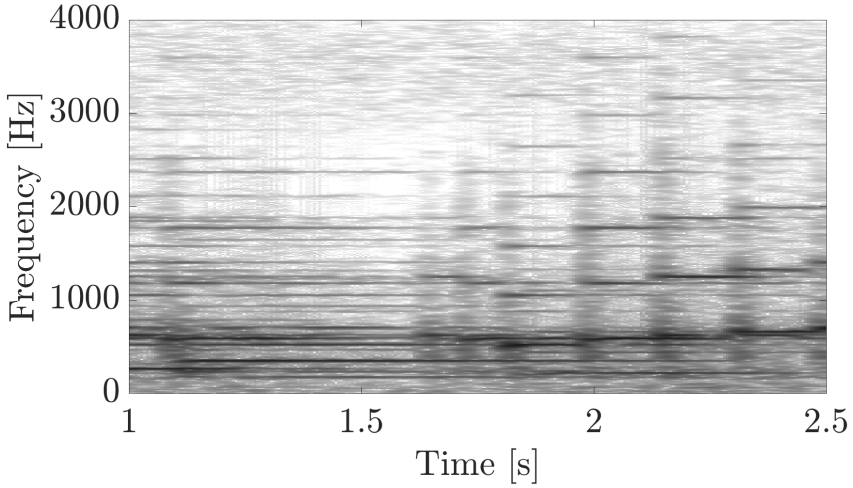



Figure 2.2: Spectrogram (85.3 ms) of a piano signal.

It is worth highlighting that the spectrogram has a linear distribution in both

[2]All spectrograms presented in this thesis are depicted in log-magnitude, with a dynamic range of 80 dB.

time and frequency axes. Therefore, for analyzing musical signals, in which the notes typically have a geometric frequency distribution,[3] this TFR exhibits too high resolution for high-frequency components and too low resolution for low-frequency components.

## 2.2 The Fan-Chirp Transform

In order to have a good (sparse) representation of a signal in a spectrogram, its windowed excerpts must be well represented by the basis in which they are projected, which is a set of complex exponentials in the case of the STFT-based spectrogram. This means that this signal must be close to stationary during that period. However, this is often a poor model for many acoustic sources whose frequency content changes very dynamically, e.g. singing voice or instruments performing vibratos or glissandos.

The fan-chirp transform (FChT) [70, 72, 73, 77, 80, 108–110] addresses this issue by 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, the representation of fast frequency variations is hopefully far superior than what the spectrogram is capable of conveying. The main idea is to use a basis composed of complex exponentials whose frequency varies linearly in time (linear chirps). When the correct frequency slope is used for the harmonic signal under analysis, all of its frequency content remains well defined, thus overcoming the energy spreading observed in a signal with fast frequency variations represented by a spectrogram.

The fan-chirp transform $X^{\mathrm{FChT}}(f, \alpha)$ of a given signal $x(t)$, in the continuous time domain, is defined in [109] as

$$X^{\mathrm{FChT}}(f, \alpha) \triangleq \int_{-\infty}^{\infty} x(t)\phi'_{\alpha}(t)\mathrm{e}^{-\mathrm{j}2\pi f \phi_{\alpha}(t)}\mathrm{d}t, \tag{2.9}$$

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

$$\phi_{\alpha}(t) = \left(1 + \frac{1}{2}\alpha t\right) t, \tag{2.10}$$

and $\alpha$ is the chirp rate parameter. Note that the instantaneous frequency $\psi(t)$ at

[3] In Western music, musical notes are disposed according to the equal-tempered scale. The minimum interval in this scale is a semitone, whose relation in frequency is given by $f_{\mathrm{note}(i+1)} = f_{\mathrm{note}(i)} \sqrt[12]{2}$, where the frequency $f_{\mathrm{note}(i)}$ related to a given note $i$ in the scale serve as a reference to the frequency $f_{\mathrm{note}(i+1)}$ of the next note in the scale. This way, the notes are distributed in a geometric fashion in frequency.

time instant $t$ is given by

$$\psi(t) = f\frac{d\phi_\alpha(t)}{dt} = f(1+\alpha t), \tag{2.11}$$

exhibiting a linear variation in time dictated by parameter $\alpha$. This means that signal $x(t)$ can be effectively decomposed into a set of linear chirps, which is suitable for tracking harmonic signals, as desired, since all the harmonic content will share the same $\alpha$.

This transform can be interpreted as a generalization of the Fourier transform, which is obtained with $\alpha = 0$. By applying the variable change $\tau = \phi_\alpha(t)$ to Equation (2.9), the time domain itself can be warped, achieving

$$X^{\mathrm{FChT}}(f,\alpha) = \int_{-1/\alpha}^{\infty} x(\phi_\alpha^{-1}(\tau))\mathrm{e}^{-\mathrm{j}2\pi f\tau}d\tau = \mathcal{F}(x(\phi_\alpha^{-1}(\tau))), \tag{2.12}$$

where $\phi_\alpha^{-1}(\tau)$ is given by

$$\phi_\alpha^{-1}(\tau) = -\frac{1}{\alpha} + \frac{\sqrt{1+2\alpha\tau}}{\alpha}. \tag{2.13}$$

The constraint $x(t) = 0$ for $t \leq -1/\alpha$ should be assured to avoid aliasing [108].

Figure 2.3 depicts the central frequency of each channel of the Fourier transform and the fan-chirp transform in the time-frequency domain. One can see that the channels of the Fourier transform are set to fixed frequencies, while the channels of the FChT all converge to a point $-1/\alpha$ in the time axis.

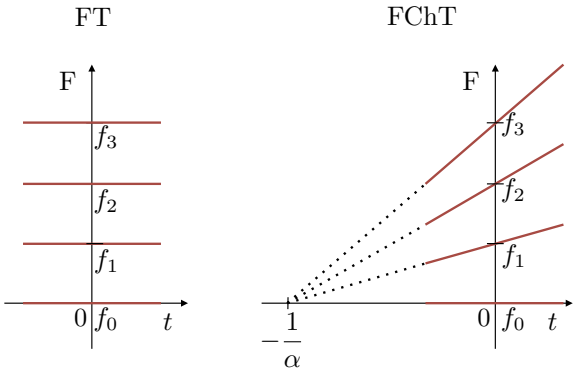



Figure 2.3: Channels of the Fourier transform and the fan-chirp transform.

In Equation (2.12), it is possible to observe that the FChT has the same formulation of the Fourier transform (Equation (2.2)), with the differences that the input signal $x(t)$ is pre-warped in time, and the inferior integration limit is changed.

A FChT-based spectrogram[4] $\mathbf{X}^{\mathrm{FChT}} \in \mathbb{R}^{K\times M}$ can be implemented by means

[4]In this thesis, the term spectrogram will be used in a wider sense to denote every TFR, allowing

of the short-time fan-chirp transform (STFChT), after resampling the input signal $x$ [109]:

$$X^{\mathrm{FChT}}_{k,m,\alpha} \triangleq \left\| \sum_{n=0}^{N-1} \tilde{x}_{\alpha,n-hm} w_n \mathrm{e}^{-j\frac{2\pi}{N}kn} \right\|^2, \tag{2.14}$$

where $\tilde{x}_n$ is the discrete version of the time warped signal $x(\phi_\alpha^{-1}(\tau))$, with the aliasing condition satisfied, and $h$ denotes the hop size. In practice, since $x(t)$ is not available, $\tilde{x}_n$ must be obtained by resampling $x_n$.[5] The constraint related to the aliasing effect restricts the usable values of $\alpha$ inside an analysis window with $N$ samples to be within the interval

$$-2F_\mathrm{s}/N \leq \alpha \leq 2F_\mathrm{s}/N. \tag{2.15}$$

With this formulation, the FChT can profit from a fast implementation of the Discrete Fourier Transform (DFT), i.e. an FFT algorithm [109].

To yield good results, the $\alpha$ parameter must be correctly estimated. This step is originally performed via an exhaustive search, in which a predetermined set of values of $\alpha$ is tested, and the choice of the best one is made by searching for the value of $\alpha$ that maximizes the salience[6] function [70, 109]. Another reliable and significantly faster way to perform this estimation, proposed in [110], consists in using the structure tensor [111, 112] image technique to estimate the direction of the frequency lines, and consequently, the parameter $\alpha$ over time. This method will be studied in Chapter 5.

Since the fan-chirp transform follows the evolution of the fundamental frequency in time, larger analysis windows can be successfully used to describe monophonic harmonic signals. To illustrate this, Figure 2.4 depicts a spectrogram of a synthetic vibrato signal, computed with 2048 samples. One can see that the frequency lines with steep slopes are poorly represented. Figure 2.5 depicts a FChT-based spectrogram of the same signal, where the frequency components can be seen much better described.

Note that this method can only well represent one monophonic harmonic sound source at a time, since the whole TFR follows a specific fundamental frequency variation. Therefore, the problem of dealing with multiple monophonic sources can only be tackled by combining multiple instances of STFChT, each one optimized for representing one source, in a way that the best representations remain in the final

one to better distinguish the time-frequency transforms from the TFRs computed with them.

[5] In practice, in order to simplify the computation of $\tilde{x}_n$, the original signal $x_n$ is resampled to have twice the original sampling frequency and a simple linear interpolation is performed to estimate $\tilde{x}_n$, according to Equation (2.13). In the current implementation, the analysis window is applied after this resampling procedure.

[6] The salience function relates the amplitude of the frequency bins with the amplitude of their harmonics. High salience values are then indicators of the presence of well-defined harmonic sources.

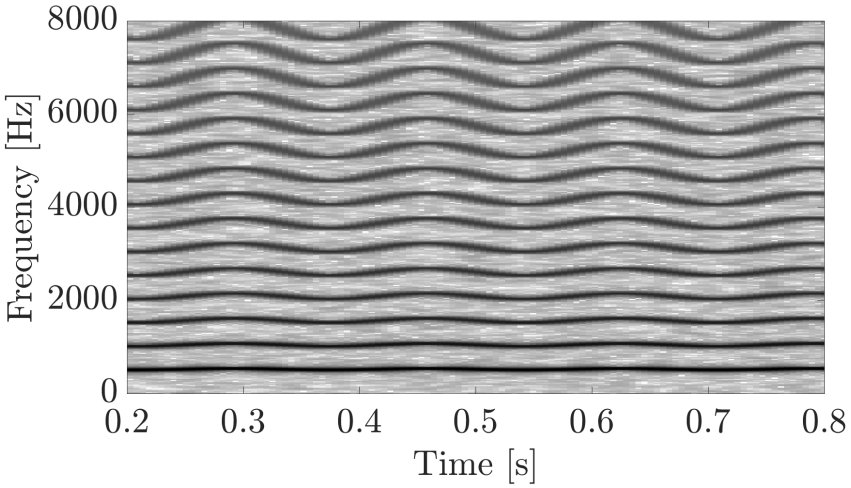



Figure 2.4: Spectrogram of a synthetic vibrato signal ($N = 2048$).

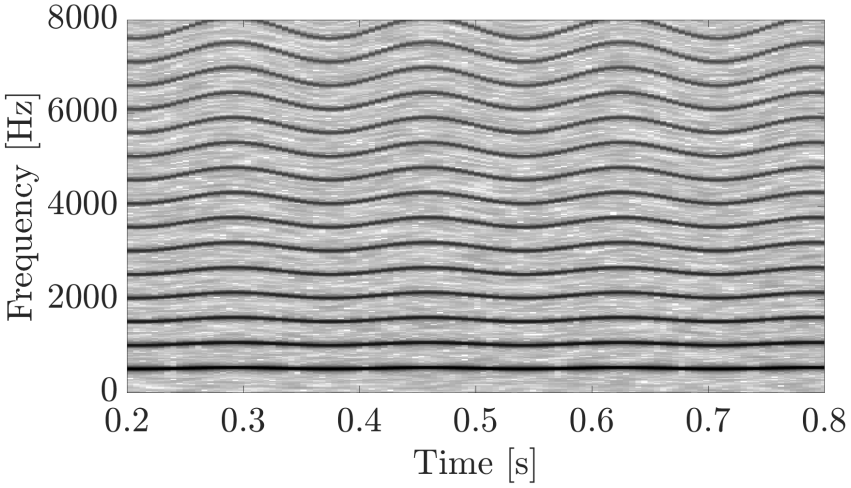



Figure 2.5: FChT-based spectrogram of a synthetic vibrato signal ($N = 2048$).

TFR. This is also addressed in Chapter 5.

## 2.3 Log-frequency Transforms

### 2.3.1 The Constant-$Q$ Transform

Another interesting time-frequency representation, which is very useful for musical signal analysis, is the constant-$Q$ transform (C$Q$T) [113–116], in which the $Q$ refers to the quality factor, or selectivity, of the filters implicitly used in the transform. This transform has the desirable property of providing logarithmic resolution[7] in frequency, which matches the way musical notes ascend in the equal-tempered scale.

[7]Here, it is useful to distinguish the frequency scale, which is related to the distribution of the center frequencies of the filters used in the transform, from the frequency resolution, which is related to the frequency width of these filters. The C$Q$T is usually represented in logarithmic scale, although it is possible to use a linear scale [116].

The quality factor can be defined as

$$Q = \frac{f_k}{\Delta f_k}, \tag{2.16}$$

where $f_k = \sqrt{f_i f_f}$ is the central frequency of the $k$-th filter with pass-band[8] between $f_i$ and $f_f$, and $\Delta f = f_f - f_i$.

The C$Q$T is an adaptation of the DFT, whose channels have a constant $\Delta f = F_\mathrm{s}/N$, where $F_\mathrm{s}$ is the sampling frequency and $N$ is the number of samples analyzed. In the case of the C$Q$T, since $Q$ is proportional to $F_\mathrm{s}$, the constant $N$ becomes

$$N_k = \frac{F_\mathrm{s}}{\Delta f_k} = \frac{Q F_\mathrm{s}}{f_k}, \tag{2.17}$$

which then varies with the channel index $k$.

With this modification, a C$Q$T can be defined by substituting $f_k/F_\mathrm{s}$ by $Q/N_k$, achieving

$$X_k^{CQT} = \frac{1}{\sqrt{N_k}} \sum_{n=0}^{N_k-1} w_{k,n} x_n \mathrm{e}^{-j2\pi Q n/N_k}, \tag{2.18}$$

where window $w_{k,n}$ is a function of $n$ and $k$, since it must follow the number of samples used per channel. For this same reason, normalization by $1/\sqrt{N_k}$ is necessary. Similarly to what was defined for the standard spectrogram, a C$Q$T-based spectrogram $\mathbf{X}^{CQT}$ can be computed as

$$X_{k,m}^{CQT} = \frac{1}{\sqrt{N_k}} \left\| \sum_{n=0}^{N_k-1} w_{k,n} x_{n-hm} \mathrm{e}^{-j2\pi Q n/N_k} \right\|^2. \tag{2.19}$$

The quality factor $Q$ can be set to provide a frequency resolution related to a given number of bins/octave $b$ as

$$Q = \frac{f_k}{\Delta f} = \frac{f_k}{(2^{1/b} - 2^{-1/b}) f_k}, \tag{2.20}$$

where $\Delta f$ is the bandwidth. For instance, if a separation of a quarter of tone is required, one can use $b = 48$, since there are 12 semitones in an octave, geometrically related. In this case, $Q = \frac{f_k}{(2^{1/48} - 2^{-1/48}) f_k} \approx 34.6$. The other free variables to be set in the C$Q$T are the minimum and maximum frequencies to be analyzed, which will define the central and limit frequencies of each channel.

When implemented as formulated here, the C$Q$T is very time-consuming. However, there exist several different approximations which dramatically reduce the cost of the C$Q$T. For instance, FFT algorithms, filter banks and multi-resolution pro-

[8]The pass-band is being considered as the region in which the frequency-response of each band-pass filter is within the interval $[-6, 0]$ dB.

cessing schemes can be used [105, 114–117]. This way, C$Q$T-based spectrograms (or approximations) can also be computed, with relatively low computational cost. In this thesis, the implementation provided in [79] was adopted, in which full rasterization is provided: only the highest frequency channel is critically sampled and all other channels are subsampled with the same rate.

Figure 2.6 depicts a C$Q$T-based spectrogram of a piano recording, in which some interesting characteristics can be observed. The C$Q$T has a fixed distribution of the harmonic content, in terms of distance in the frequency axe, since the harmonics are always multiples of the fundamental frequency and the frequency resolution is logarithmic, as can be observed in Figure 2.6, where the first three harmonics of four different notes are highlighted. On the downside of this transform is the poor representation of transient information at low frequencies.

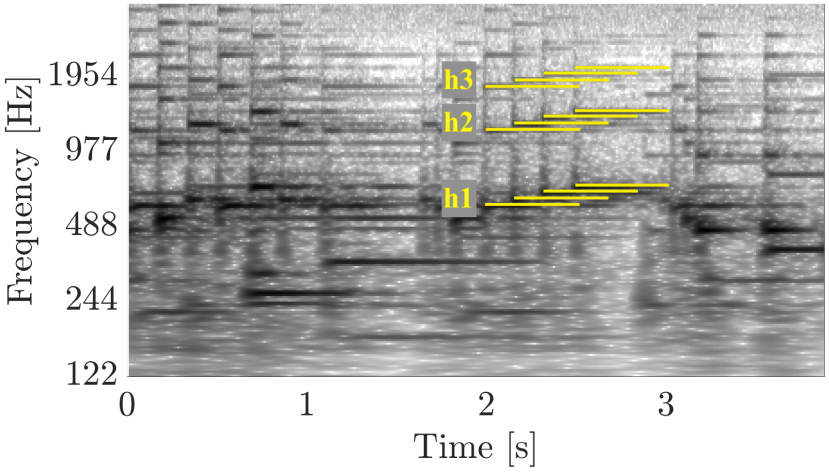



Figure 2.6: Equal harmonic distribution for different notes.

### 2.3.2 Harmonic C$Q$T

The fact that the harmonic distances are note-independent in a C$Q$T representation can be explored in a convolutive approach to note searching, since a fixed two-dimensional filter can be used. Many procedures make use of a tensor in which each layer comprises a certain range of the C$Q$T spectrogram, in such a way that the harmonics of each note are aligned throughout the layers. This can be performed by computing C$Q$Ts having different minimum frequency $f_{\min}i$ and maximum frequency $f_{\max}i$ for each layer $i$. Since this can be time-consuming, a C$Q$T spectrogram can be computed comprising the maximum range necessary for all layers and interpolated to fit the specific range of each layer. This tensor representation is the so-called harmonic C$Q$T (HC$Q$T), and was proposed in [93]. It is useful especially for convolutional neural networks (CNNs), which benefit from this structure. A

visual example of this approach is seen in Figure 2.7. Note that the fundamental frequency of each highlighted note present in the left spectrogram aligns with its second harmonic in the second spectrogram, with the third harmonic in the third spectrogram, and so on. The HCQT is then formed by the region of each spectrogram inside the dashed-line area.

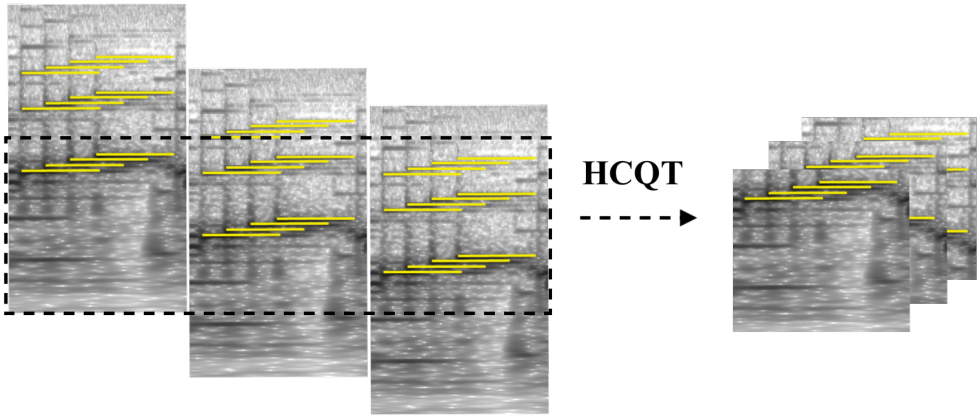



Figure 2.7: Equal harmonic distribution for different notes.

### 2.3.3 Variable-$Q$ Transform

Among many variants of algorithms for the computation of the C$Q$T, there is also a TFR that exhibits uniformly-distributed frequency bins in log-frequency, whose $Q$ factors obey a linear function [79]. In this transform, the bandwidth $\Delta f$ is given by

$$\Delta f = (2^{1/b} - 2^{-1/b})f + \kappa, \tag{2.21}$$

where $\kappa$ is a constant parameter provided to set the variation of $Q$ (see [79] for details). With this parameter, the bandwidth is no longer proportional to the frequency, and it allows, for instance, the computation of TFRs whose bandwidths are constant on the auditory critical-band scale, with a smooth variation of $Q$, or even a common C$Q$T if $\kappa = 0$. This transform provides particularly interesting results whenever preserving a good time resolution at low frequencies is desirable.

Figure 2.8 depicts two spectrograms[9] of a signal containing piano and vocal, with parameters $\kappa = 0$ (equivalent to C$Q$T) and $\kappa = 30$, respectively. Note that the non-stationary components of the lower part of the spectrum, which comprises

[9]In this work, the implementation provided in [79] was used for the computation of the V$Q$T.

the first harmonics of the voice and the lower piano notes, are significantly better defined in the spectrogram using $\kappa = 30$, at the expense of having thicker stationary frequency components.

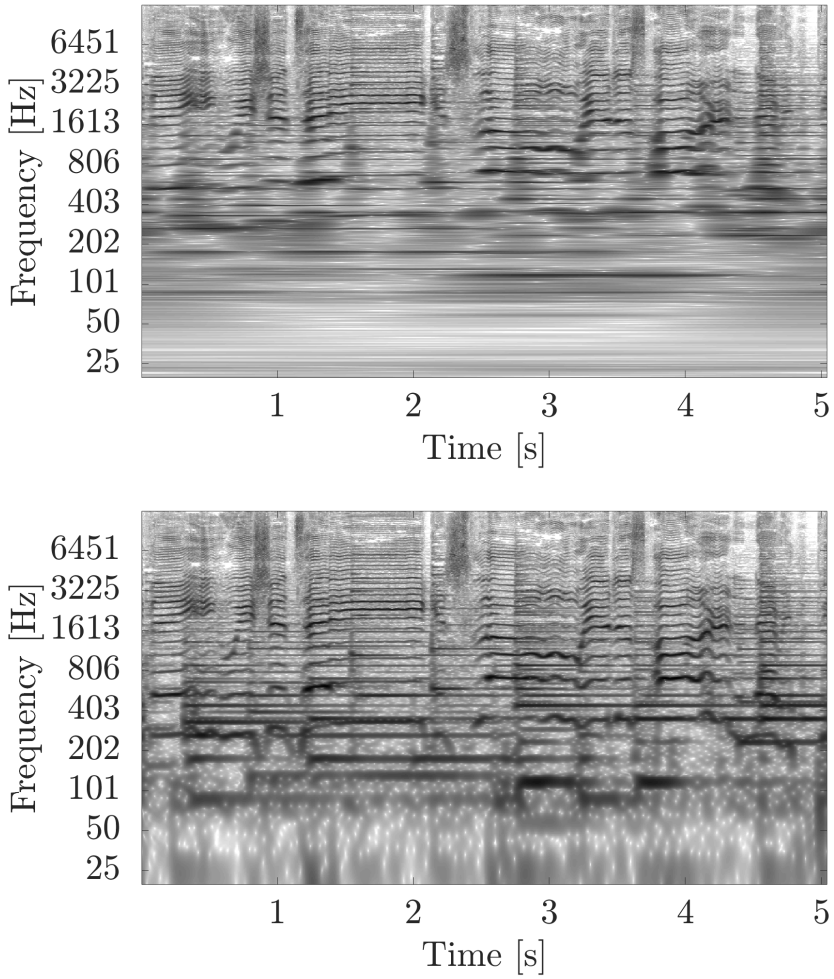



Figure 2.8: Spectrograms of a signal containing piano and voice, with parameters $\kappa = 0$ (equivalent to C$Q$T) and $\kappa = 30$, respectively.

## 2.4 Concluding Remarks

In this chapter, some TFRs were presented. First, the well-known STFT-based spectrogram was derived. This representation presents linear frequency resolution and requires a low computational power, due to the FFT algorithm used to compute the spectrum of each time frame. The FChT-based spectrogram is a variation of this TFRs that is able to represent a harmonic sound source whose fundamental frequency varies continually in time, by modelling the frames as having linear frequency variations. This method also presents linear frequency resolution and can only properly represent sound sources which share the same $\alpha$ parameter.

Other three representations were presented, having log-frequency scale: the C$Q$T, the V$Q$T and the HC$Q$T. The C$Q$T also presents logarithmic resolution in frequency, while the V$Q$T does not maintain a constant $Q$; instead, the frequency resolution in log scale decreases for low frequencies, preserving a better time resolution in such areas. As mentioned, these representations are useful as they fit the geometric distribution of musical notes in the equal-tempered scale. The HC$Q$T is a tensor which stacks regions of C$Q$Ts in such a way that the harmonics of the sound sources remain aligned throughout the layers, allowing for a multi-layer processing of the TFRs, typically found in systems based on deep neural networks. Every log-scale representation is suitable for this kind of stacking procedure.

On the following chapters, methods for combination of TFRs will be presented.

# Chapter 3

# Bin-wise Combination of Spectrograms

Spectrograms computed with different parameters (analysis window length, bins per octave, $\alpha$ chirp rate, etc.) are capable of best representing certain aspects of audio signals, often in a complementary fashion. Nevertheless, it is desirable to have one representation which could gather the beneficial aspects of all the representations available. Following this idea, spectrogram combination methods using different approaches are studied.

In order to combine transforms with different characteristics, e.g. spectrograms computed with analysis windows of different lengths, all bins with same line and column indexes must be related to the same time-frequency bins throughout all TFRs, and therefore all TFRs must deliver the same number of samples in frequency and time domains. Also, all TFRs must have the same energy $E$. The dimensionality can be equalized either via 2-D interpolation or via zero-padding and some time-alignment procedure, while the energy match can be assured by simple energy scaling. The 2-D interpolation[1] is used in this work, to facilitate the implementation using different types of spectrogram.

All combination methods in this work will be presented as the combination of spectrograms stacked in a generic tensor $\underline{\mathbf{X}} \in \mathbb{R}^{K \times M \times P}$, with $K$ frequency bins, $M$ time frames, and $P$ different spectrograms; the element at bin $(k, m)$ of the $p$-th spectrogram is denoted by $X_{k,m}[p]$. In this chapter, bin-wise combinations of spectrograms (along the $p$ dimension)[2] will be studied.

[1] In this thesis, linear 2-D interpolation was used, since it works for all kinds of frequency scale.

[2] Here, all spectrograms are computed using the Hamming window, which is one of the most applied functions for this purpose. Such a choice, although arbitrary, has no significant effect on the results.

## 3.1 The Numerical Mean

This first combination method studied is the numerical mean (NM) [118], which is the solution that minimizes the mean squared error between the combination result and the individual spectrograms. Using the appropriate cost function,

$$X_{k,m}^{\mathrm{NM}} = \arg\min_{\hat{X}_{k,m}} \frac{1}{P} \sum_{p=1}^{P} \sum_{k=0}^{K-1} \sum_{m=1}^{M} (\hat{X}_{k,m} - X_{k,m}[p])^2, \tag{3.1}$$

subject to the energy constraint

$$\|\mathbf{X}^{\mathrm{NM}}\|_1 = \sum_{k=0}^{K-1} \sum_{m=1}^{M} X_{k,m}^{\mathrm{NM}} = E, \tag{3.2}$$

where $\|\cdot\|_1$ denotes the entrywise $L_1$ norm of a matrix and

$$E = \sum_{k=0}^{K-1} \sum_{m=1}^{M} X_{k,m}[p], \forall p. \tag{3.3}$$

This constrained optimization problem can be solved using the method of Lagrange multipliers by finding the critical points of the Lagrangian:

$$L(\hat{\mathbf{X}}, \lambda) = \frac{1}{P} \sum_{p=1}^{P} \sum_{k=0}^{K-1} \sum_{m=1}^{M} (\hat{X}_{k,m} - X_{k,m}[p])^2 + \lambda (\sum_{k=0}^{K-1} \sum_{m=1}^{M} \hat{X}_{k,m} - E), \tag{3.4}$$

where scalar $\lambda$ is the Lagrange multiplier to be determined. The solution is then obtained by making the derivative of the Lagrangian with respect to $\hat{X}_{k,m}$ equal to zero:

$$\frac{dL(\hat{\mathbf{X}}, \lambda)}{d\hat{X}_{k,m}} = 2\hat{X}_{k,m} + \lambda - \frac{2}{P} \sum_{p=1}^{P} X_{k,m}[p] = 0, \tag{3.5}$$

which leads to

$$\hat{X}_{k,m} = -\frac{\lambda}{2} + \frac{1}{P} \sum_{p=1}^{P} X_{k,m}[p]; \tag{3.6}$$

the energy constraint is satisfied with $\lambda = 0$, yielding the NM solution

$$X_{k,m}^{\mathrm{NM}} = \frac{1}{P} \sum_{p=1}^{P} X_{k,m}[p]. \tag{3.7}$$

This means that the sample at bin $(k, m)$ at the combined representation will be computed by averaging the samples of all the representations available at bin $(k, m)$.

To visualize the resulting combination, two test signals were used: one comprised of a sinusoid with frequency 1 kHz and another comprised of an impulse at 0.5 s, both

with sampling frequency of 44.1 kHz. Three spectrograms, using analysis windows with 1024, 2048, and 4096 samples, respectively, were computed for these test signals, and then combined according to Equation (3.7). Figure 3.1 depicts the spectrum of a time frame of such spectrograms for the sinusoidal signal, and Figure 3.2 depicts the evolution of a frequency bin over time for the impulse signal. Zero-padding with a factor of 8 and a hop size of 64 samples were used in order to provide a sufficient number of samples to illustrate the signals in detail.

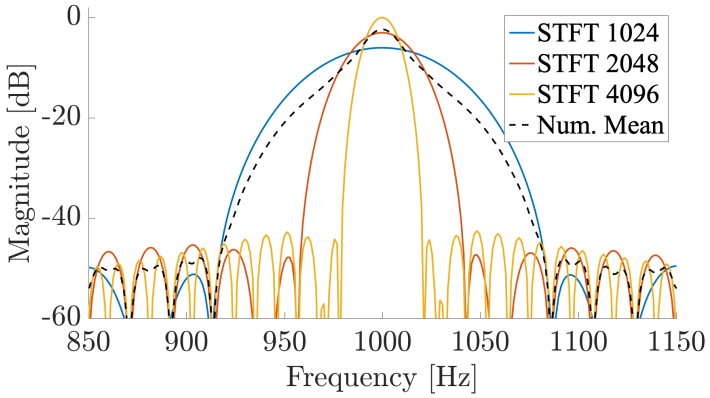



Figure 3.1: Spectrum of a time frame for a sinusoid (1 kHz): different STFT's and the numerical mean combination.

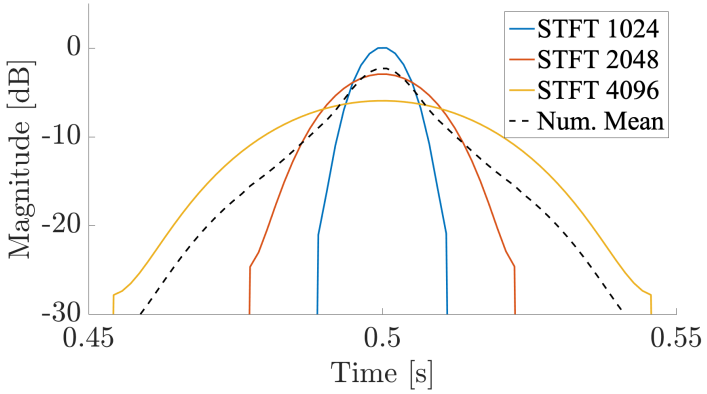



Figure 3.2: Time evolution of a frequency bin for an impulse at 0.5 s for different STFT's and the numerical mean combination.

In the first image, one can see the general behaviour of what happens frequency-wise: the frequency peak is better defined by the combined representation than by the spectrogram computed with a short analysis window, but the spectrogram computed with a long window is still the best representation in this context; on the other hand, as for the impulse, seen in the second figure, the spectrogram with a short window provides the best representation for this signal, as expected.

In these examples, the ideal combined representation should be as close as possible to the most sparse spectrograms available, i.e. should provide the most pronounced peaks possible. Although the NM method is far from ideal, it still provides

a good overall time-frequency representation.

## 3.2 The Reciprocal Mean

The reciprocal mean (RM) combination is also an optimum solution, which minimizes the Itakura-Saito error [118]. The cost function for each bin is given by

$$J_{k,m} = \frac{1}{P}\sum_{p=1}^{P}\left[\frac{\hat{X}_{k,m}}{X_{k,m}[p]} - \ln\frac{\hat{X}_{k,m}}{X_{k,m}[p]} - 1\right], \tag{3.8}$$

whose unconstrained solution is normalized to match the energy constraint of Equation (3.2). Deriving this cost function and equating it to zero

$$\frac{dJ_{k,m}}{d\hat{X}_{k,m}} = \frac{1}{P}\sum_{p=1}^{P}\left[\frac{1}{X_{k,m}[p]} - \frac{1}{\hat{X}_{k,m}}\right] = 0, \tag{3.9}$$

one gets the following relation:

$$\frac{1}{P}\sum_{p=1}^{P}\frac{1}{X_{k,m}[p]} = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{\hat{X}_{k,m}} = \frac{1}{\hat{X}_{k,m}} \tag{3.10}$$

and the reciprocal mean (RM) solution

$$X_{k,m}^{\mathrm{RM}} = c^{\mathrm{RM}}\left(\frac{1}{P}\sum_{p=1}^{P}X_{k,m}[p]^{-1}\right)^{-1}, \tag{3.11}$$

where the term $c^{\mathrm{RM}}$ just equalizes the global energy to match with the energy $E$ of the individual spectrograms, and is given by

$$c^{\mathrm{RM}} = \frac{E}{\left\|\left(\frac{1}{P}\sum_{p=1}^{P}X_{k,m}[p]^{-1}\right)^{-1}\right\|_1}. \tag{3.12}$$

Again, the sinusoid and impulse signals were used to illustrate the results of this combination procedure, which can be seen in Figures 3.3 and 3.4. As can be observed, the reciprocal mean yields much better results than the numerical mean, producing well defined peaks that are close to the best representation among the ones available for each signal. It is worth noting that, differently from the numerical mean, the RM representation goes to zero whenever one of the representations goes to zero. This can be observed in Figure 3.4.

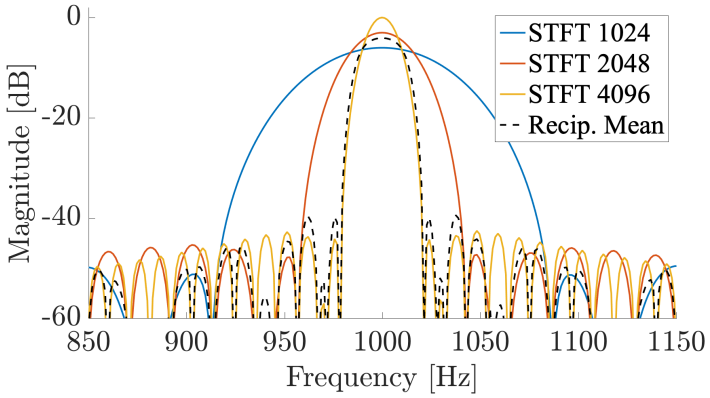



Figure 3.3: Spectrum of a time frame for a sinusoid (1 kHz): different STFT's and the reciprocal mean combination.

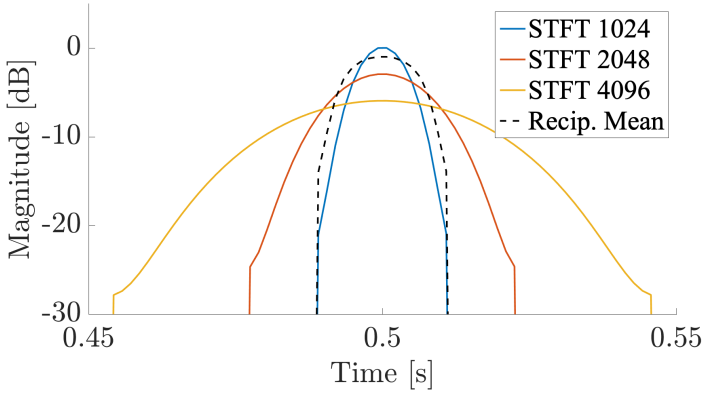



Figure 3.4: Time evolution of a frequency bin for an impulse at 0.5 s for different STFT's and the reciprocal mean combination.

## 3.3 The Geometric Mean

In [87], it is shown that the geometric mean (GM) results in an optimal combination of spectrograms in the sense of minimum mean cross-entropy[3] between the combined spectrogram and the individual ones [119], expressed by

$$X^{\mathrm{G}}_{k,m} = \arg\min_{\hat{X}_{k,m}} \frac{1}{P} \sum_{p=1}^{P} \sum_{k=0}^{K-1} \sum_{m=1}^{M} \hat{X}_{k,m} \log \frac{\hat{X}_{k,m}}{X_{k,m}[p]}, \tag{3.13}$$

also subject to the energy constraint of Equation (3.2). The sum along the $p$ dimension can be transformed into a product inside the logarithm

$$X^{\mathrm{G}}_{k,m} = \arg\min_{\hat{X}_{k,m}} \sum_{k=0}^{K-1} \sum_{m=1}^{M} \hat{X}_{k,m} \log \frac{\hat{X}_{k,m}}{\prod_{p=1}^{P} X_{k,m}[p]^{\frac{1}{P}}}, \tag{3.14}$$

[3]The cross-entropy minimization is a well known concept in the field of information theory, which relates two probabilities distributions allowing one to obtain the least impaired joint density, given only partial information [87].

and, once again, the method of Lagrange multipliers can be used to solve this optimization problem, by the following Lagrangian up to constant terms:

$$L(\hat{\mathbf{X}}, \lambda) = \sum_{k=0}^{K-1} \sum_{m=1}^{M} \hat{X}_{k,m} \log \frac{\hat{X}_{k,m}}{\prod_{p=1}^{P} X_{k,m}[p]^{\frac{1}{P}}} + \lambda(\sum_{k=0}^{K-1} \sum_{m=1}^{M} \hat{X}_{k,m} - E). \tag{3.15}$$

The derivative of the Lagrangian with respect to $\hat{X}_{k,m}$ is made equal to zero:

$$\frac{dL(\hat{\mathbf{X}}, \lambda)}{d\hat{X}_{k,m}} = \log \frac{\hat{X}_{k,m}}{\prod_{p=1}^{P} X_{k,m}[p]^{\frac{1}{P}}} + 1 + \lambda = 0. \tag{3.16}$$

The minimum mean cross-entropy solution is then yielded by solving this equation for $\hat{X}_{k,m}$,

$$X_{k,m}^{\mathrm{GM}} = b^{-(1+\lambda)} \prod_{p=1}^{P} X_{k,m}[p]^{\frac{1}{P}}, \tag{3.17}$$

where $b$ is the base of the logarithm. The geometric mean solution is then obtained by applying the energy constraint to determine $\lambda$,

$$X_{k,m}^{\mathrm{GM}} = c^{\mathrm{GM}} \prod_{p=1}^{P} X_{k,m}[p]^{\frac{1}{P}}, \tag{3.18}$$

where

$$c^{\mathrm{GM}} = \frac{E}{\| \prod_{p=1}^{P} X_{k,m}[p]^{\frac{1}{P}} \|_1} \tag{3.19}$$

is the energy-matching constant.

This method was developed heuristically in [120] as a way to combine spectrograms processed with different window lengths, preserving important visual features of each spectrogram.

The same test signals were used to compare the GM combination with the standard spectrograms, which are shown in Figures 3.5 and 3.6. Again, as expected, the combined curves are good candidates to represent the overall behaviour of all the spectrograms, but the resulting curves are less sparse than the ones from the RM combination. Nevertheless, the combination is still a better overall representation than any of the spectrograms considering both time and frequency resolutions.

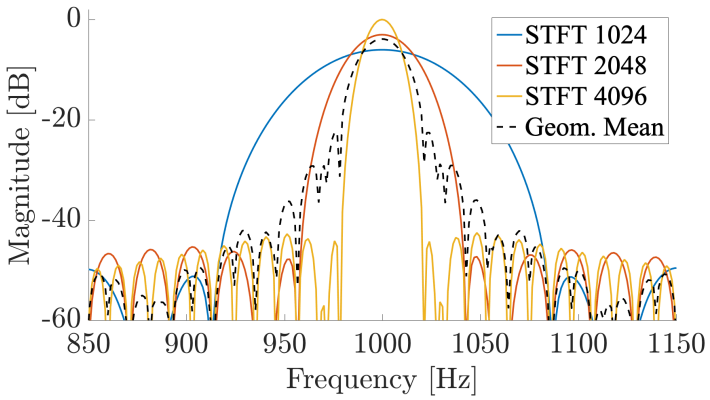



Figure 3.5: Spectrum of a time frame for a sinusoid (1 kHz): different STFT's and the geometric mean combination.

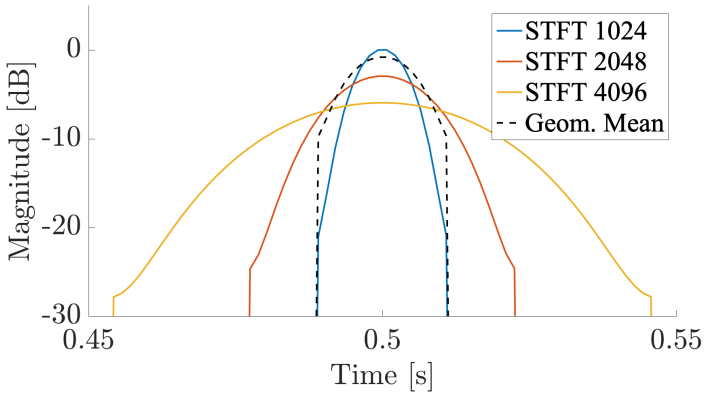



Figure 3.6: Time evolution of a frequency bin for an impulse at 0.5 s for different STFT's and the geometric mean combination.

## 3.4 The Minimax

Another optimum combination can be obtained by minimizing the maximum cross-entropy across the spectrograms [87],

$$X_{k,m}^{\mathrm{MM}} = \arg\min_{\hat{X}_{k,m}} \max_{p} \sum_{k=0}^{K-1} \sum_{m=1}^{M} \hat{X}_{k,m} \log \frac{\hat{X}_{k,m}}{X_{k,m}[p]}, \tag{3.20}$$

subject to the energy constraint of Equation (3.2). Note that, in order to solve this problem, N optimizations must be performed (one for each spectrogram in the set). Jensen's inequality can be used to reduce the dimensionality of this problem to a single optimization calculation, by setting an upper bound

$$\max_{p} \sum_{k=0}^{K-1} \sum_{m=1}^{M} \hat{X}_{k,m} \log \frac{\hat{X}_{k,m}}{X_{k,m}[p]} \leq \sum_{k=0}^{K-1} \sum_{m=1}^{M} \max_{p} \hat{X}_{k,m} \log \frac{\hat{X}_{k,m}}{X_{k,m}[p]}. \tag{3.21}$$

Now, the problem can be reformulated as

$$X_{k,m}^{\mathrm{MM}} = \arg\min_{\hat{X}_{k,m}} \sum_{k=0}^{K-1} \sum_{m=1}^{M} \max_{p} \hat{X}_{k,m} \log \frac{\hat{X}_{k,m}}{X_{k,m}[p]}, \tag{3.22}$$

which is again subject to the same energy constraint of the other methods. Due to the monotonicity of the logarithm, this problem is equivalent to

$$X_{k,m}^{\mathrm{MM}} = \arg\min_{\hat{X}_{k,m}} \sum_{k=0}^{K-1} \sum_{m=1}^{M} \hat{X}_{k,m} \log \frac{\hat{X}_{k,m}}{\min_p X_{k,m}[p]}. \tag{3.23}$$

Using the Lagrangian, setting its derivative to zero, and applying the same energy constraint as done previously, we have the minimax (MM) solution, given by

$$X_{k,m}^{\mathrm{MM}} = c^{\mathrm{MM}} \min_{p} X_{k,m}[p], \tag{3.24}$$

with energy-matching constant

$$c^{\mathrm{MM}} = \frac{E}{\| \min_p X_{k,m}[p] \|_1}. \tag{3.25}$$

Figures 3.7 and 3.8 depict the results of such a combination, compared to the test signals. As can be observed, the results are very sparse, since the lower valued samples are selected to compose the combined representation, but at the expense of flattering the top of the peaks. Note that the combined curves match the lowest samples, but there is a constant offset applied, which happens due to the energy matching.

## 3.5 The Sample-Weighted Geometric Mean

The sample-weighted geometric mean (SWGM), proposed in [75], is a combination that sits in between the geometric mean and the minimax solutions, and one of the contributions of this thesis. The main idea is to combine the spectrograms in such a way that the lowest valued sample takes precedence, increasing the sparsity. In fact, this principle can be observed in both the minimax and the geometric mean combinations. The minimax applies absolute importance to the samples having minimum values; in the geometric mean, whenever a sample is close to zero, the result tends to be close to zero.

Our solution aims at controlling this effect by performing a weighted geometric mean, giving more weight to smaller values by means of a weighting function which depends on the samples to be weighted themselves. The SWGM method can be

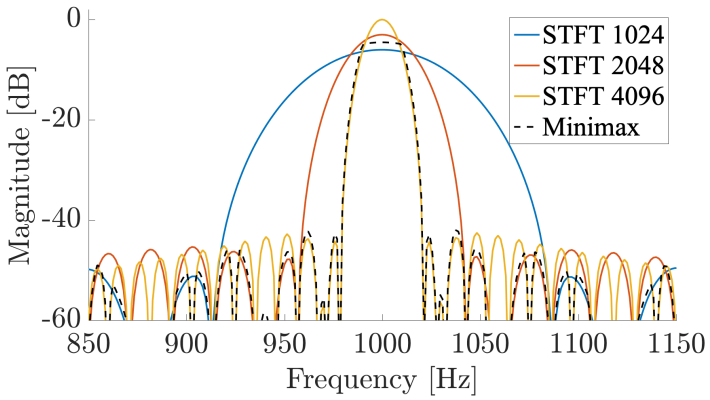



Figure 3.7: Spectrum of a time frame for a sinusoid (1 kHz): different STFT's and the minimax combination.

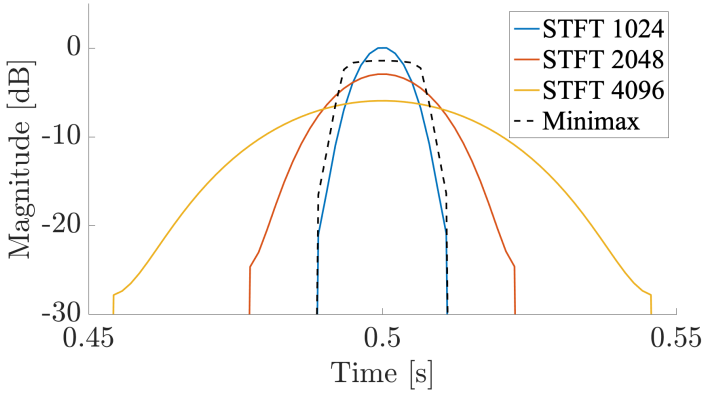



Figure 3.8: Time evolution of a frequency bin for an impulse at 0.5 s for different STFT's and the minimax combination.

described by

$$X_{k,m}^{\text{SWGM}} = c^{\text{SWGM}} \left( \prod_{p=1}^{P} X_{k,m}[p]^{\gamma_{k,m}[p]} \right)^{\frac{1}{\sum_p \gamma_{k,m}[p]}}, \tag{3.26}$$

where $\gamma \in \mathbb{R}^{K \times M \times P}$ is the weighting function aforementioned

$$\gamma_{k,m}[p] = \left( \frac{\prod_{l \neq p} X_{k,m}[l]^{\frac{1}{P-1}}}{X_{k,m}[p]} \right)^{\beta}, \tag{3.27}$$

where $\beta \in \mathbb{R}_{\geq 0}$ is a constant intended to regulate the weighting process, and

$$c^{\text{SWGM}} = \frac{E}{\left( \prod_{p=1}^{P} X_{k,m}[p]^{\gamma_{k,m}[p]} \right)^{\frac{1}{\sum_p \gamma_{k,m}[p]}}} \tag{3.28}$$

is the energy-matching gain.

The weight applied to the $p$-th spectrogram is the ratio between the geometric

mean of the samples of the remaining spectrograms,[4] and the sample at bin $(k, m)$ of the $p$-th spectrogram. This necessarily results in larger weights for lower samples.

This fraction is also raised to the power of $\beta$ to allow for some control over the weights applied to the mean. Small values of $\beta$ provide smooth transitions, i.e. lower peak distortion, with $\beta = 0$ corresponding to the GM combination. With $\beta > 0$, the weights take effect, and a sharper spectrogram is achieved, exhibiting higher sparsity as $\beta$ grows. Assigning $\beta = \infty$ is equivalent to computing the MM solution. Usually, setting $0 \leq \beta \leq 0.5$ provides the best results for most cases.[5]

In order to illustrate this principle, the same spectrograms used for the previous combination methods were combined using the GM, the MM and the SWGM with different configurations of $\beta$. Figures 3.9 and 3.10 depict the resulting combinations, without energy compensation, to facilitate observing the boundaries. Here, one can see the transition between the GM and the MM combinations as $\beta$ increases. It is worth highlighting the fact that the SWGM combinations can be almost as sparse as the MM combinations, with the advantage of better preserving the shape of the top of the peak. This happens because the discrepancy between the samples near the peak for the different spectrograms is not as high as in the side-lobes. For this reason, the resulting side-lobes can get very close to the MM solution ones, while better keeping the shape of the main lobe.

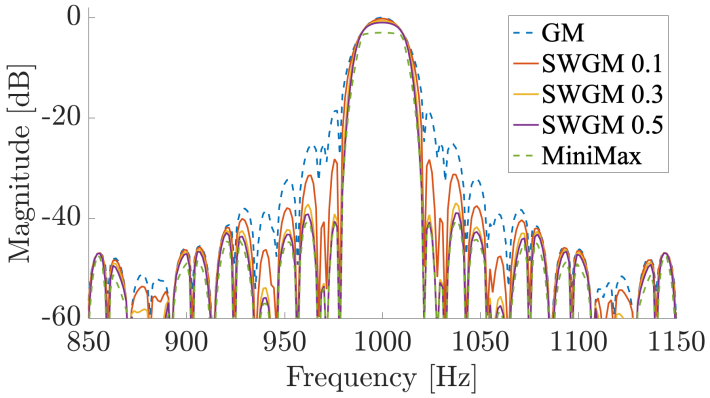



Figure 3.9: Spectrum of a time frame for a sinusoid (1 kHz): SWGM with different configurations of $\beta$, GM, and MM combinations, without global energy compensation.

It is important to note that since the weighted geometric mean is performed by raising the samples to the power of the weighting function and such function can reach high values, it can easily lead to numeric problems. An easy way to avoid this

[4]Here, there is not the necessity of performing the geometric mean. Any operation that varies continually with the samples' values and results in a value between them will work.

[5]Differently from [75], in which combination is performed over the compressed version of the spectrograms, here the energy spectrograms are combined, hence the difference in values of $\beta$. This different approach was chosen for the sake of homogeneity in relation to the other methods.

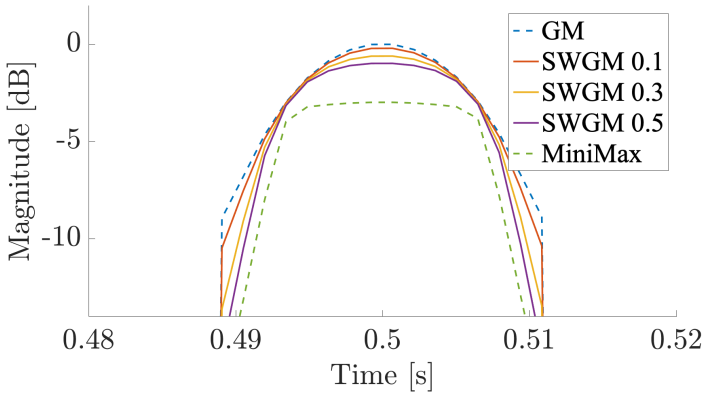



Figure 3.10: Time evolution of a frequency bin for an impulse at 0.5 s: SWGM with different configurations of $\beta$, GM, and MM combinations, without global energy compensation.

problem is setting an upper bound for $\gamma$, e.g. $\gamma \leq 20$.

Figures 3.11 and 3.12 depict the SWGM combination using $\beta = 0.5$, now applying the energy matching gain. As can be seen, the results are very sparse and close to the best spectrogram for each case.

## 3.6 Concluding Remarks

All the methods presented in this chapter combine different representations in order to enhance time and frequency resolutions simultaneously. In fact, it is important to highlight the fact that such procedures do not break the uncertainty principle, which only constrains the individual standard deviations in time and frequency for a signal and its Fourier transform [91], but not to a combination of spectra computed from signals (frames) with different time lengths. In other words, the uncertainty principle does not hold for all kinds of time-frequency representation. Furthermore, the results of the combination will never provide peaks in frequency which are narrower than the ones provided by spectrograms with long windows or have better time definition than what is provided by the spectrogram computed with short windows.

It is also important to mention that, since the spectrograms are interpolated to provide the same dimensions and be combined, the new TFRs cannot be inverted back to time domain, or at least not perfectly. Therefore, these combination procedures can be very useful whenever a time-frequency representation is needed, but only for analysis purposes.

In Figures 3.13, 3.14 and 3.15, all the combination methods studied in this chapter are presented, for comparison, using the same three spectrograms as before and compensated in energy.

As mentioned, the NM and the GM methods provide the lowest overall definition.

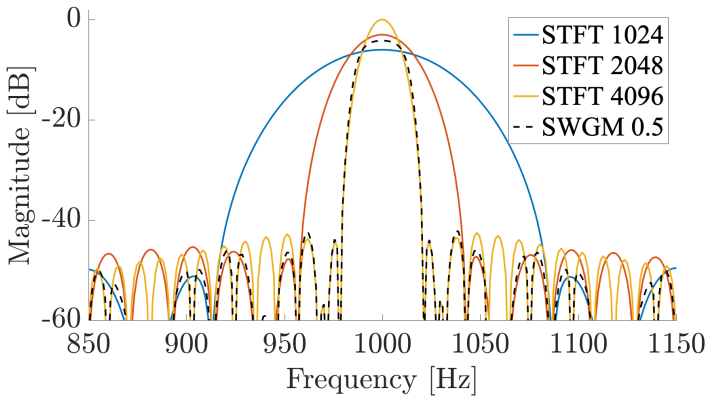



Figure 3.11: Spectrum of a time frame for a sinusoid (1 kHz): different STFT's and the SWGM combination.

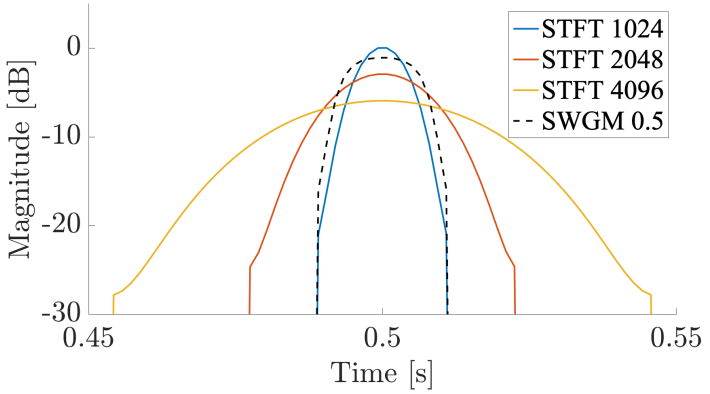



Figure 3.12: Time evolution of a frequency bin for an impulse at 0.5 s for different STFT's and the SWGM combination.

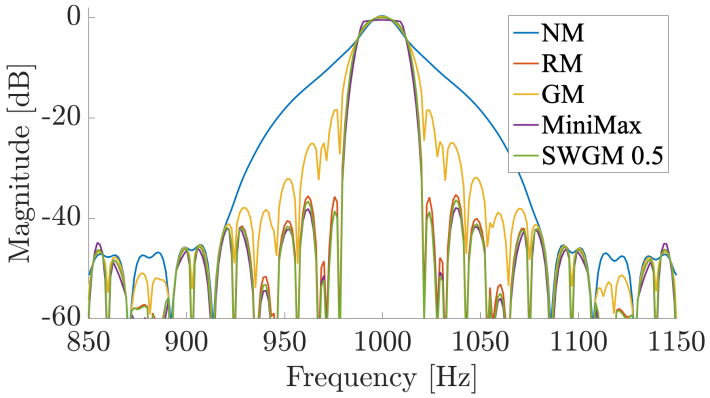



Figure 3.13: Spectrum of a time frame for a sinusoid (1 kHz): all combination methods.

As for the other methods, the SWGM provides slightly better results than the RM in terms of sparsity, while both present much better definition of the top part of the peaks, i.e. peak distortion, than the MM. One can state that among the methods presented, the SWGM provides the best results, immediately followed by the RM

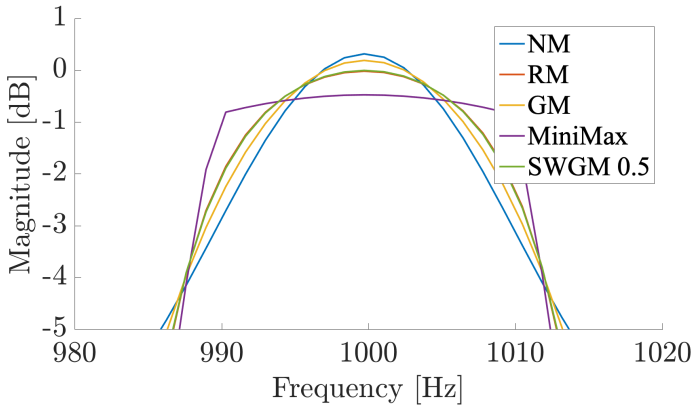



Figure 3.14: Spectrum of a time frame for a sinusoid (1 kHz): all combination methods (zoom).

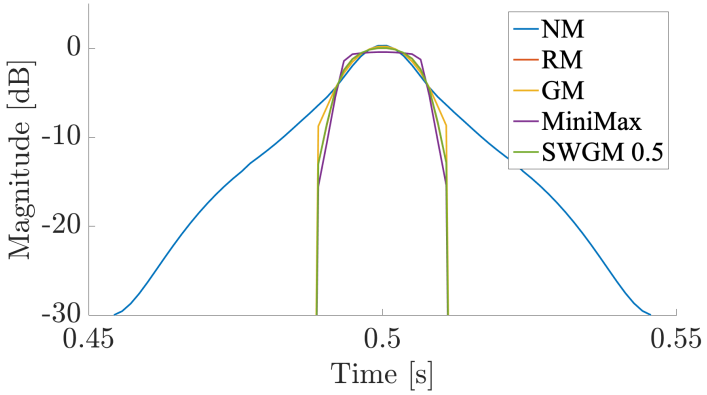



Figure 3.15: Time evolution of a frequency bin for an impulse at 0.5 s: all combination methods.

combination.

In order to illustrate the effect of combining TFRs of a real audio recording, an excerpt of a Brazilian song with a rich instrumentation was used. Three spectrograms with window sizes of 21.3 ms, 42.6 ms, and 85.3 ms were combined via the SWGM method, using $\beta = 0.5$. These four spectrograms can be seen in Figure 3.16. As previously discussed, the spectrogram with short analysis window provides better resolution for transient information, and the one with long analysis window better defines stationary information. Clearly, the resolution of the combined spectrogram is better than any of the individual STFT-based spectrograms.

Figure 3.17 depicts the combination of spectrograms of the same signal, this time computed using the C$Q$T. As can be seen, the combination of such spectrograms also produces good results.

Although the bin-wise combination methods studied can provide good overall representations, the resulting TFR may degrade important information, under certain conditions. In almost all of them, the peak values are attenuated and, depending

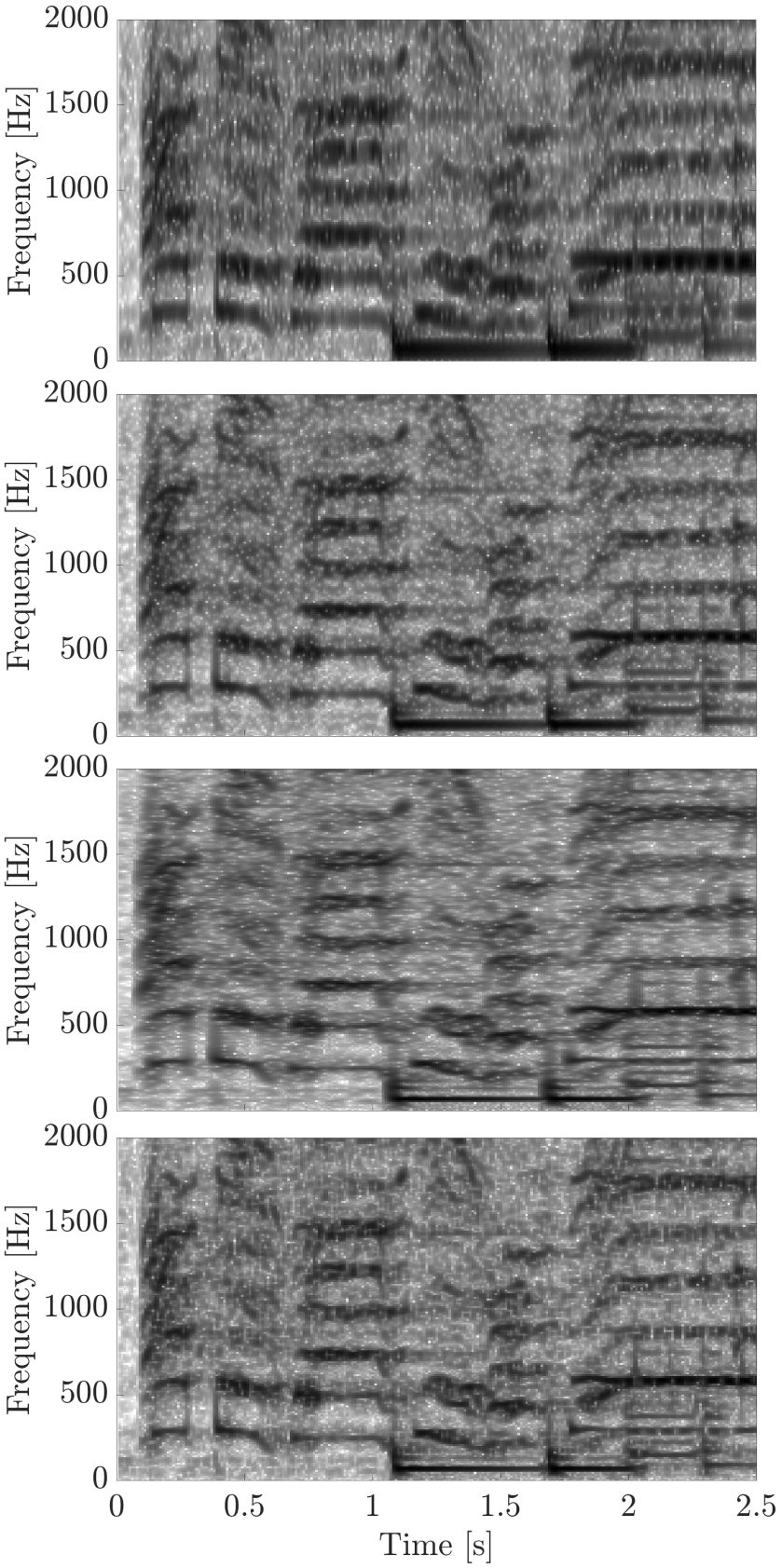



Figure 3.16: Spectrograms of a Brazilian song: STFT (21.3 ms, 42.6 ms, and (85.3 ms)) and SWGM ($\beta = 0.5$), respectively.

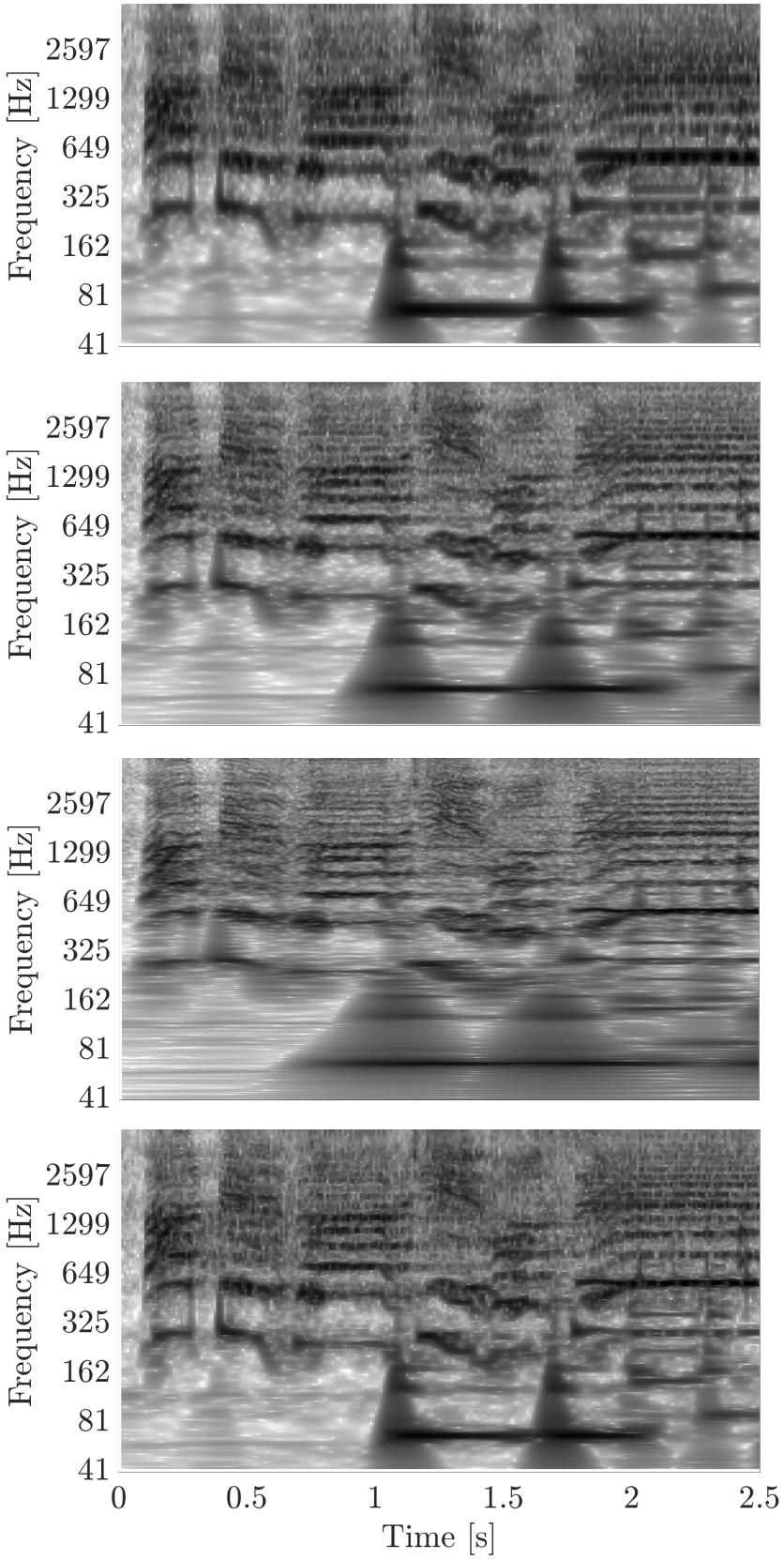



Figure 3.17: Spectrograms of a Brazilian song: C$QT$ spectrograms (12, 24 and 48 bins/octave) and SWGM ($\beta = 0.5$), respectively.

on how poorly concentrated they appear in some TFR provided, this attenuation can be strong, since the best combination methods tend to give results close to the minimum values of the given TFRs. For instance, in a fast frequency chirp, the information may be severely attenuated if a long analysis window is used. As a consequence, a sparser representation is guaranteed at the end, but the energy of some components may show some loss. One advantage of this approach is the relatively low computational burden associated with it.

The combination methods of the following chapters were designed to overcome this problem by considering the context of each bin and assessing which among the representations available better represents the signal for that specific time-frequency location. Such methods provide better overall results, at the expense of requiring much higher computational resources.

# Chapter 4

# Combinations of Spectrograms Based on Local Information

In this chapter, a different approach to the combination of TFRs is studied. Here, the methods use information related to the region around the time-frequency bins under analysis to combine different TFRs. The attempt is to minimize energy spreading using some criterion to select the best representations throughout the time-frequency plane, avoiding the degradation of frequency components inherent to the bin-wise combination methods. Two main methods will be presented: the local sparsity method (LS) [74], along with its smoothed version—the smoothed local sparsity method (SLS) [72], which are contributions of this thesis, and the Lukin-Todd's Method [83].

## 4.1 The Local Sparsity Method

The LS combination method is based on the quantification of the relevance of each time-frequency bin $(k, m)$ of the given TFRs in terms of its local sparsity. It starts with a region selection procedure performed by a two-dimensional window $\mathbf{W}^{\mathrm{S}}$ with $N^{\mathrm{S}}$ elements around bin $(k, m)$ of each $p$-th TFR to be combined. This windowed region, denoted by $\tilde{\mathbf{X}}^{k,m}[p]$, must have odd numbers of lines and of columns, in order to be effectively centered at bin $(k, m)$. The sparsity of each region $\tilde{\mathbf{X}}^{k,m}[p]$ is then measured via the Gini index (or Gini ratio) [121]

$$\mathrm{G}(\tilde{\mathbf{X}}) = 1 - 2\sum_{j=1}^{N^{\mathrm{S}}} \frac{\tilde{x}_j}{||\tilde{\mathbf{x}}||_1} \left( \frac{N^{\mathrm{S}} - j + \frac{1}{2}}{N^{\mathrm{S}}} \right), \tag{4.1}$$

where $\tilde{\mathbf{x}} = [\tilde{x}_1 \ \tilde{x}_2 \ \ldots \ \tilde{x}_{N^{\mathrm{S}}}]$ is a vector containing the elements of the input matrix $\tilde{\mathbf{X}}$ in ascending order of magnitude. The Gini index is a reliable way to assess (or evaluate) the sparsity of a vector, and is a well known measure of inequality in wealth

distribution [121]. Its output is limited to the interval $[0,\ 1-1/N^{\mathrm{S}}]$, where 0 denotes a group of samples with equally distributed energy and $1-1/N^{\mathrm{S}}$ means maximum energy concentration. With infinite elements, the index can range [0,1]. Note that we are dealing with samples that can only assume values greater than or equal to zero. This index satisfies all the desirable criteria of a measure of sparsity [121]:

1. Robin Hood - Transferring energy from a more energetic sample to a less energetic sample, assuming the later will not become more energetic than the former, decreases sparsity.

2. Scaling - Multiplying all the samples by a constant factor does not affect the overall sparsity: sparsity is scale invariant.

3. Rising Tide - Adding a positive constant to all samples decreases sparsity, since it reduces the relative inequality between them.

4. Cloning - Adding samples to the set by cloning it results in the same sparsity of the original set: sparsity is invariant under cloning.

5. Bill Gates - As one sample becomes infinitely more energetic than the others, the sparsity becomes as high as possible.

6. Babies - Adding zero-energy samples to a population with non-zero total energy increases sparsity.

It is worth noting a disadvantage of this method, in which, in comparison to the approach presented in the previous chapter, this measure of a figure-of-merit regarding a group of samples around a given time-frequency bin under analysis requires much higher computational resources.

After computing the local sparsity (i.e. the Gini index given in Equation (4.1) for the regions around bins $(k, m)$) of all representations, a matrix $\hat{\mathbf{P}}$ is constructed to indicate which among the available TFRs exhibits the highest local sparsity, for each time-frequency bin $(k, m)$:

$$\hat{P}_{k,m} = \mathrm{argmax}_p \mathrm{G}(\tilde{\mathbf{X}}^{k,m}[p]). \tag{4.2}$$

This way, an optimum representation in terms of local sparsity could be computed by $X_{k,m}[\hat{P}_{k,m}]$. This is not yet the LS combination method, which also includes a local energy compensation; but this guarantees, under certain conditions, that for any frequency line, the representation which provides the most resonant peak will be chosen to represent that time-frequency region, without interference of the other representations, hence producing a peak with the best possible shape.

Figures 4.1 and 4.2 depict the curves obtained for the sinusoid and impulse test signals, as in the previous chapter, for two spectrograms with different window lengths along with their combinations using the SWGM and the LS$^-$ (i.e. LS without local energy compensation).

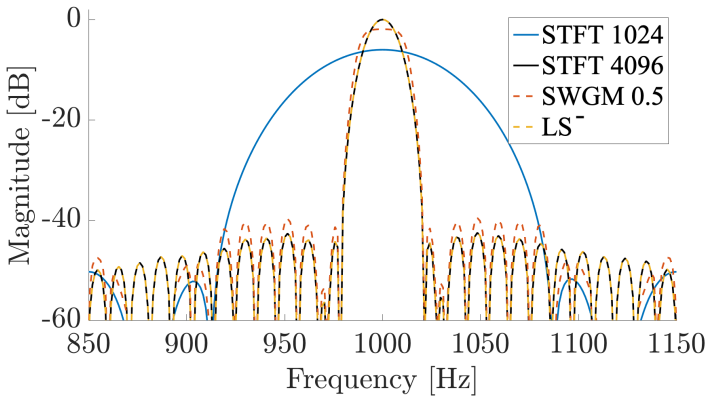



Figure 4.1: Spectrum of a time frame for a sinusoid (1 kHz): STFTs, SWGM and LS$^-$.

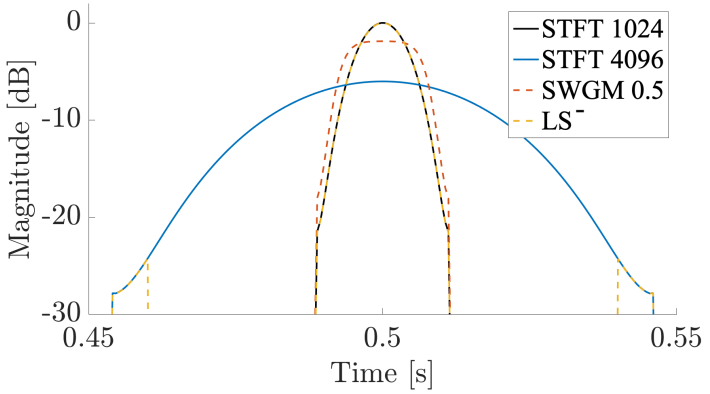



Figure 4.2: Time evolution of a frequency bin for an impulse at 0.5 s for STFTs, SWGM and LS$^-$.

As can be observed in Figures 4.1 and 4.2, the peak obtained by the LS$^-$ curve perfectly follows the sparser curve. In Figure 4.1, the peak-to-sidelobe ratio produced by this combination is nearly 6 dB higher than for the SWGM combination. An interesting detail to be noticed is the effect of the window $\mathbf{W}^{\mathrm{S}}$ when a bin far from the peak is being evaluated: the spectrogram with longer tail is selected in the combination, originating the artifacts that can be observed at the bottom of Figure 4.2.

Nevertheless, this procedure does not yield good results throughout the whole time-frequency map. For instance, consider the beginning of a musical note with a fast transient, which is typically comprised of an initial attack followed by the harmonic components. Right before the attack, representations computed with longer windows (which provide higher frequency resolution) spread frequency information

backwards in time. This can be observed in Figure 4.3, where two spectrograms of a note played in an acoustic guitar are shown. The red dashed line shows the onset instant.

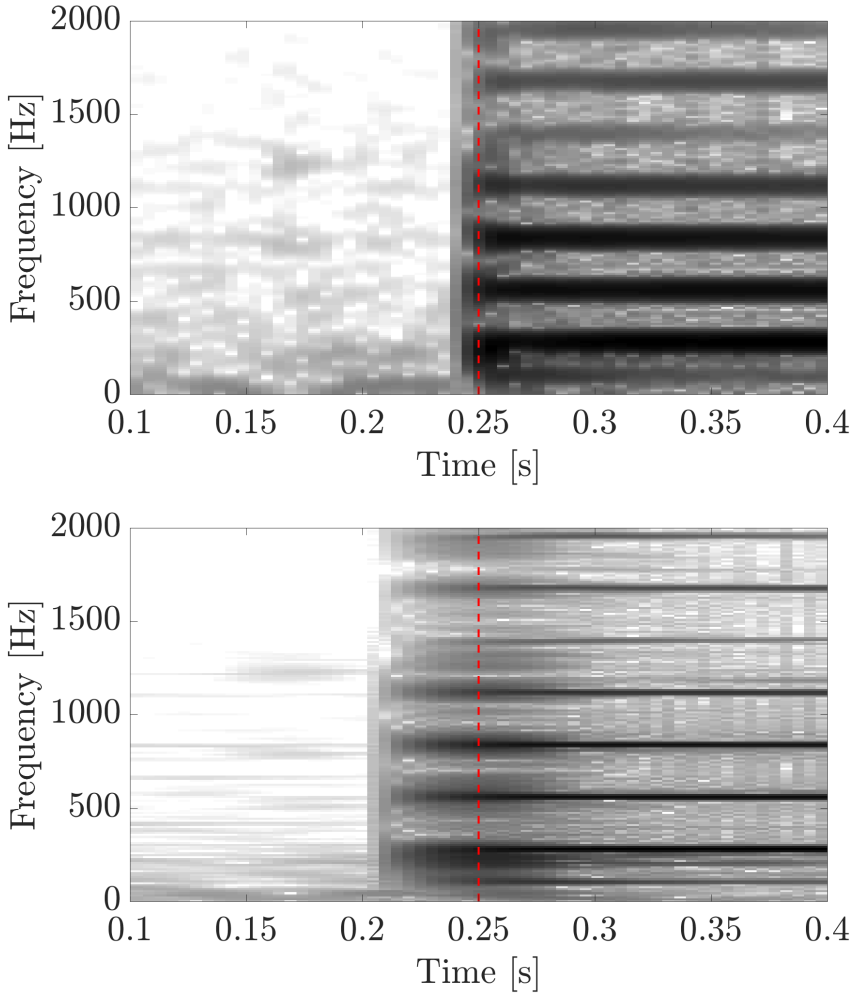



Figure 4.3: Spectrograms of a note played in an acoustic guitar, using short (21.3 ms) and long (85.3 ms) window sizes, respectively.

Comparatively, the regions before the attack, when computed with shorter analysis window (top figure), show much less energy smearing. This can result in a very evenly distributed energy, depending on the background noise, hence probably with lower local sparsity than the same regions in the other spectrogram. As a consequence, undesirable regions from the spectrogram computed with longer window are selected to compose the combined TFR, as can be observed in Figure 4.4. In general, the TFR which best represents a certain region of the signal, in the time-frequency domain, is also the one that best concentrates the signal's energy; but the energy spread, in some cases, may produce a greater local sparsity, as can be seen in this example.

A local energy measure is then used to compensate for such issue. In the prob-

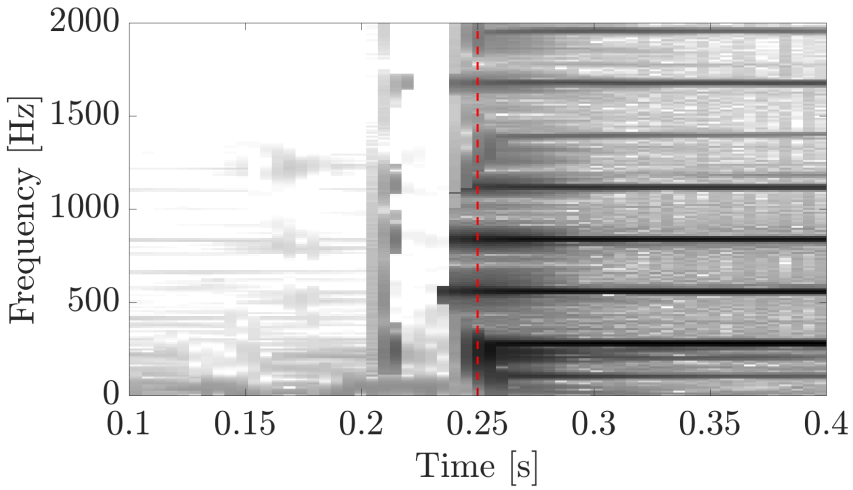



Figure 4.4: LS$^{-}$ combination of a note played in an acoustic guitar.

lematic regions observed, the representation with lower local energy best represent the input signal; meanwhile, after the attack of a note, the local energy is roughly the same for all given TFRs close to frequency lines. Our solution is then to multiply the value of each sample $X_{k,m}[\hat{P}_{k,m}]$ by the local energy ratio $\breve{E}$, which compares the minimum local energy around bin $(k, m)$ among all TFRs with the local energy of $X_{k,m}[\hat{P}_{k,m}]$:

$$\breve{E}_{k,m}[\hat{P}_{k,m}] = \frac{\min_p \sum_{i,j}(\breve{\mathbf{X}}^{k,m}_{i,j}[p])}{\sum_{i,j}(\breve{\mathbf{X}}^{k,m}_{i,j}[\hat{P}_{k,m}])}, \tag{4.3}$$

where $\breve{\mathbf{X}}^{k,m}[p]$ is a matrix comprised of the region around bin $(k, m)$ of the $p$-th TFR under consideration selected by applying a rectangular analysis window $\mathbf{W}^{\mathrm{E}}$. This tends to attenuate the energy of those regions exhibiting energy leakage caused by large windows, while preserving the shape of the frequency components pre-selected by the local sparsity criterion. As a consequence, the application of this principle combines the fine frequency resolution and continuity inherent to the use of long windows with the time precision attained by short windows. In order to avoid numerical problems, the local energy is lower-bounded, e.g. by $10^{-8}$.

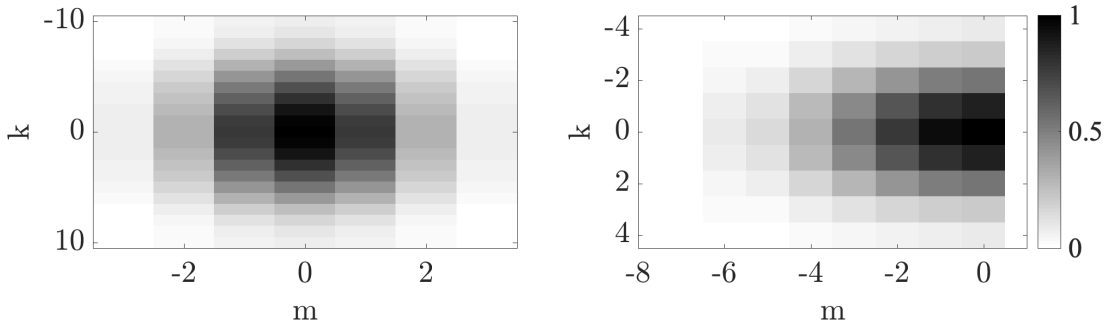
The LS combination $\mathbf{X}^{\mathrm{LS}}$ is then described as

$$X^{\mathrm{LS}}_{k,m} = c^{\mathrm{LS}} X_{k,m}[\hat{P}_{k,m}]\breve{E}_{k,m,}[\hat{P}_{k,m}], \tag{4.4}$$

where $c^{\mathrm{LS}}$ is an energy-matching constant.

The window $\mathbf{W}^{\mathrm{S}}$, responsible for the local sparsity computation, is designed as a two-dimensional Hamming window. For computing the local energy in order to provide the best results at initial transients, an asymmetric window $\mathbf{W}^{\mathrm{E}}$ is used. It is also composed of a two-dimensional Hamming window, but with its right half zeroed. This way, only the energy before the given time-frequency bin is taken into

account. Figure 4.5 depicts both windows.




(a) $\mathbf{W}^{\mathrm{S}}$. (b) $\mathbf{W}^{\mathrm{E}}$.

Figure 4.5: Analysis windows used for computing the local sparsity and the local energy ratio.

The dimensions of $\mathbf{W}^{\mathrm{S}}$ and $\mathbf{W}^{\mathrm{E}}$ are related to the resolutions of the assembled TFRs. Considering that the $P$ TFRs to be combined have been computed with analysis windows of lengths $N_1, N_2, ..., N_P$ in ascending order, the frequency width for $\mathbf{W}^{\mathrm{S}}$ can be sized, e.g., as $10N_P/N_1$. This leaves this window enough room for including the frequency components present in all representations. For $\mathbf{W}^{\mathrm{E}}$, using a smaller frequency width, e.g. $4N_P/N_1$, usually produces good results. As to the time span, something between $N_1$ and $N_2$ can be used, e.g. a length of $\approx 50\,\mathrm{ms}$.

Figure 4.6 depicts the resulting combination, now using the local energy compensation. As can be observed, the energy smearing was successfully corrected, being attenuated to values compatible with the noise floor ($\approx$ -70 dB). As for the sinusoid and impulse signals, the results are roughly the same, and can be seen in Figures 4.7 and 4.8. The major difference that can be observed is the correction of the artifact before the impulse and a slight distortion in the impulse signal, which happens due to the asymmetric nature of the local energy analysis window.

One problem related to this type of method is the interference of harmonic components close to each other in situations where the different representations to be combined show different resolutions for such components, i.e. there is no TFR to properly represent all the harmonic components in such region. Since one representation must be chosen due to its higher local sparsity, necessarily one or more harmonic components will not be well represented in such regions of interference. This effect is attenuated when using the modified version of the LS method, namely, the smoothed local sparsity method.

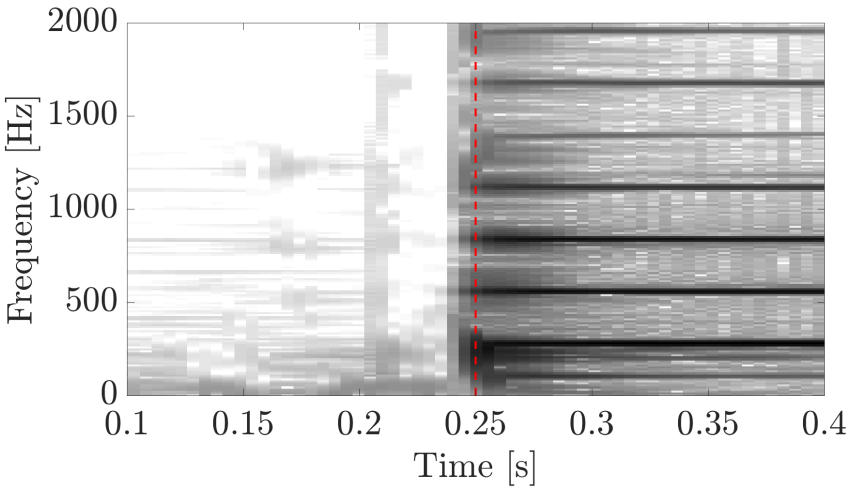



Figure 4.6: LS combination of a note played in an acoustic guitar.

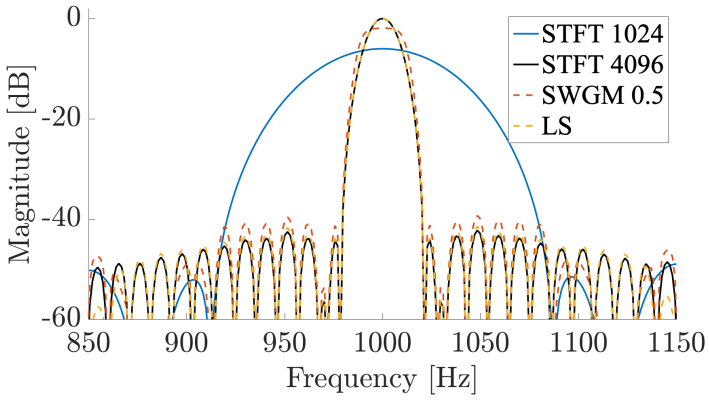



Figure 4.7: Spectrum of a time frame for a sinusoid (1 kHz): STFTs, SWGM and LS.

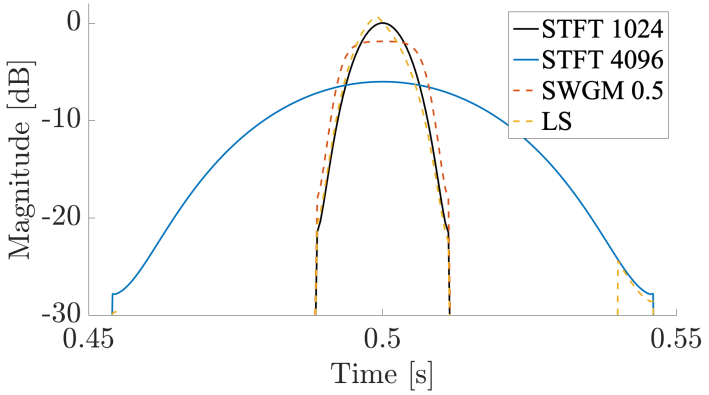



Figure 4.8: Time evolution of a frequency bin for an impulse at 0.5 s for STFTs, SWGM and LS.

## 4.2 The Smoothed Local Sparsity Method

The Smoothed Local Sparsity method (SLS) is a modification of the LS method intended to achieve smoother combinations of TFRs as well as mitigate some artifacts

by providing soft transitions between different resolutions.

The modification consists in weighting each sample according to its sparsity and combining all samples in dimension $p$ for each bin $(k, m)$, instead of generating an index-matrix of optimal samples $\hat{\mathbf{P}}$. In order to generate such weights, a tensor $\underline{\mathbf{S}}$ is constructed with elements

$$S_{k,m}[p] = \left( \frac{\mathrm{G}(\tilde{\mathbf{X}}^{k,m}[p])}{\prod_{l \neq p} \mathrm{G}(\tilde{\mathbf{X}}^{k,m}[l])} \right)^{\zeta}. \tag{4.5}$$

Note that $\underline{\mathbf{S}}$ is essentially determined by ratios comparing the sparsity measure of bin $(k, m)$ in each TFR $p$ with the product of all remaining sparsity measures for the same bin, which necessarily achieves its highest value for $p$ related to the TFR with highest local sparsity. Since the difference between the sparsity measures of different TFRs is usually small, an exponent $\zeta$ is used to amplify them.

The final representation $\mathbf{X}^{\mathrm{SLS}}$ is then defined as

$$X^{\mathrm{SLS}}_{k,m} = c^{\mathrm{SLS}} \frac{\sum_p X_{k,m}[p] S_{k,m}[p] \breve{E}_{k,m}[p]}{\sum_p S_{k,m}[p]}, \tag{4.6}$$

where $c^{\mathrm{SLS}}$ matches the overall energy. This combination performs a mean of the samples through dimension $p$ (which indexes the TFRs) weighted according to their local sparsities, for each bin $(k, m)$. This provides a smoother final result, since there are no more hard switching between different TFRs. High values for $\zeta$, e.g. $\zeta \geq 50$, are preferable choices, as they tend to give high priority to the best-resolution TFR, but still allow for smooth transitions. There is no optimum value for such parameter. By using $\zeta = \infty$, this solution matches the original local sparsity combination method, since it is equivalent to applying weight 1 to the locally-sparser samples, and 0 to the others.

In Figures 4.9 and 4.10 one can see the comparison between the SLS and the LS methods for the test signals. For the sinusoid, the results are nearly indistinguishable, and match the representation which best suits the input signal; for the impulse, the results are similar, with a difference in gain. Also, in the right part of the image one can see that the subtle change in the choice of the representation that occurred in the LS combination, causing that artifact, is no more present in the SLS combination curve. This phenomenon happens when, in some point of transition, a representation starts to exhibit the highest local sparsity. This is not possible in the SLS combination due to the continuous nature of the combination: all representations are taken into account and have some influence in the final result. Note that the fact that the shape of these peaks is virtually preserved is the best scenario possible, since our goal is to just remove artifacts related to abrupt transitions.

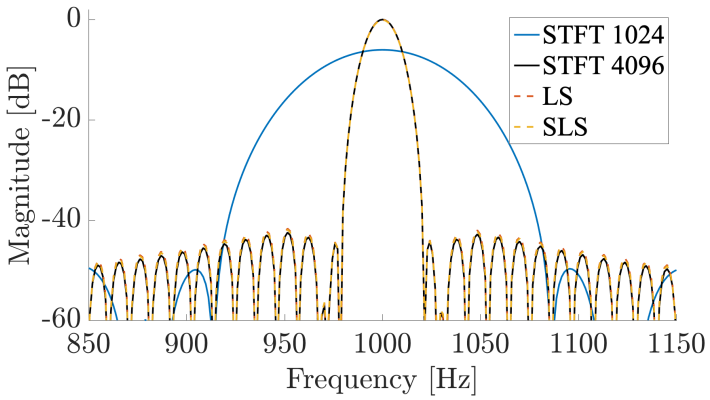



Figure 4.9: Spectrum of a time frame for a sinusoid (1 kHz): STFTs, LS and SLS.

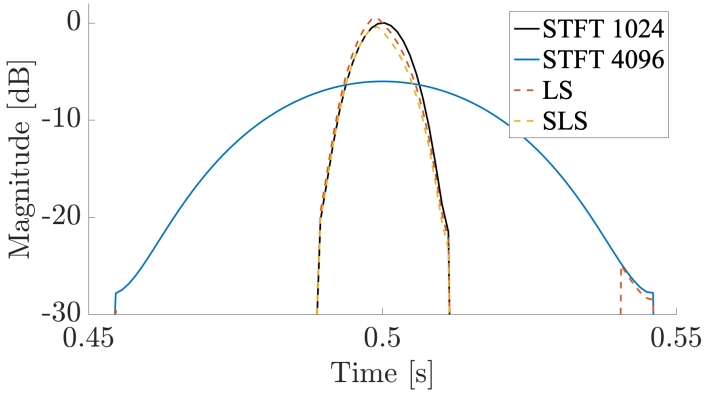



Figure 4.10: Time evolution of a frequency bin for an impulse at 0.5 s for STFTs, LS and SLS.

In order to compare the combination of spectrograms of a real signal, Figure 4.11 depicts the LS and the SLS combinations, respectively, obtained from three spectrograms computed with short (21.3 ms), medium (42.6 ms), and long (85.3 ms) window sizes. The SLS combination was configured with $\zeta = 70$. The input signal contains percussive instruments and a *cuíca*, a Brazilian instrument which can produce nearly harmonic sounds with fast frequency variation. As desired, the SLS combination yields a smoother representation, with less artifacts, which are created by hard switching between spectrograms. Inside the highlighted regions one can see the presence of vertical artifacts, which are diminished in the SLS spectrogram. Such subtle peaks in energy could lead, for instance, to false onset detections.

This method does not necessarily yield better overall representations in terms of sparsity, since the LS method optimizes the local sparsity of each time-frequency bin, but the resulting TFR can be a better input to several methods that rely on TFRs to extract information about the audio signal.

One issue inherent to both LS and SLS methods is that, although they are not limited to combine a specific type of spectrogram, they do not yield good results for

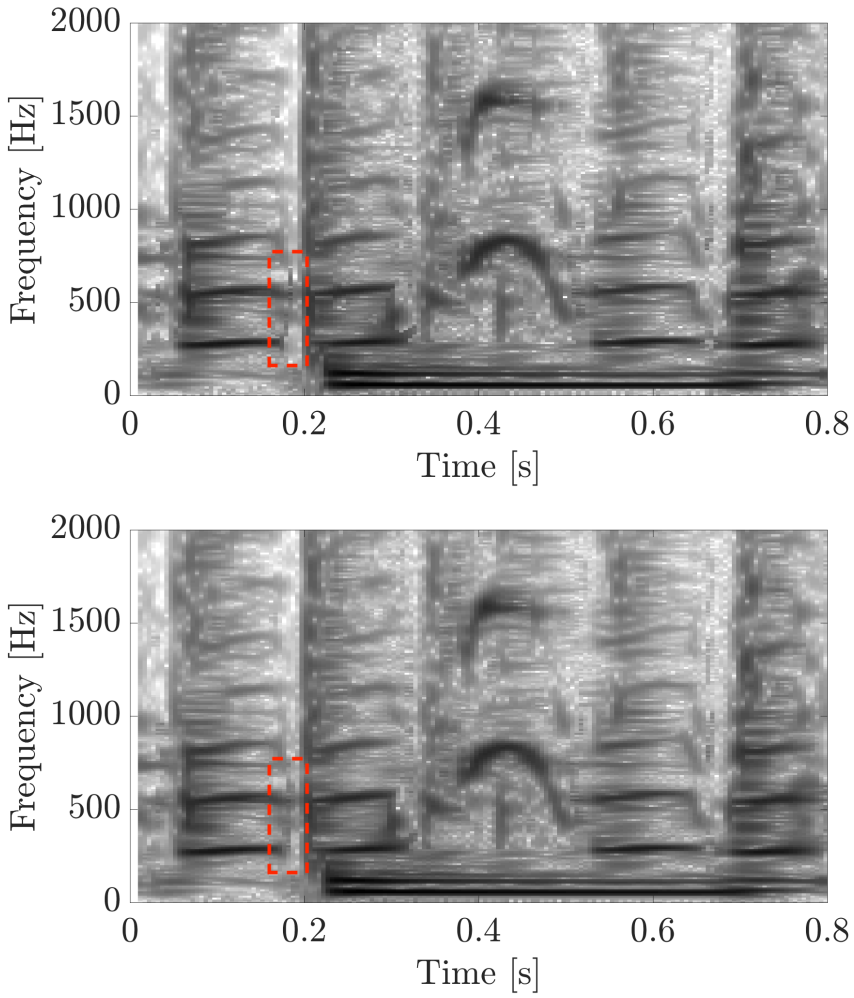



Figure 4.11: Combined spectrogram using the LS and SLS methods, respectively.

C$Q$T-based spectrograms, since the 2-dimensional analysis windows have fixed size, while the frequency resolution is not linear in this case. These methods should be modified to have variable window sizes, in such a way to fit the input representations.

## 4.3 The Lukin-Todd's Method

In [83], a combination method, to which we will refer as Lukin-Todd's method (LT), based on an energy smearing criterion was devised. There, the best time-frequency bin of each spectrogram, in Mel scale[1], is chosen by measuring how smeared the energy inside a given frame is, also penalizing the amount of energy present. Since this procedure also produces harsh transitions between different spectrograms, this energy smearing measure acts as a weight in a soft combination procedure, to overcome this effect.

[1]The Mel scale was defined, based on some experiments, to represent the way humans perceive linear pitch variation [122].

Differently to what is performed in the LS method, the LT combination starts by selecting the region $\tilde{\mathbf{X}}^{k,m}[p]$ around bin $(k, m)$ of the $p$-th TFR to be combined, without applying an edge-softener window. The energy smearing function is then measured by the function [83]

$$\mathrm{L}(\tilde{\mathbf{X}}) = \frac{\sum_{j=1}^{N^{\mathrm{S}}} j\tilde{x}_j}{\sqrt{\sum_{j=1}^{N^{\mathrm{S}}} \tilde{x}_j} + \epsilon}, \tag{4.7}$$

where $\tilde{\mathbf{x}} = [\tilde{x}_1\ \tilde{x}_2\ \ldots\ \tilde{x}_{N^{\mathrm{S}}}]$ is a vector containing the elements of the input matrix $\tilde{\mathbf{X}}$ in descending order and $\epsilon$ is a small constant used to prevent divisions by zero. The numerator of the ratio evaluates the first moment of the distribution of the energy magnitudes, while the denominator normalizes the function by the square root of the total energy of that region. Note that, since the square root is applied, this function does not totally compensates for the total energy, thus indicating higher smearing for more energetic regions. This helps reducing energy smearing, in a similar way to what is performed by the local energy compensation in the LS and SLS methods.

After that, the final representation is obtained by computing a weighted mean, using the smearing measure:

$$X_{k,m}^{\mathrm{LT}} = c^{\mathrm{LT}} \frac{\sum_p X_{k,m}[p] L_{k,m}[p]^{-\eta}}{\sum_p L_{k,m}[p]^{-\eta}}, \tag{4.8}$$

where $L_{k,m}[p] = \mathrm{L}(\tilde{\mathbf{X}}^{k,m}[p])$ and $\eta$ is a constant used to exaggerate the differences in energy smearing, set to 8 in [83].

In order to compare the LT method with the others, it was implemented as a combination of standard spectrograms, instead of Mel-scale spectrograms, which are based on the former. This does not affect the quality of the combination criterion and allows one to better observe the differences between the methods. Figures 4.12 and 4.13 depict the resulting curves using the same simple test signals for the LT and SLS methods. As can be seen, for these signals, the LT method results in nearly perfect combination, matching the best representation possible for each case.

However, the energy smearing at the attack of a note is not well treated by this method, as can be observed in Figure 4.14. Here, the attack is not so well defined as in the case of the SLS method, which is depicted in Figure 4.15. This problem, which is specifically solved by using a special asymmetric window to normalize the local energy in the LS and SLS methods, cannot be addressed here with the same precision; the symmetric nature of the analysis window results in such less precise attacks.

Figure 4.16 depicts the same signal with percussive instruments used before, now combined using the LT method. The result is a very smooth representation, with

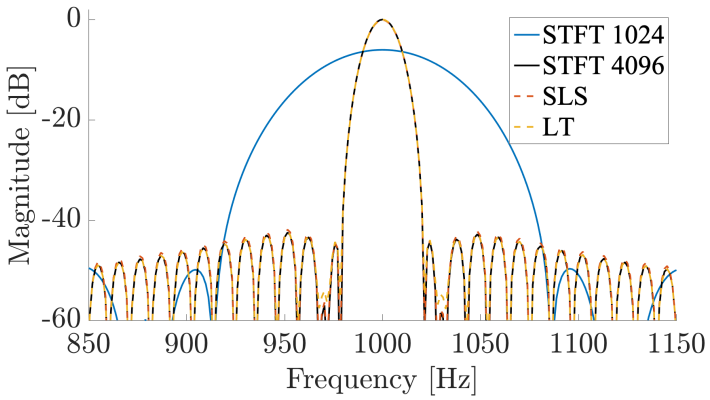



Figure 4.12: Spectrum of a time frame for a sinusoid (1 kHz): STFTs, SLS and LT.

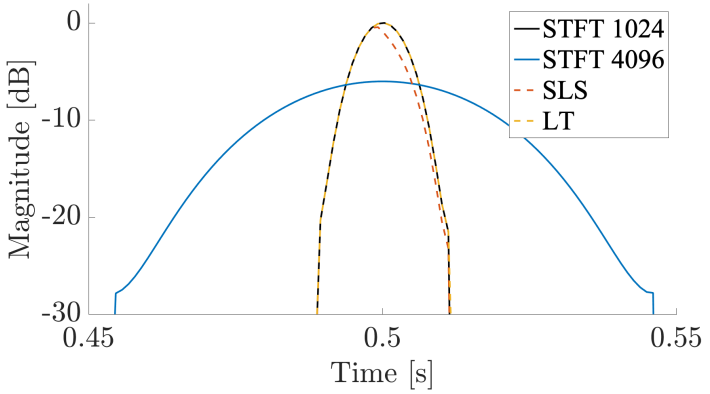



Figure 4.13: Time evolution of a frequency bin for an impulse at 0.5 s for STFTs, SLS and LT.

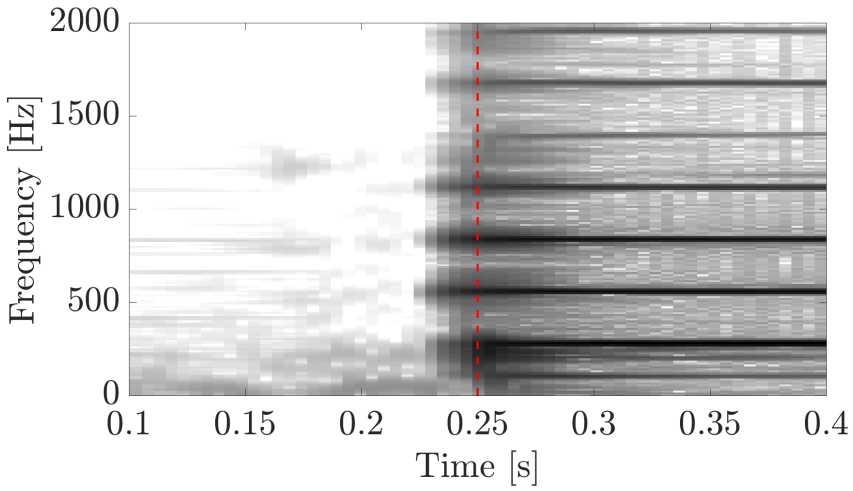



Figure 4.14: LT combination of a note played in an acoustic guitar.

the attacks less well defined, as mentioned. A small region is selected in order to compare one more time the energy smearing in attacks, this time of a percussive instrument. In Figure 4.17 one can see this selected region for the SLS and the LT

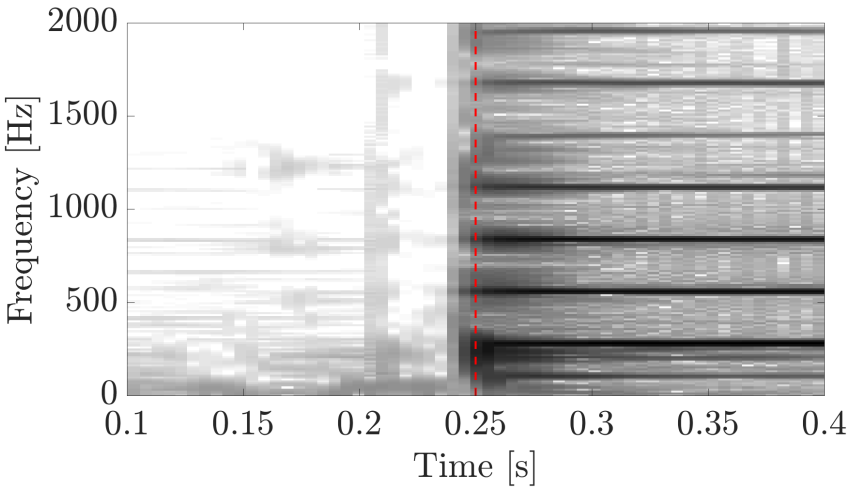



Figure 4.15: SLS combination of a note played in an acoustic guitar.

combinations, respectively. As shown before, this attack is much less smeared in the SLS combination than in the LT one.

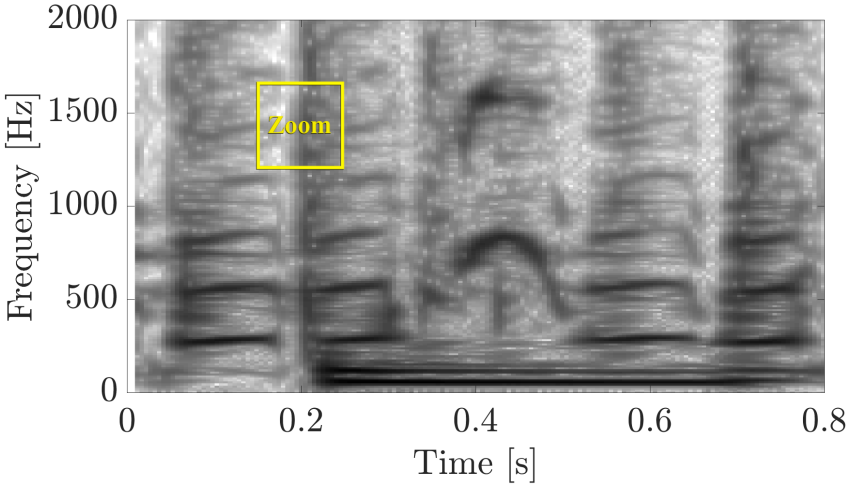



Figure 4.16: Combined spectrogram using the LT method.

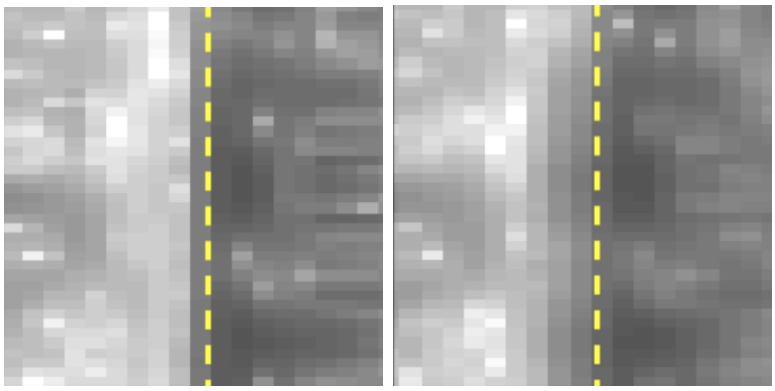


Figure 4.17: Zoom of the combined spectrogram using the SLS and the LT method, respectively. The attack is much better defined in the SLS combination.

## 4.4 Concluding Remarks

In this chapter, the two methods presented follow basically the same strategy: use local figures-of-merit to guide combinations of representations available. This approach has the advantage of selecting the best representation for each region of the time-frequency plane at the cost of having high computational complexity, since a group of samples must be sorted and processed for the computation of the local measures for each time-frequency bin. It may be possible to use different figures-of-merit that require less operations, for instance by subsampling the regions in such a way that similar results are achieved.

Another point that may be useful to stress is that representations that use different analysis window lengths provide frequency lines having different peak values, since they have the same energy and larger windows will present higher energy concentration in frequency.[2] Therefore, if the absolute peak value is a useful feature for a given application, this kind of combination procedure may not provide consistent results, since there are regions (especially where there are frequency slopes) where different resolutions will be used interchangeably.

In the next chapter, another approach will be presented: an image processing technique will be used to indicate the direction of the frequency lines present in a spectrogram. This tool provides useful information, which makes possible a low-cost computation of FChT-based spectrograms and a combination procedure based on the directions estimated. This last method provides excellent results and requires much less computational power, compared to the SLS and LT methods.

[2]This example is considering stationary signals, only to illustrate the principle discussed.

# Chapter 5

# Combinations Based on Image Analysis

In this chapter, two methods [72, 73] for generating high-definition TFRs are studied. Both make use of an image processing technique, namely, the structure tensor [54, 111, 112], which is used to compute the frequency slope parameter $\alpha$ of the FChT—as first proposed in [110]. This is performed by analyzing a spectrogram of the given signal and identifying the direction of the frequency lines, which are then processed in order to provide the parameter $\alpha$ needed for the computation of the FChT. Since an exhaustive search is no longer needed, this method circumvents the high computational complexity required to estimate the correct values for parameter $\alpha$. This chapter is strongly based on [72, 73], which are original contributions of this thesis.

## 5.1 The Structure Tensor

### 5.1.1 Computation of the Structure Tensor

Initially, $\hat{\mathbf{X}}$, which is a compressed and limited in dynamic range version of the spectrogram $\mathbf{X}$ of the input signal is built by modifying each bin as bellow:

$$\hat{X}_{k,m} = \max\left\{1 + \frac{10}{R}\log_{10}\frac{X_{k,m}^2}{X_{\max}}, 0\right\}, \tag{5.1}$$

were

$$X_{\max} = \max_{k',m'} X_{k',m'}^2 \tag{5.2}$$

and $R$ is the desired dynamic range. This procedure[1] produces a spectrogram with values limited between $[0, 1]$ with a dynamic range of $R\,\mathrm{dB}$ with regards to the original signal.

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

$$\hat{\mathbf{X}}^{\mathrm{m}} = \hat{\mathbf{X}} * \mathbf{D}, \tag{5.3}$$

$$\hat{\mathbf{X}}^{\mathrm{k}} = \hat{\mathbf{X}} * \mathbf{D}^{\mathrm{T}}, \tag{5.4}$$

where $\mathbf{D}$ is a discrete differentiation operator, more specifically the Sobel-Feldman operator [123]

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \tag{5.5}$$

and $*$ denotes the 2-dimensional convolution.

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

$$\hat{\mathbf{X}}^{\mathrm{mm}} = [\hat{\mathbf{X}}^{\mathrm{m}} \odot \hat{\mathbf{X}}^{\mathrm{m}}] * \mathbf{G} \tag{5.6}$$

$$\hat{\mathbf{X}}^{\mathrm{mk}} = \hat{\mathbf{X}}^{\mathrm{km}} = [\hat{\mathbf{X}}^{\mathrm{k}} \odot \hat{\mathbf{X}}^{\mathrm{m}}] * \mathbf{G} \tag{5.7}$$

$$\hat{\mathbf{X}}^{\mathrm{kk}} = [\hat{\mathbf{X}}^{\mathrm{k}} \odot \hat{\mathbf{X}}^{\mathrm{k}}] * \mathbf{G}, \tag{5.8}$$

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 $\sigma_{\mathrm{m}}$ and $\sigma_{\mathrm{k}}$ in time- and frequency-index directions, respectively, intended to soften the frequency lines and provide more continuous transitions.[2] Matrix $\hat{\mathbf{X}}^{\mathrm{mm}}$ contains information related to temporal (horizontal) variation in the image, $\hat{\mathbf{X}}^{\mathrm{kk}}$ contains information about frequency (vertical) variation, and $\hat{\mathbf{X}}^{\mathrm{mk}}$ and $\hat{\mathbf{X}}^{\mathrm{km}}$ convey both. Here, all convolutions are applied preserving the original shape of the spectrogram, hence the initial and final samples generated in the convolution are neglected.

Now, each time-frequency bin $(k, m)$ has a group of four other values related to it: $\hat{X}^{\mathrm{mm}}_{k,m}$, $\hat{X}^{\mathrm{kk}}_{k,m}$, $\hat{X}^{\mathrm{mk}}_{k,m}$, and $\hat{X}^{\mathrm{km}}_{k,m}$. Together, such bins form a structure tensor element

[1] This approach is different from what is performed in the literature [54], in which the full spectrogram is used.

[2] Also, as will be discussed in Section 5.3, this smoothing filter is responsible for providing anisotropy measures which indicate the linearity of the region around the bin under scrutiny.

$\mathbf{T}_{k,m}$, which is a $2 \times 2$ symmetric and positive semi-definite matrix:

$$\mathbf{T}_{k,m} = \begin{bmatrix} \hat{X}_{k,m}^{\mathrm{mm}} & \hat{X}_{k,m}^{\mathrm{km}} \\ \hat{X}_{k,m}^{\mathrm{mk}} & \hat{X}_{k,m}^{\mathrm{kk}} \end{bmatrix}. \tag{5.9}$$

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 an anisotropy measure, which indicates how focused on a certain direction is the edge near a given bin, as shown in the following section.

### 5.1.2 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 $\mathbf{v}_{k,m} = [v_{k,m}^1, v_{k,m}^2]^{\mathrm{T}}$ 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

$$\theta_{k,m} = \arctan\left(\frac{v_{k,m}^2}{v_{k,m}^1}\right) \in [-\pi/2, \pi/2], \tag{5.10}$$

with $v_{k,m}^1$ being the horizontal (temporal) component and $v_{k,m}^2$ 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

$$C_{k,m} = \begin{cases} \left(\dfrac{\mu_{k,m} - \lambda_{k,m}}{\mu_{k,m} + \lambda_{k,m}}\right)^2, & \hat{X}_{k,m} > 0 \\ 0, & \hat{X}_{k,m} = 0. \end{cases} \tag{5.11}$$

The anisotropy measure will then be related to how directional is the given neighbourhood of bin $(k, m)$ and will only be computed for time-frequency bins within the dynamic range $R$. A straight frequency line will provide maximum difference between the eigenvalues, and hence $C_{k,m}$ will be close to 1, while curved lines yield smaller values of $C$. In order to have a softer transition along consecutive values of $C$, especially when transitioning to regions of $C = 0$, a small 2-D Gaussian smoothing

filter having standard deviations of 0.5 samples is used.

In [54, 72, 73, 110], a threshold is used to limit the range of what should be considered anisotropic by the restriction $\mu_{k,m} + \lambda_{k,m} \geq \varepsilon$, in order to increase robustness against background noise, with the normal spectrogram (in dB) being used instead of $\hat{\mathbf{X}}$. Figure 5.1 depicts, as an example, a region of a TFR where there exists a frequency line, with the structure tensor vectors depicted in blue dashes and whose magnitude is related to their $C$, for the two different approaches mentioned, i.e. using the standard spectrogram and the anisotropy restriction compared to the spectrogram limited in dynamic range. As one can observe, in our new approach (shown on the right), the anisotropy tends to be more focused on the frequency lines. This characteristic will be even more useful in the MRFCI method, to be studied in Section 5.3, where the combination procedure depends on the anisotropy.

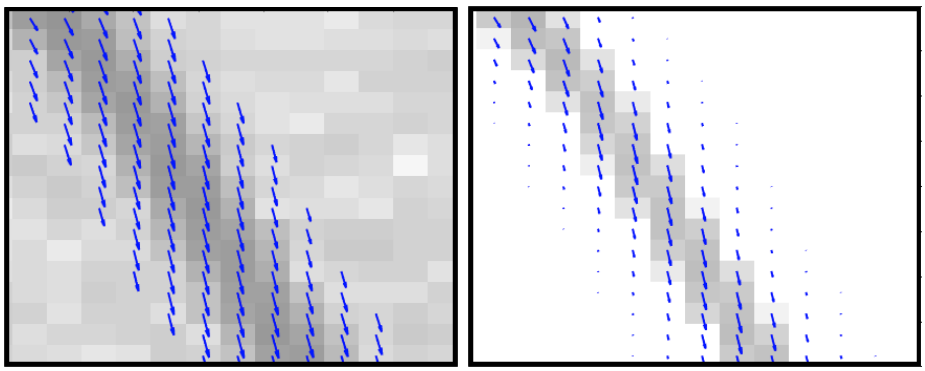


Figure 5.1: Structure tensor computed for a region of a TFR comprising a frequency line: standard approach and the proposed modified version, respectively.

Although in this new approach there is still one parameter to be set, which is the dynamic range $R$ of the spectrogram, we believe this is a more easily interpretable and more controllable parameter to set. Another advantage is that our restriction is much more insensitive to parameters such as analysis window length, hop size and $\mathbf{G}$ dimensions, while the original restriction directly depends on all these factors. As a consequence, the modified version is expected to provide much more predictable and reliable results.

### 5.1.3 Computation of $\alpha$

Since the angles $\theta$ 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 $\alpha$, which is related to the analog time-frequency domain; therefore, a transformation must be performed in order to compute the set of $\alpha$'s from a set of $\theta$'s.

Let the angle $\vartheta$ be the continuous time-frequency domain version of the angle $\theta$, 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 $\nu^2 = v^2 F_\mathrm{s}/N$, where $F_\mathrm{s}$ is the sampling rate, $h$ is the hop-size of the STFT, and $N$ is the number of samples used in the Fourier transform. Figure 5.2 depicts the geometrical relation between $\vartheta$ and $\alpha$.

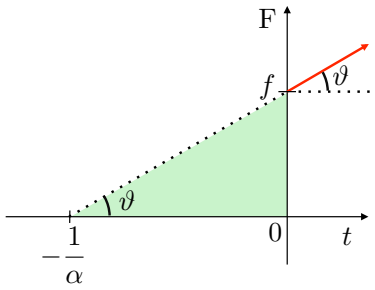



Figure 5.2: Geometrical relation between the orientation angle $\vartheta$ and variable $\alpha$, in the continuous time-frequency domain.

By analyzing the triangle highlighted in green in Figure 5.2, one can verify that

$$\tan\vartheta = f\alpha, \tag{5.12}$$

which using the aforementioned conversions can be written as

$$\tan\vartheta = \frac{\nu^2}{\nu^1} = \frac{v^2 F_\mathrm{s}/N}{v^1 h/F_\mathrm{s}} = \tan\theta \frac{F_\mathrm{s}^2}{Nh}. \tag{5.13}$$

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

$$\alpha_{k,m} = \tan\theta_{k,m} \frac{F_\mathrm{s}}{hk} = \frac{v^2_{k,m} F_\mathrm{s}}{v^1_{k,m} hk}. \tag{5.14}$$

Therefore, by performing this conversion one has an $\alpha_{k,m}$ related to each $\theta_{k,m}$ estimated from the spectrogram. This conversion will be used in both methods presented in the following.

## 5.2 The Frame-Based Method for Estimation of Main Directions

The frame-based method for estimation of main directions (FEMD) [72] aims to create a high-resolution TFR by combining different FChT-based spectrograms generated using the structure tensor to estimate multiple simultaneous $\alpha$'s. This means that it makes possible the use of the FChT to create a high-resolution representation of simultaneous sources, which is not possible using a single FChT-based spectrogram. The inclusion of normal spectrograms (equivalent to $\alpha = 0$) is suitable for

the description of sources with stable pitch. The combination of the different representations can then be performed by any of the combination procedures studied.

### 5.2.1 Estimating Multiple $\alpha$'s

Considering that $Z$ simultaneous values of $\alpha$ are required to properly describe the whole TFR, one should be able to detect $Z$ different $\alpha$'s for each time frame.[3] Note that this does not mean that there are necessarily $Z$ sound sources present in the audio signal, since many sources may just require $\alpha = 0$ to be properly described, or the different sources may not appear simultaneously.

In order to detect the most prominent $\alpha$'s, a distribution is estimated for each time frame $m$. A one-dimensional Gaussian function is generated for each frequency bin $k$, centered at $\alpha_{k,m}$, with fixed standard deviation $\sigma$, and weight $C_{k,m}$; then, all Gaussians generated for that certain frame $m$ are summed up. As a consequence, the most anisotropic vectors, which are the most relevant ones, have higher influence in the distribution produced. It is worth mentioning once again that, for harmonic sources, every harmonic component share the same $\alpha$, hence a single source will provide a single peak in the distribution.

After that, the values of $\alpha$ corresponding to the highest peaks in the distribution are selected as the best candidates, one for each time frame $m$ and source $z$, to form a matrix $\mathbf{A} \in \mathbb{R}^{Z \times M}$. For now, the algorithm receives as input from the user the total number of $\alpha$ estimates $Z$, but some procedure could be performed in order to automatically estimate it.

An example of distribution obtained for one frame of a synthetic signal is given in Fig. 5.3, in blue. The signal consisted of two synthetic vibratos with different fundamental frequencies and frequency modulation rates. In this particular frame, both fundamental frequencies increased with estimated chirp rates $\alpha_1 = 0.38$ and $\alpha_2 = 1.06$, respectively, as indicated by the red crosses.

Since the distribution is comprised of a sum of Gaussian functions, relevant values of $\alpha$ accumulated close to each other may or may not result in a single peak in the distribution, according to the value of $\sigma$. This parameter impacts the quality of the overall result, since too large a $\sigma$ unduly integrates independent peaks into a single $\alpha$, while too small a $\sigma$ improperly discriminates similar values of $\alpha$. Moreover, since low peaks will probably lead to spurious results, a threshold relative to the highest peak is adopted to limit the range within which valid peaks are allowed to be found. When searching for the $Z$ candidates in the $\alpha$ distribution, if the number $Z'$ of peaks found within the range defined by the threshold is smaller than $Z$, the $\alpha$ relative to the highest peak is attributed to the missing $Z - Z'$ candidates.

[3] Note that here we are referring to $Z$ FChT-based representations; since the number of representations $P$ to be combined will also include common spectrograms, $P > Z$.

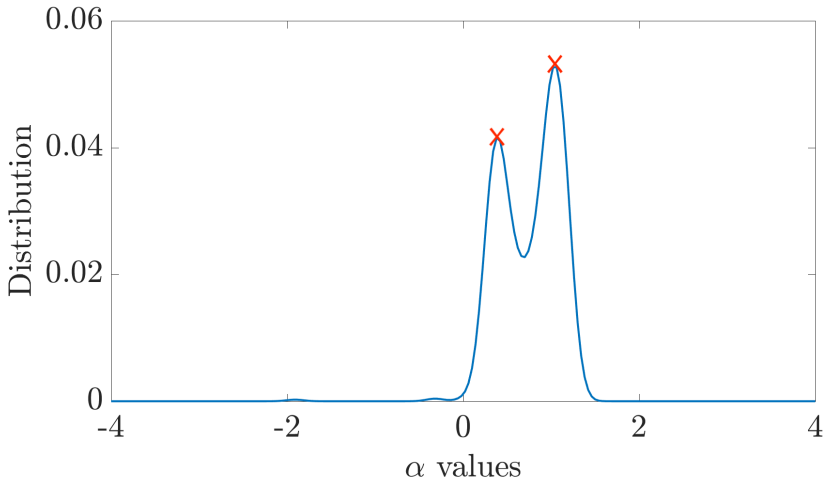



Figure 5.3: Estimated distribution for one frame of a synthetic signal (in blue) and estimated $\alpha$ (in red).

Then, for each frame $m$ and index $z$, an instance of the FChT will be generated; at the end, they will all be combined to form a single TFR. The peaks found in the $\alpha$ distribution are initially estimated and ordered by their values, which means that there is not necessarily a correspondence between the $z$ indexes and the sound sources throughout the time frames $m$.

The $\alpha$ estimation procedure described usually produces some outliers, which may yield spurious discontinuities in the FChT spectrograms and, therefore, probably in the final TFR. Hence, some sort of smoothing procedure is advisable.

### 5.2.2 Filtering the Estimated $\alpha$'s

Aiming at producing a smoother evolution of $\alpha$ along time dimension $m$, i.e. along the rows of $\mathbf{A}$, a simple filtering procedure is performed. As mentioned, the estimates are ordered, for each time frame $m$, in such a way that the elements $A_{z,m}$ decrease with $z$; this way, if $\alpha$ estimates relative to different sources cross each other, their indexes $z$ will be exchanged after the cross-point.

After this sorting, two filtering stages are applied through time dimension $m$, for each index $z$: an outlier removal by means of a median filter, and a final smoothing filtering using a Hann window. With a sufficiently small smoothing filter, the deviations produced at the different $\alpha$ cross-points are small enough to be neglected. This is corroborated by the fact that the fine resolution produced by the FChT has a relatively small sensitivity to the values of $\alpha$ used. Empirically, both median and smoothing filters with length of 5 samples have shown to yield good results.

This procedure produces a set $\bar{\mathbf{A}} \in \mathbb{R}^{Z \times M}$, comprised of $\alpha$'s which evolve smoothly in time. Despite of the resolution adopted for the spectrogram used for the structure tensor, one can generate FChTs with any desired resolution, as long as

synchronism is guaranteed. One could also adopt another hop size, but this would require resampling $\bar{\mathbf{A}}$. Finally, $Z$ FChT-based spectrograms are computed for each $\bar{A}_{z,m}$, generating a TFR tensor $\underline{\mathbf{X}}^{\mathrm{FChT}} \in \mathbb{R}^{K \times M \times Z}$.

Figure 5.4 shows an example of $\alpha$ estimation where a mixture of two violins was analyzed. Two TFRs are depicted: (i) using the filtering stage, and (ii) not using any filtering in the $\alpha$ estimation procedure. Note that the TFR computed from smoothed $\alpha$ estimates does not exhibit the vertical artifacts produced by abrupt $\alpha$ transitions observable in the bottom TFR. The third plot depicts the estimates obtained: blue crosses show the original $\mathbf{A}$, while the filtered $\bar{\mathbf{A}}$ is depicted in red. Note that the outliers present in $\mathbf{A}$ are removed and much smoother estimates are obtained. As can be seen, for this example, two $\alpha$'s were estimated, hence two different instances of FChT-based spectrograms were combined using the SLS method.

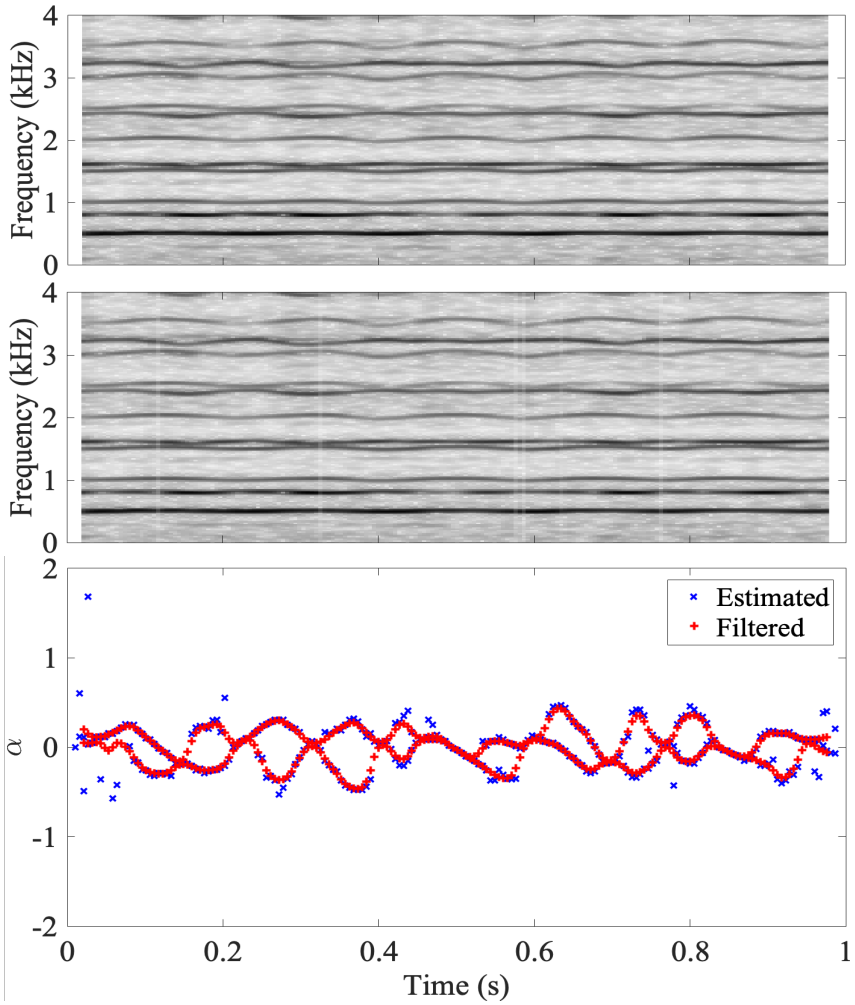



Figure 5.4: From top to bottom: two TFRs generated by the proposed method for two violins performing a vibrato, the first with and the second without the $\alpha$ filtering stage; and their respective estimates, $\bar{\mathbf{A}}$ (filtered, in red) and $\mathbf{A}$ (non-filtered, in blue).

### 5.2.3 System Overview

A flow chart of the proposed system is depicted in Figure 5.5. The first procedure consists in computing a compressed magnitude spectrogram $\hat{\mathbf{X}}$ of the input signal $x$, for which an appropriate window size must be selected, considering the nature of the signal. After many experiments, a length of 21.3 ms ($N_1 = 1024$ samples) have shown to be a good generic starting point; using longer window sizes tend to degrade fast chirps, and therefore degrade the $\alpha$ estimation procedure, while using shorter windows provide too low frequency resolution. A hop size of $N_1/4$ is also a good compromise, even if this choice did not show a significant effect on the quality of the estimation procedures within the range $N_1/8$ to $N_1/2$. After that, the spectrogram is used as the input image for the structure tensor procedure, which produces as outputs a set of angles $\theta$ and anisotropy measures $C$. For each time frame $m$, a distribution using a sum of Gaussians is generated and a set of $Z$ frequency slopes $\alpha$ (in matrix $\mathbf{A}$) are estimated. Then, the filtering procedure is applied, and a vector of values of $\bar{\alpha}$ (in matrix $\bar{\mathbf{A}}$) is provided. At this point, the $Z$ instances of FChT are computed for the input signal $x$ and stacked together with additional spectrograms into tensor $\underline{\mathbf{X}}$, after a simple linear 2-D interpolation has been applied to these TFRs in order to match their dimensions. Finally, all these TFRs are combined using some combination method, e.g. the SLS, studied in Chapter 4.2. At the end, the system gives as a result the final combined TFR $\mathbf{X}^{\text{Comb}}$.

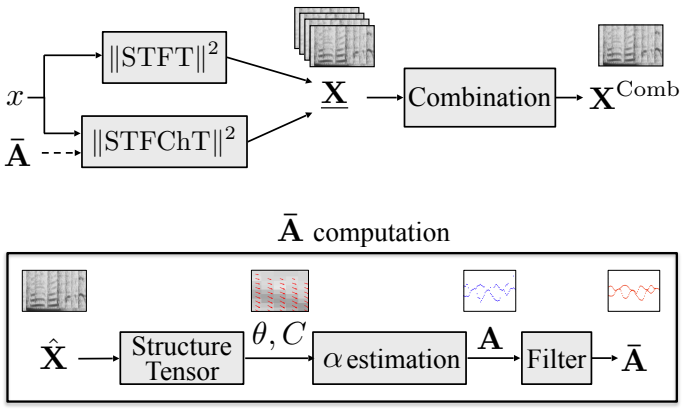



Figure 5.5: Flow chart of the FEMD method.

Although the system is tailored for dealing with frequency varying signals, e.g. harmonic instruments or vocals performing vibratos, the combination procedure can successfully deal with virtually any kind of musical input signal, thanks to the inclusion of spectrograms. In fact, it is fairer to say that the proposed system differs from the usual combination procedures because it uses, in addition to spectrograms, TFRs based on the FChT, which improve the representation of harmonic frequency chirps. Anyway, it is a robust and adaptable tool for generating high-resolution TFRs.

### 5.2.4 Proof of Concept

Here, some audio signals will be processed and analyzed to illustrate some characteristics of the proposed method. All signals were sampled at $F_S = 48\,\text{kHz}$; the spectrogram required to compute the structure tensor used a 1024 samples long (21.3 ms) analysis window with a hop size of 256 samples; the 2-D Gaussian smoothing filter $\mathbf{G}$ (see Equation (5.6)) of the structure tensor was $9 \times 9$ samples long with standard deviations $\sigma_m = 1.5$ ($\approx 8\,\text{ms}$) and $\sigma_k = 1.5$ ($\approx 70\,\text{Hz}$), which are minimum values to effectively reduce interference from background noise; the dynamic range chosen was $R = 40\,\text{dB}$, which is enough to separate the frequency lines of interest from the background noise; in the estimation of the $\alpha$ distribution, the standard deviation of the Gaussians was $\sigma = 0.07$, which assures that peaks close to each other would not overlap each other, yet a smooth distribution would still be generated; only peaks not lower than 5% of the highest peak in the estimated distribution were considered relevant; and the median and smoothing filters applied to $\alpha$ were 5 samples long, which experimentally has proven to be a minimum filter length to attenuate the problems observed in most scenarios.

Using the guidelines recommended in Chapter 4, the combinations were performed with the SLS method configured with $\zeta = 70$, which is a high value that guarantees a sufficiently high prevalence of the best bin in the combination; analysis window $\mathbf{W}^{\text{S}}$ with 6 bins ($\approx 30\,\text{ms}$) of time span and 21 bins ($\approx 492\,\text{Hz}$) of frequency span; and analysis windows $\mathbf{W}^{\text{E}}$ with 6 bins of time span and 9 bins ($\approx 211\,\text{Hz}$) of frequency span. Besides a set of FChT-based spectrograms computed using the $\alpha$ estimated using the proposed method, both the spectrograms used for computation of the structure tensor (window length of 21.3 ms, $N_1 = 1024$) and a spectrogram with the same resolution of the FChT ones (window length of 42.6 ms, $N_2 = 2048$) were included in the combination.

The first example was computed with two synthetic harmonic signals with sinusoidal $f_0$ variation and additive white Gaussian noise, with SNR = 50 dB. These signals represent what typically occurs when a vibrato is performed on a musical instrument or by a vocalist. By summing two signals with this characteristic, one obtains several points where frequency lines cross each other, which are usually the most difficult time-frames regarding the estimation of $\alpha$. In Figure 5.6 one can see the standard spectrogram obtained (top), and the resulting combined TFR (bottom), with the system configured to estimate two simultaneous $\alpha$'s.

As desired, a sharper TFR was achieved by the proposed method, with a very fine overall resolution, as expected when one uses the FChT with the correct values of $\alpha$. It is worth noting that, for low frequencies, the results are nearly the same in both TFRs, since the frequency slopes are low. However, for high frequencies, the

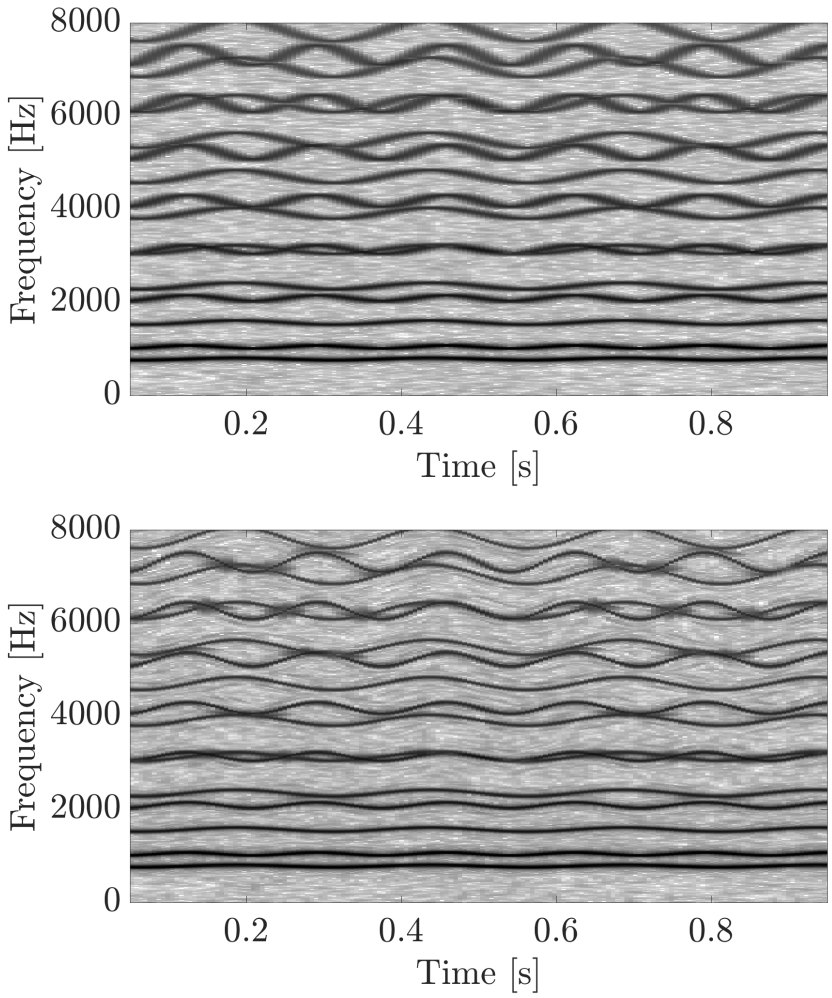



Figure 5.6: TFRs of two synthetic vibrato signals. Standard spectrogram and the resulting combination of FChT spectrograms, respectively.

results are drastically different.

This signal is a good example to evidence the system's capability of simultaneously estimating multiple $\alpha$'s and the problem of combining TFRs in points where two frequency lines with very distinct $\alpha$'s cross each other [74]. In such regions, a darker shade surrounding the frequency lines can be observed, which is caused by the energy smearing that necessarily occurs in this situation, since there is no representation that well represents both components.

In Figure 5.7 one can see the very good $\alpha$ estimates achieved. Note that occasional outliers occurred, and the filtering procedure was effective to overcome this issue. Besides, one can see a region where the estimation procedure was not capable of picking the correct values of $\alpha$, as probably the peaks obtained on the $\alpha$ distribution were too close to each other. This region is circled by a green-dashed-line.

Now, some real audio recordings representative of the signals this method is

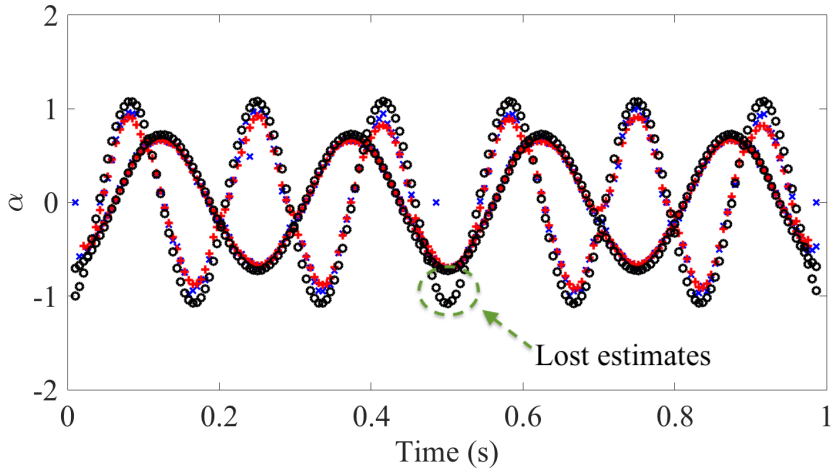



Figure 5.7: Estimated $\alpha$'s of a pair of synthetic vibrato signals: theoretical (in black), estimated (in blue), and filtered (red) values. In green, a region where the correct $\alpha$'s could not be estimated.

intended to deal with will be examined. The first signal contains an excerpt of an opera piece, comprised of a voice singing along with a full orchestra. Typically, vocal lines in opera present very fast frequency variations, such as accentuated vibratos and glissandos. In Figure 5.8 one can see the TFRs computed for this signal. The combined version yields a much better definition of the vocal frequency lines, while maintaining the same resolution of the STFT for the stationary components, since they share a common window size. Here, the system was also configured to estimate two simultaneous $\alpha$'s.

The other real-world signal processed is an excerpt of two violins performing glissandos in opposite directions summed with a snare drum. The snare drum produces a sharp and very wide-band sound at its attack, followed by some body resonance, mainly at low and low-mid frequencies. This signal is useful to visualize the method's robustness when dealing with both harmonic and percussive sources simultaneously. The system here was asked to estimate three $\alpha$'s per frame. Figure 5.9 depicts the initial spectrogram and the final combined TFR for this signal. As can be seen, the violins' partials could be much more clearly represented by the latter, along with the drums sound. The snare drum onset can be seen as a vertical pattern around 0.2 s, with a noticeable energy concentration at low frequencies. In the proposed TFR, the attack is sharper than in the spectrogram version, while the body resonance of the drum (horizontal frequency lines that follow after the attack) is still clear. This happens thanks to the inclusion of the spectrogram with shorter window in the combination procedure. As an overall observation, one can state that the combined TFR provides a better time-frequency resolution.

As mentioned, other combination procedures can be used in this method. For instance, the SWGM, which is a much simpler combination in terms of computational

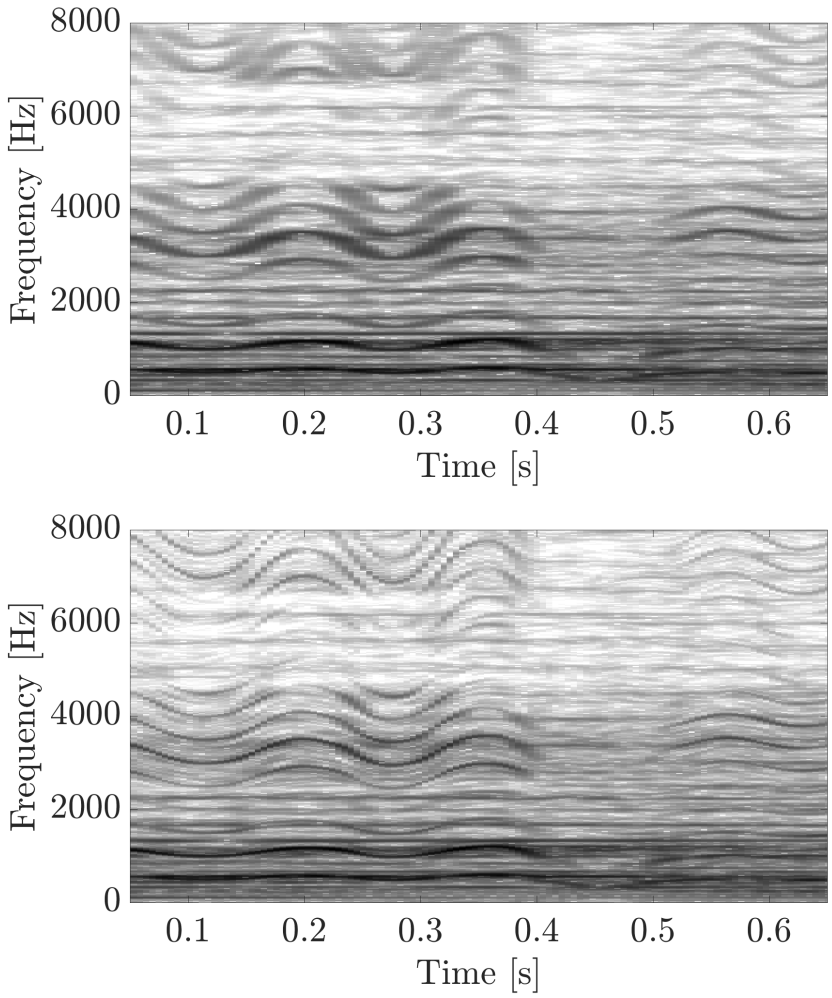



Figure 5.8: TFRs of a vocal with orchestra. Standard spectrogram and the resulting combination of FChT spectrograms, respectively.

burden, can be successfully employed, at the expense of providing a less sharp representation. Figure 5.10 depicts the resulting combination for this same signal, now using the SWGM, while Figure 5.11 depicts a zoomed region of the three representations for more detailed visual comparison. Note that the lines are, in general, less well defined, as there are more representations to be combined and the discrepancy of energy spread is severe, causing the peak distortion discussed in Chapter 3.5. The choice of the combination method will then depend on the system which will use the representation and on the computational resources available.

## 5.3 The Multi-Resolution Fan-Chirp Interpolation Method

In this section, a different approach for combining TFRs is presented, namely, the multi-resolution fan-chirp interpolation (MRFCI) [73]. Now, instead of using the

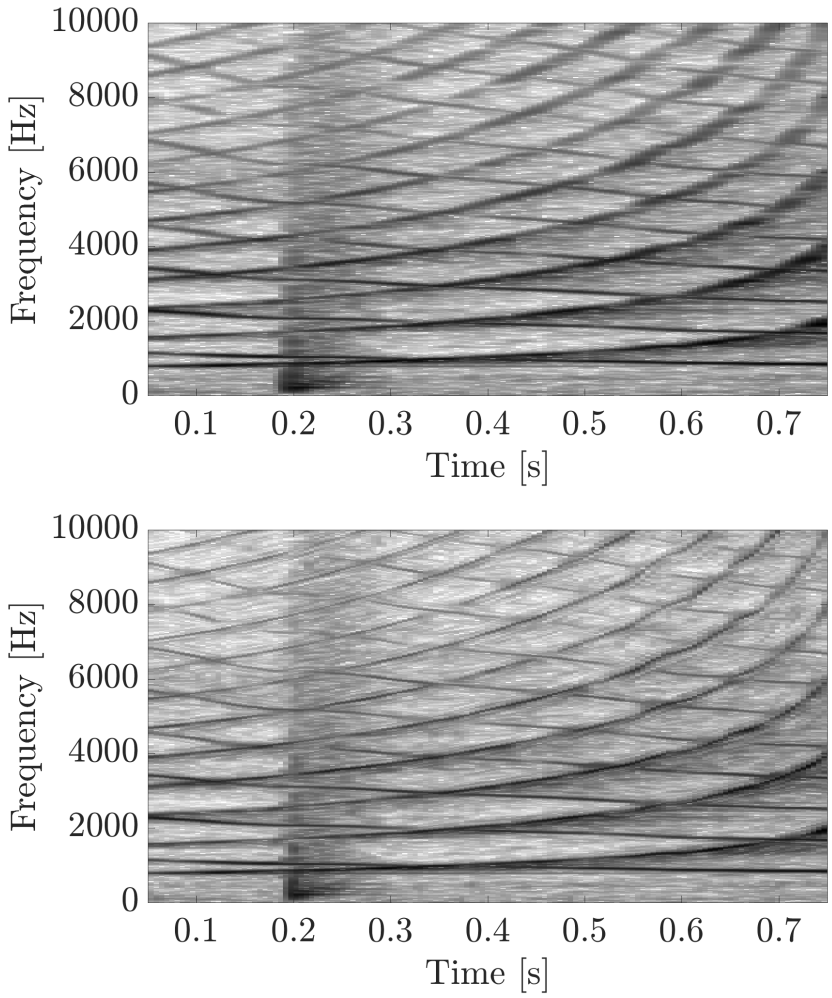



Figure 5.9: TFRs of two violins and a snare-drum. Standard spectrogram and SLS combinations, respectively.

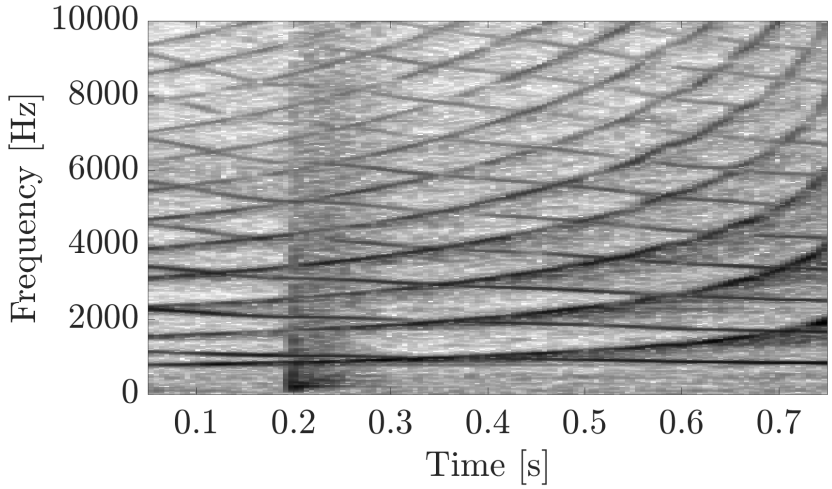



Figure 5.10: TFRs of two violins and a snare-drum. Standard spectrogram, SWGM and SLS combinations, respectively.

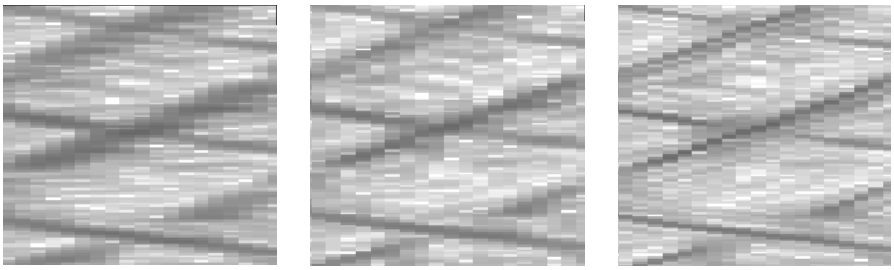


Figure 5.11: Zoomed region of TFRs comprising two violins and a snare-drum. Standard spectrogram, SWGM and SLS combinations, respectively.

information provided by the structure tensor to estimate directions and generate TFRs to be combined, such geometric information is used to guide a combination procedure which builds the final TFR from a pre-computed dictionary of FChT-based spectrograms with different $\alpha$ chirp rates and analysis window sizes. This method is then more flexible, as it does not require the pre-setting of the number of simultaneous $\alpha$'s to be estimated and adapts the resolution of the analysis window to the region. Besides, it demands much less computational power, due to the simple interpolation procedure performed to produce the final TFR, when compared to the FEMD using the SLS method.

### 5.3.1 Principles of the Method

The structure tensor outputs, i.e. the set of angles $\theta$ and the set anisotropy measures $C$, comprise the information of direction and steepness of the region in such direction, respectively, for each time-frequency bin. As an example, Figure 5.12 depicts a small region of the spectrogram of an audio signal with blue arrows representing vectors pointing at direction $\theta$, 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 5.12, two different regions are highlighted: the arrows inside region 1 present smaller magnitudes 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 5.1.1): a too small-dimension $\mathbf{G}$ induces smaller regions, which favors a linear model, and thus decreases the effect observed. Also, it is worth noting that the magnitudes ($C$) vary smoothly over the whole time-frequency domain, which will assure smooth transitions between different TFRs in the combined result.

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

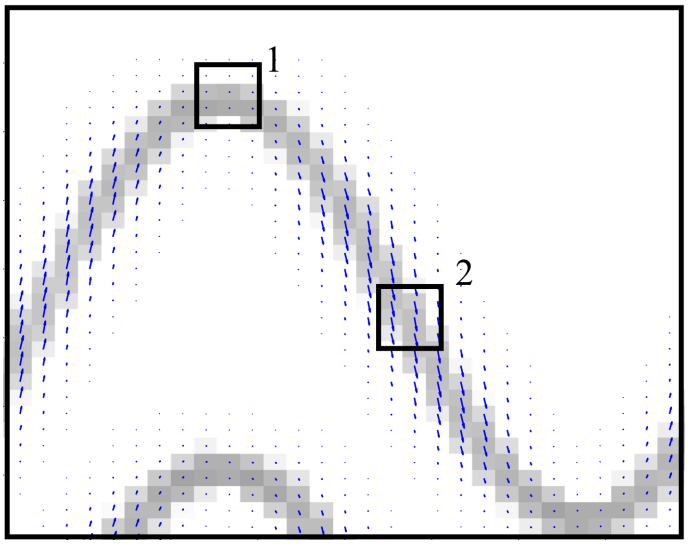



Figure 5.12: Vectors in $\theta_{k,m}$ directions with magnitudes $C_{k,m}$.

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 $\alpha$ for a such a long 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 best analysis window length to be used; and parameter $\alpha$ can then be used to choose the best fan-chirp representation for each time-frequency bin. In the end, the method consists in performing a linear combination of time-frequency bins of the best candidates among a set of FChT-based spectrograms with different $\alpha$'s and analysis window lengths.

A flow-chart of the MRFCI method is depicted in Figure 5.13. Firstly, the audio signal $x$ is processed to generate a set of different FChT-based spectrograms using predetermined sets of chirp slopes and analysis window sizes, $\boldsymbol{\alpha}$ and $\mathbf{N}$, respectively. Then, all TFRs are interpolated[4] 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 $\mathbf{C}$, indicating the preferable window sizes, are computed. Finally, a simple linear combination of the TFRs in $\underline{\mathbf{X}}$ is performed according to $\mathbf{A}$ and $\mathbf{C}$ for each time-frequency bin (as will be described in Section 5.3.3), resulting in the combined TFR $\mathbf{X}^{\text{Comb}}$.

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

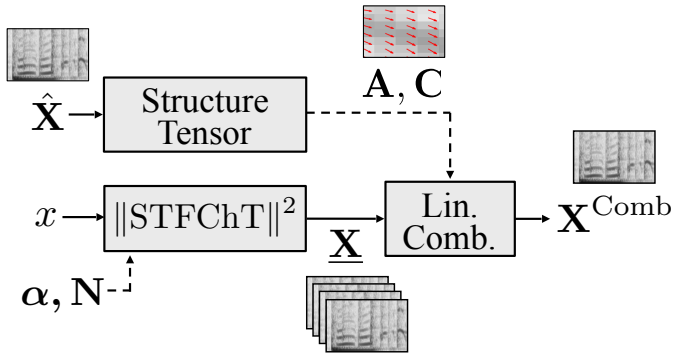



Figure 5.13: Flow-chart of the MRFCI method.

### 5.3.2 Computation of the Dictionary Tensor

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. As an example, Figure 5.14 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 spectrograms with short windows will be used in the combination procedure, and angles that indicate the presence of tonal information, which will be represented by FChT-based spectrograms with proper window length and parameter $\alpha$. These two angular regions are indicated in Figure 5.15, similarly to what is done in [54], where a percussive/tonal source separation procedure is performed.

For computation of the optimum $\alpha$'s distribution, an equally spaced distribution of angles $\vartheta$ is adopted, in order to minimize the energy smearing in tonal regions. Consider that the angular region $[0, \vartheta^{\max}]$ will be divided into $I$ parts. This maximum analog angle is related to an $\alpha^{\max}$ by the same relation described in Equation (5.12), which indicates that the analog angle $\vartheta$ is proportional to $\tan\alpha$. Since

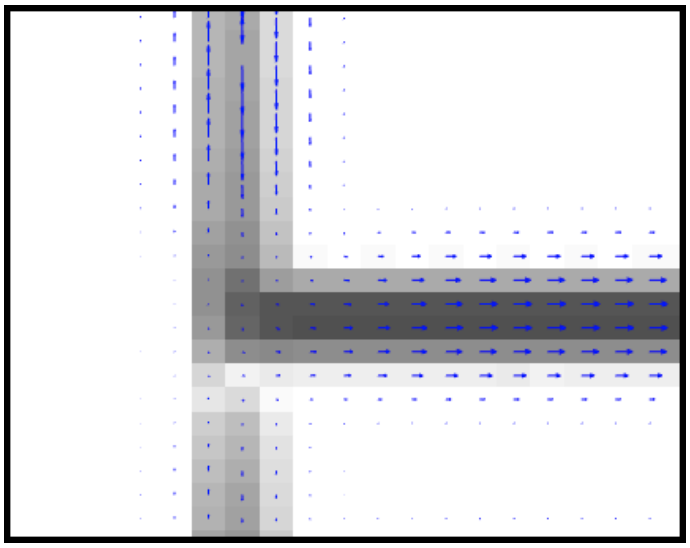


Figure 5.14: Spectrogram: onset of a harmonic pulse. Vectors in $\theta_{k,m}$ directions with magnitudes $C_{k,m}$.

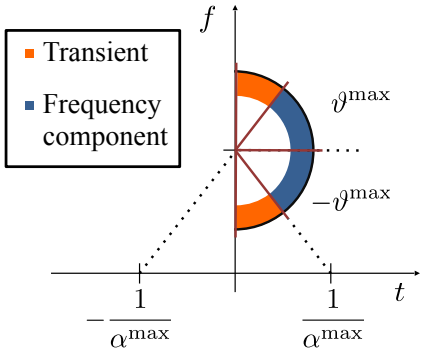



Figure 5.15: Angular regions associated to transient and tonal information.

parameter $\alpha$ better describes the behavior of varying harmonic frequency content,[5] 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 (2.15), since there is a range of values of $\alpha$ 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 (5.12),

$$\tan(\vartheta^{\max}) = f\alpha^{\max}. \tag{5.15}$$

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

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

$$\vartheta^{\max} = \arctan(\alpha^{\max}), \tag{5.16}$$

and

$$\vartheta_i = i\frac{\vartheta^{\max}}{I} = i\frac{\arctan(\alpha^{\max})}{I}. \tag{5.17}$$

Finally, we can project a linear distribution of $\vartheta$ into $\alpha$ by computing $\alpha_i$ as

$$\alpha_i = \tan(\vartheta_i) = \tan(i\arctan(\alpha^{\max})/I), \tag{5.18}$$

and the set of $\alpha$'s that we shall use to compute the FChT-based spectrogram symmetrically spans this distribution with positive and negative values:

$$\boldsymbol{\alpha} = [-\bar{\alpha}_I, -\bar{\alpha}_{I-1}, \ldots, -\bar{\alpha}_1, \bar{\alpha}_0, \bar{\alpha}_1, \ldots, \bar{\alpha}_{I-1}, \bar{\alpha}_I]. \tag{5.19}$$

For choosing the best distribution of analysis window lengths $\mathbf{N} = [N_1, N_2, \ldots, N_J]$, since the FFT algorithm is used, powers-of-two are a preferable choice. This criterion is used to choose the elements of $\mathbf{N}$, optimizing this way the computational cost regarding this parameter. The parameters to be set are, then, $\mathbf{N}$, following the aforementioned criterion, and the number $I$ of parameters $\alpha$. A set of TFRs is then composed of several instances of FChT-based spectrograms using the combinations of $\mathbf{N}$ and $\boldsymbol{\alpha}$, and a spectrogram computed with $N_1$ (for the transients). Note that the sets of FChT-based spectrograms also include common 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 $K_J$[6] frequency bins after the interpolation, and must be synchronized. In the present implementation, the same hop size is used for computing all TFRs, but, as mentioned before, this does not guarantee a proper time alignment between different spectrograms, reason why the time-wise interpolation (or a previous time shift in $x$) is also necessary. The set of parameters $\alpha$ 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 $\theta$ to $\alpha$ 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 $N_j$ and a chirp rate parameter $\bar{\alpha}_i$. Since the transient

[6]It is worth mentioning once more that the number of frequency bins in the spectrogram is $K = 1 + N/2$, hence $K_J = 1 + N_J/2$.

information will be represented by a spectrogram originally having $K_1$ frequency bins, it is allocated at the first and last positions in dimension $\alpha$ for all layers $j$, and therefore it will not be necessary to compute FChT-based spectrograms using the first and last values of $\alpha$, i. e. $\bar{\alpha}_I$ and $-\bar{\alpha}_I$. Figure 5.16 depicts the tensor $\underline{\mathbf{X}}$, where groups of TFRs with different $\alpha$'s are illustrated clustered according to the original number of frequency bins $K_j$.

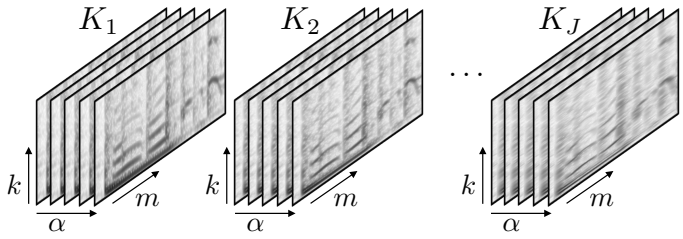



Figure 5.16: Scheme of the four-dimensional TFR dictionary tensor.

### 5.3.3 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 5.17 depicts an example of the weights related to the parameters $\alpha$, $\lambda^\alpha$, for $I = 2$. The weight $\lambda^\alpha_i$ is applied to the $i$-th layer of $\underline{\mathbf{X}}$, so the centered weight in the image, $\lambda^\alpha_0$, in black, is related to the layers in $\underline{\mathbf{X}}$ which were computed with $\bar{\alpha}_0 = 0$, the others in blue, $\lambda^\alpha_1$ and $\lambda^\alpha_{-1}$, are related to the layers computed with $\bar{\alpha}_1$ and $\bar{\alpha}_{-1}$, and the last ones, $\lambda^\alpha_2$ and $\lambda^\alpha_{-2}$, in orange, are related to the STFT computed with short window, 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, $\lambda^C$ (depicted in Figure 5.20) will be used for weighting the layers of $\underline{\mathbf{X}}$ along dimension $j$, which is related to $K$. In this example, the weighting function of $\lambda^C_2$ is centered in 0.5, while the others are set to the minimum and maximum extremes.

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

$$X^{\text{Comb}}_{k,m} = \sum_{j=1}^{J} \sum_{i=-I}^{I} \lambda^C_{k,m;j} \lambda^\alpha_{k,m;i} X_{k,m;j,i}. \tag{5.20}$$

As mentioned at the beginning of the chapter, the computation of the structure tensor as proposed in this thesis provides much better results for the MRFCI method, since the anisotropy measure is more focused on the frequency lines. This

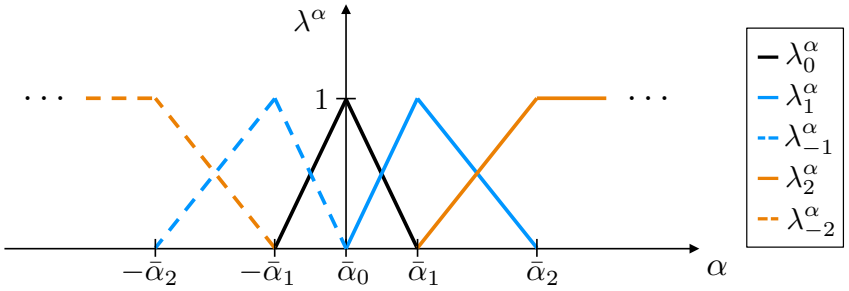



Figure 5.17: Example of the weights used for combining TFRs with different $\alpha$'s ($I = 2$).

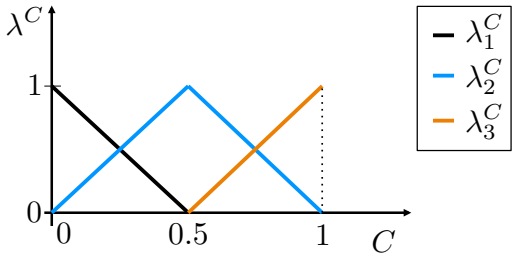



Figure 5.18: Example of the weights used for combining TFRs with different $K$'s ($J = 3$).

is especially critical in regions where there are fast frequency variations. The high anisotropy at bins outside the regions of interest cause energy smearing, since large analysis windows will be used to compose such bins in the final TFR, and this will make them integrate the surrounding energy. This effect can be observed in Figure 5.19, where the test signal gradually increases the frequency variations. The combination computed with the standard structure tensor (on the left) clearly exhibits more energy smearing.

### 5.3.4 Practical Considerations

In practice, CPU processing can be saved by using only $\alpha = 0$ for small window sizes, e.g. $N = 1024$ ($\approx 21$ ms), achieving very similar results. Another practical consideration relates to 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 corresponding memory space can be freed up. This approach is also very appropriate for parallel processing.

Regarding the size of the two-dimensional smoothing filter $\mathbf{G}$, which has a direct influence on the choice of the analysis window to be used, as mentioned, it needs to be large enough to encompass a region compatible with the maximum analysis

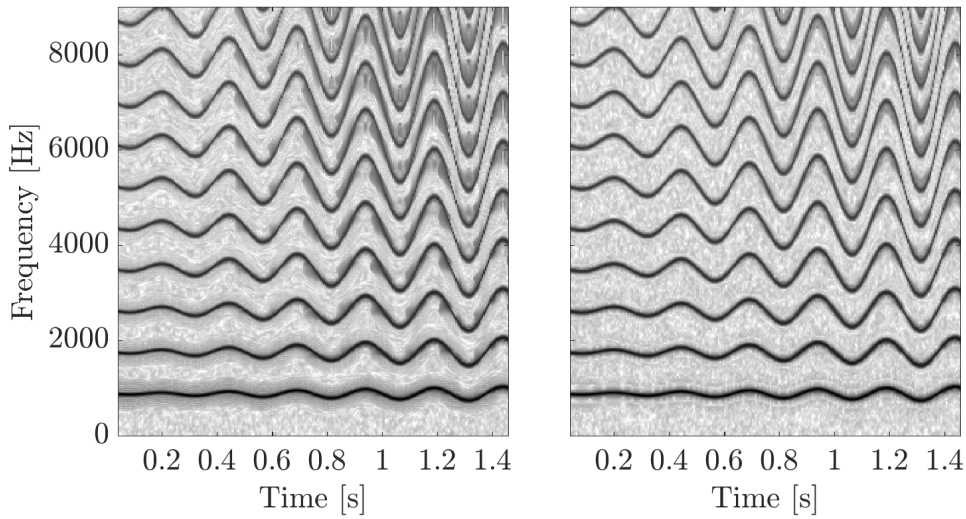



Figure 5.19: Varying vibrato: MRFCI combinations using dictionaries of $I = 7$ and $N \in \{1024, 2048, 4096\}$, computed using standard structure tensor and the proposed modified structure tensor, respectively.

window size, and therefore providing a good estimation of the best analysis window to be used. Since a Gaussian filter is being used, its dimensions are related to its standard deviations $\sigma_k$ (frequency) and $\sigma_m$ (time). Experimentally, using $\sigma_k$ relative to a frequency range around 100 Hz and $\sigma_m$ relative to $1/4$ of the length of the larger analysis window yields excellent results.

Also, in order to reduce backwards smearing of attacks, asymmetrical analysis windows having a longer tail on the left side can be used for the longer windows. Although the structure tensor indicates the presence of the attacks and assigns short windows for such regions, having frequency components before the attacks may lead to the use of large windows, which can create energy smearing artifacts. Nevertheless, this has the side effect of decreasing the frequency resolution provided by the long window. In the current implementation, the asymmetric windows are computed by concatenation of the first half of a Hanning window computed with $N$ samples, and the second half of a Hanning window computed with $N/2$ samples.

For choosing the transition points for the weights $\lambda^C$, setting the regions according to the relation of window sizes has shown to be a good approach. For instance, Figure 5.20 depicts a good configuration of $\lambda^C$ if the window sizes $N \in \{1024, 2048, 4096\}$ are to be used, tending to yield the minimum energy smearing possible for this case. If a window of 3072 samples were also included, its weight $\lambda^C$ would be centered in $C = 0.75$. Experimentally, the difference that results from the inclusion of such a window is negligible.

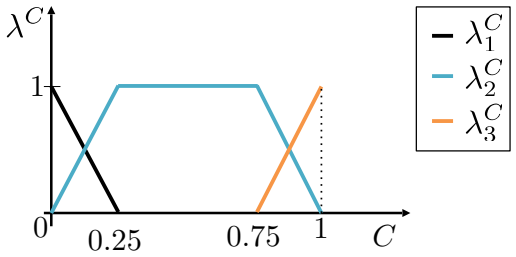



Figure 5.20: Example of the weights used for combining TFRs with using $\mathbf{N} = [1024, 2048, 4096]$.

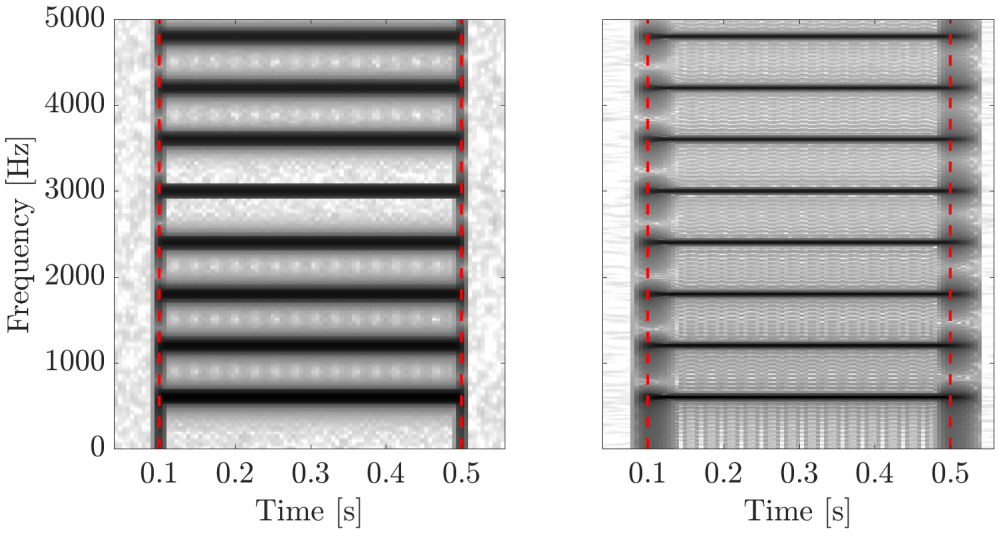
### 5.3.5 Proof of Concept

In order to assess the performance of the proposed method, both synthetic and real-world audio signals were analyzed. All input signals had sampling rate $F_s = 48000\,\text{Hz}$. The system was set according to the following configuration. In the structure tensor procedure, the analysis windows of the spectrogram had length $N = 1024$ (21.3 ms); in the smoothing two-dimensional filter $G$, $\sigma_k$ corresponded to 100 Hz and $\sigma_m$ to 21.3 ms; and the dynamic range used in the analysis spectrogram was $R = 50$. The analysis window sizes for the FChT-based spectrograms were chosen as $N \in \{1024, 2048, 4096\}$ (21.3, 42.6 and 85.3 ms); in order to reduce backwards energy smearing, asymmetric analysis windows were used for the computation of FChT-based spectrograms with $N_3$; $\alpha^{\max} = 23.4$ resulted from the application of Equation (2.15); and all TFRs were computed with hop size $h = 256$ samples.

As a proof of concept, synthetic signals were selected to assess the method 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 5.21(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 5.21(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.

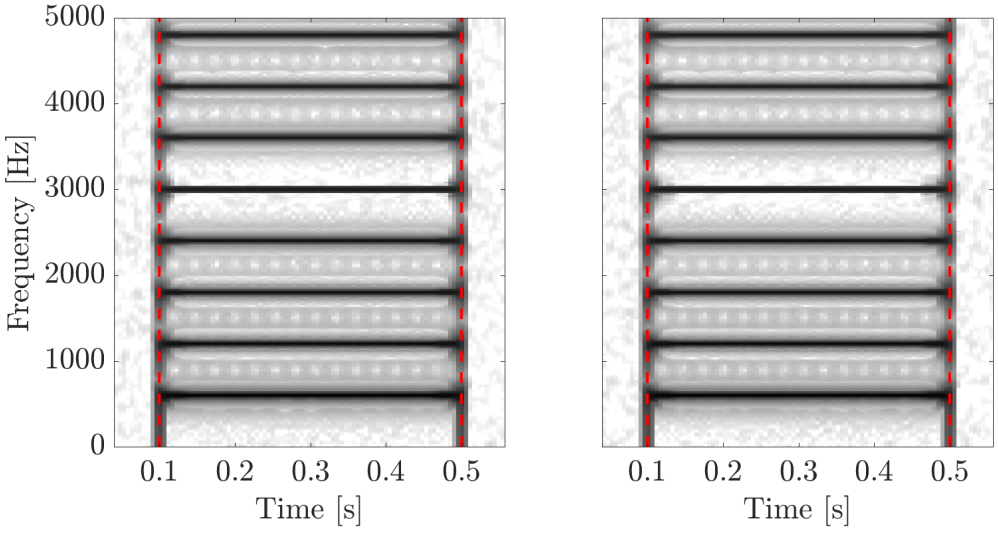
As can be clearly observed, the two TFRs computed with the proposed method yielded nearly identical results, combining the time precision provided by the first spectrogram with the frequency resolution of the second one. Since the frequency lines present in this signal are well represented by an FChT-based spectrogram with $\alpha = 0$ (i.e. a spectrogram), increasing the number of FChTs 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




(a) Spectrogram: $N = 1024$.

(b) Spectrogram: $N = 4096$ (asymmetric).



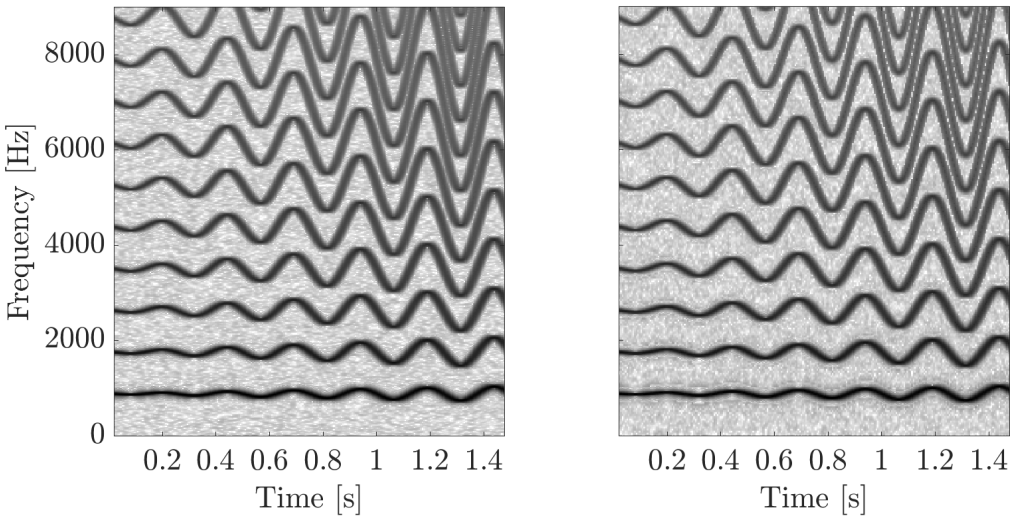

(c) MRFCI: $I = 1$.

(d) MRFCI: $I = 5$.

Figure 5.21: Spectrograms computed with different window sizes and MRFCI combinations with different $I$ computed for a pulse composed of harmonically related sinusoids. Onset and offset are indicated by the red-dashed lines.

(SNR = 50 dB). This signal allows one to verify the capability of handling a wide variety of $\alpha$'s. The results are depicted in Figure 5.22, 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.

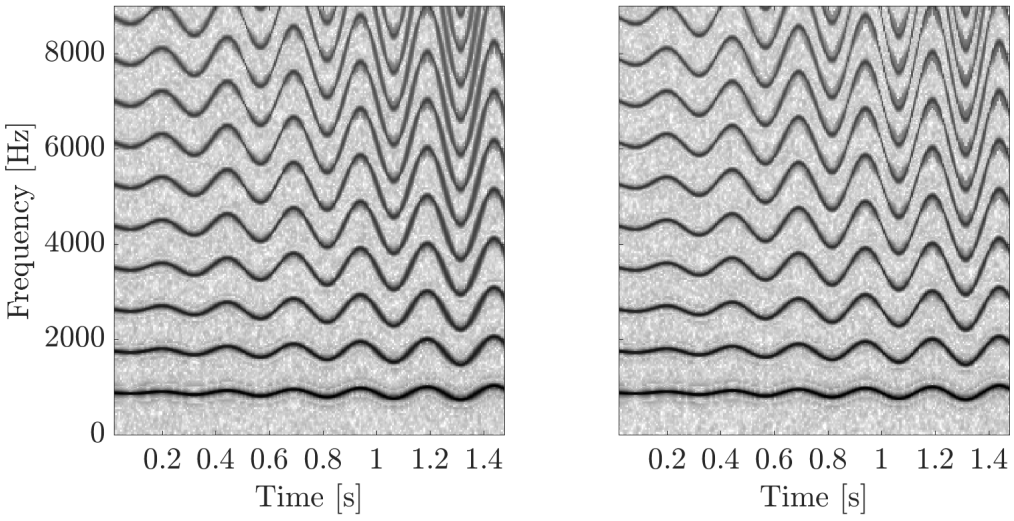
The last synthetic signal is a sum of two harmonic signals having different sinusoidal variations of $f_0$, with additive white Gaussian noise (SNR = 50 dB). For




(a) Spectrogram: $N = 2048$.




(b) MRFCI combination: $I = 1$.




(c) MRFCI combination: $I = 3$.




(d) MRFCI combination: $I = 5$.

Figure 5.22: Varying vibrato: spectrogram and the MRFCI combinations, using dictionaries of different $I$'s and $N \in \{1024, 2048, 4096\}$.

comparison, Figure 5.23 depicts the TFR obtained using the FEMD method, with the SLS combination procedure, and the combined TFR using the MRFCI method, for which $I = 7$ and $N \in \{1024, 2048\}$ was used (to provide a fair comparison). The resulting TFR represents the input signal with a similar definition, and very smooth transitions can be observed. In particular, the regions where crossings of frequency lines are better represented, and there is a slightly higher contrast between the frequency lines and the background noise.

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 5.24 depicts its original spectrogram ($N = 2048$) and the combined TFR,

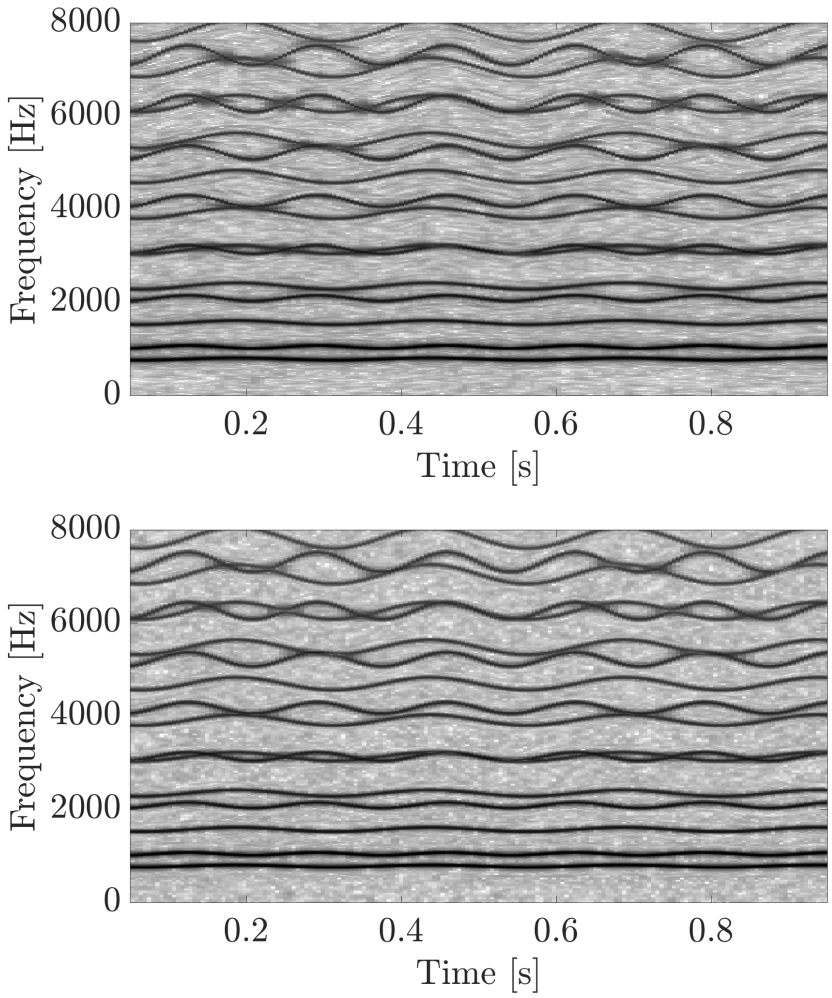



Figure 5.23: TFRs of two synthetic vibrato signals. At the top, the representation obtained via FEMD procedure using the SLS combination, and at the bottom, the MRFCI combination ($I = 7$).

using $I = 7$ and $N \in \{1024, 2048, 4096\}$. Once again, the TFR generated by the MRFCI method provides a clear high-resolution representation, which clearly represents both the piano and the singing vocal—which is performing a very fast melisma. The contrast is also enhanced, in this example.

## 5.4 Concluding Remarks

In this chapter, two methods for generating high-definition TFRs were presented. Both methods make use of the structure tensor technique, which indicates the direction of frequency lines, and thus allow for the fast computation of the frequency slope parameter $\alpha$, used in the FChT.

The FEMD provides a set of FChT-based spectrograms, which are computed using the $\alpha$'s estimated. Then, these instances are combined using some combination method, such as the methods presented before. As for the MRFCI, it is

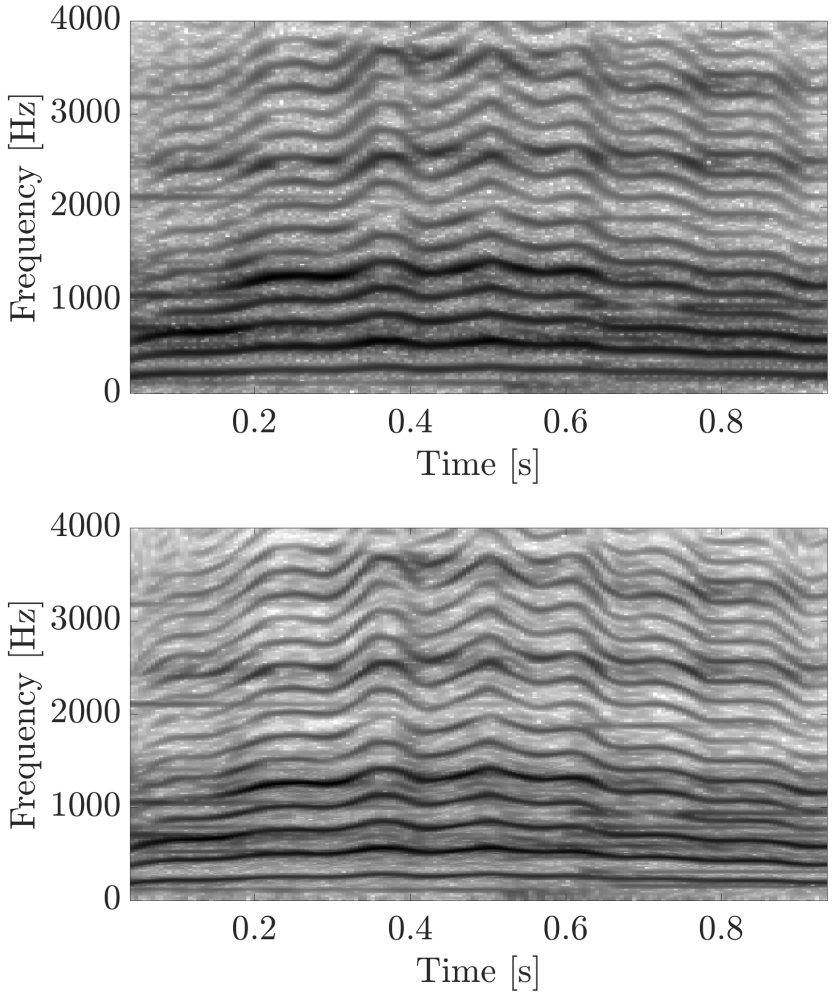



Figure 5.24: Vocal and piano: spectrogram ($N = 2048$) and MRFCI combination ($I = 7$ and $N \in \{1024, 2048, 4096\}$), respectively.

a combination method which performs linear combination of samples present in a multi-resolution dictionary of representations based on the FChT, and this combination procedure is guided by the information provided by the structure tensor. This last method has proven to be very flexible and has a good tradeoff regarding the results achieved and its computational cost, which is much smaller than for the combinations based on local information.

In the next chapter, the experiments that will be presented compare all the approaches for generating high-definition TFRs in the context of melody analysis.

# Part II

# Music Information Retrieval: Experiments, Applications and Tools

# Chapter 6

# Experiments on Main Melody Analysis

This chapter contains experiments concerning the application of some TFRs among the methods studied in this thesis for melody analysis. The first set of experiments aims to characterize the TFRs when representing real-world melodic sources, while the second one is an attempt to apply different TFRs to a state-of-the-art system [41] that estimates dominant melodies from mixtures.

## 6.1 Time-Frequency Representations of Main Melody Signals

The objective of this first set of experiments is to use the most controlled environment possible to assess the performance of the different TFR combination methods when representing melodic signals, which is the main motivation of this thesis.

Here, the MDB-melody-synth dataset [94] is used. It contains solo tracks of main melody sources, i.e. tracks containing vocals and melodic instruments, with virtually perfect f0 annotations, which were created by resynthesizing the original signals to match the automatic annotations.

The synthesis is performed by an improved sinusoidal model that follows the amplitudes of the original signal and scale them in frequency according to the $f_0$ annotation; the unvoiced segments, which are annotated as $f_0 = 0$, are muted (for details, see [67, 94]). In [94], the authors show that the results yielded by different current state-of-the-art algorithms for melody extraction and multi-$f_0$ using the resynthesized dataset are statistically equivalent to the ones yielded when using manually corrected annotations with the original signals. Two files of the dataset were not used in the experiments: the song 'EthanHein_1930sSynthAndUprightBass', which

has a bass track[1] as main melody, and the song entitled 'MusicDelta_GriegTrolltog', whose annotations are unusable due to very poor estimations given the complexity of the signal. In total, 63 signals were used for the experiments.

Since the annotations perfectly match the audio signals, they can be used as guides to indicate the precise frequency of each harmonic and the exact location of each onset. Hopefully, this will provide us with useful information to estimate a statistical behaviour of each TFRs for main melody signals.

In order to save processing power and storage, the audio signals were downsampled to 22.5 kHz, which provides a frequency spectrum from 0 to 11.25 kHz. This frequency band is enough to comprise at least 10 harmonics for all signals. The TFRs computed for the experiments are the following:

**STFT** – STFT-based spectrograms computed with window sizes of 21.3, 42.6 and 85.3 ms ($N \in \{512, 1024, 2048\}$ samples, considering the new sampling rate);

**STFT-SWGM** – the three STFT spectrograms combined using the SWGM method ($\beta = 0.5$);

**STFT-SLS** – the three STFT spectrograms combined using the SLS method ($\zeta = 70$);

**FEMD-SLS** – the three STFT spectrograms and one FChT-based spectrogram, computed with the FEMD method and having analysis window of 42.6 ms ($N = 1024$ samples), combined using the SLS combination method ($\zeta = 70$);

**FEMD-SWGM** – the three STFT spectrograms with one FChT-based spectrogram, computed with the FEMD method and having analysis window of 42.6 ms ($N = 1024$ samples), combined using the SWGM combination method ($\beta = 0.5$);

**MRFCI** – spectrogram computed with the MRFCI method, using $I = 7$, and symmetrical analysis windows with durations of 21.3, 42.6 and 85.3 ms ($N \in \{512, 1024, 2048\}$ samples);

All parameters not mentioned here were set the same way they were described in the examples present in their respective chapters.

### 6.1.1 Experiment 1: MDB-Melody-Synth Dataset

The first experiment consists in computing, for all annotated voiced excerpts, the energy distribution around $f_0$ and its harmonics within a margin of 100 Hz, and

[1]The bass track was not included in the experiment since the distance between its harmonics was too small.

then computing the average energy distribution. This average peak format will summarize information of different melodic sources of the entire dataset.

In order to have a consistent peak estimation for all frames, only the harmonics 2 to 9 were evaluated, in such a way that the margins were always inside the time-frequency plane. Figure 6.1 depicts an example of a vocal present in the dataset, represented using the MRFCI method, along with its $f_0$ annotation (red line) and its corresponding margins (blue lines).

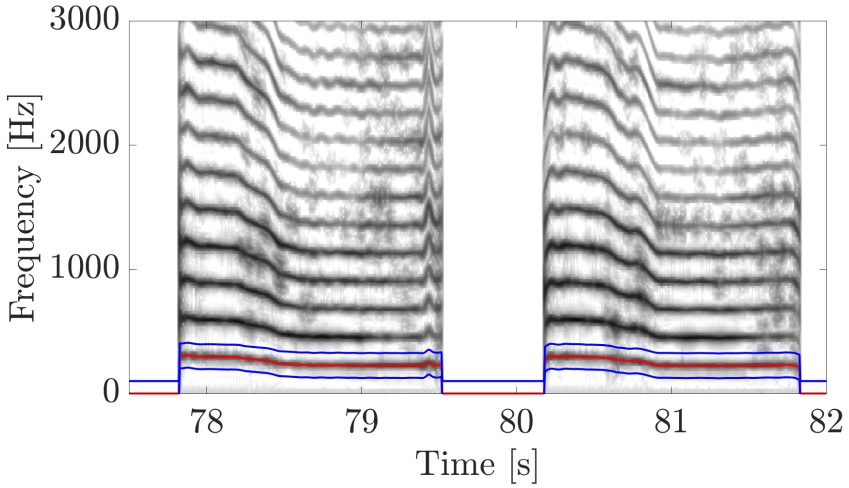



Figure 6.1: Resynthesized vocal signal represented with the MRFCI method, along with the annotated $f_0$ line (red) and the margins considered for the computation of the average peak (blue).

As can be seen, the annotation follows exactly the fundamental frequency. Also, it is worth mentioning that non-harmonic energy of the voice signal, which can be clearly seen between the frequency lines, is still present; therefore, this synthesis does not only considers the tonal frequency lines, but it addapts the whole frequency spectrum of the original signal to match the annotation. Such information can be linked to noise or the formants of the voice, i.e. the response of the vocal tract, which changes according to the vowel pronounced.

The peak average computation starts by interpolating the frequency samples of each frame to compute the peak format having multiples of the annotated $f_0$ as center, which will not necessarily fall into a frequency index of the representations. The first and the last 5 frames of each voiced segment are not taken into account, avoiding the energy spread in frequency inherent in onsets and offsets. Then, for each harmonic, the vectors containing the frequency peaks are stacked into a matrix; after that, the average peak of each harmonic is computed and normalized, and finally the average of such peaks is computed. Figure 6.2 depicts one of the matrices mentioned, which has the peaks disposed column-wise, in dB. One can see the energy concentration in the middle (horizontally), but there is also the presence of many

vectors with energy smearing (vertical dark patterns). Also, it is worth mentioning that the frequency lines have a large variation in amplitude, since the signal is a reproduction of a natural voice; hence, the average peaks will take all of these facts into account.

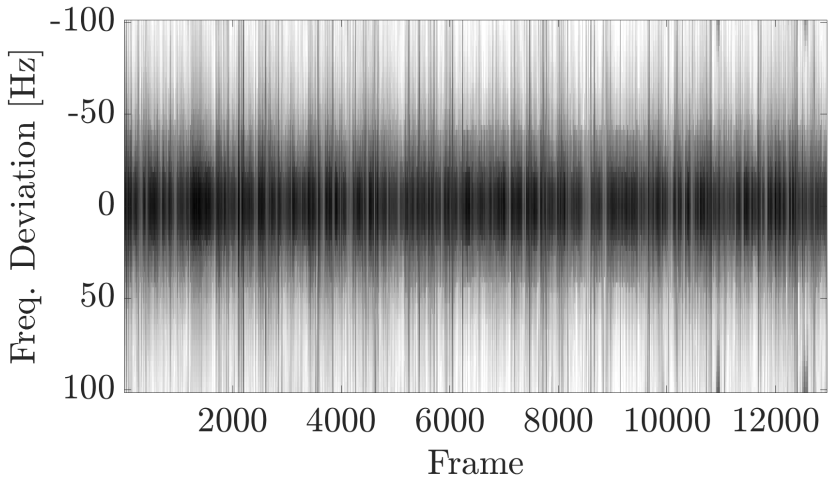



Figure 6.2: Matrix containing all the frequency peaks disposed side by side for one harmonic of a vocal signal.

Two figures-of-merit were chosen to compare the different average curves: bandwidth within $-3\,\mathrm{dB}$ (BW) and average dynamic range (DR), in dB, which is the average difference between the central peak and the valleys.[2] Figure 6.3 depicts the average frequency peaks, in dB, for all spectrograms, with peak normalization. The results are summarized in Table 6.1.

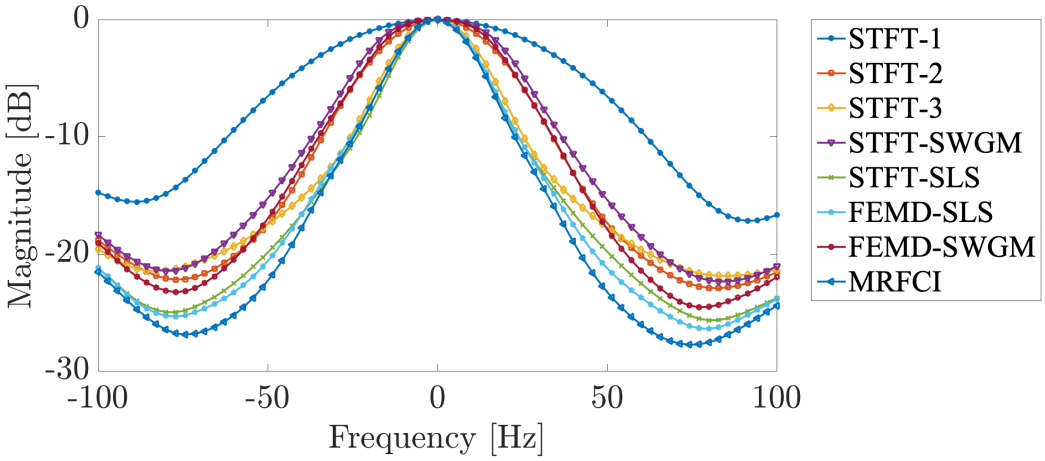



Figure 6.3: Normalized average frequency peaks, in dB, for all spectrograms.

As expected, the average peak for the STFT-based spectrogram results the worst of all. In this matter, it has the only advantage of not causing much energy smearing

[2]The curves tend to have asymmetric valleys due to the tendency of the harmonics' amplitude to decrease with frequency.

| | BW [Hz] | DR [dB] |
|---|---|---|
| STFT-1 | 68.4 | 16.4 |
| STFT-2 | 36.2 | 22.6 |
| STFT-3 | 25.2 | 21.7 |
| STFT-SWGM | 41.9 | 21.9 |
| STFT-SLS | 22.1 | 25.3 |
| FEMD-SLS | 23.3 | 25.9 |
| FEMD-SWGM | 37.7 | 23.9 |
| MRFCI | 22.8 | 27.3 |

Table 6.1: Results for average peaks using the MDB-melody-synth dataset.

in very non-stationary parts of the signal, which are not that frequent. Among the three STFT-based spectrograms, the one with larger window size seems to be the best option, on average. This sounds reasonable, since a large proportion of this type of audio signal is considerably stationary. Nevertheless, the non-stationary parts of the signal contributes to attenuate the average DR.

As for the combination of these spectrograms, the average curve obtained using the SWGM method resulted in a similar dynamic range of the best STFT, although the peak distortion inherent to this method led to a relatively large bandwidth. For this kind of signal, the SWGM combination will provide a more controlled energy smearing in non-stationary regions than the STFT using large analysis window, at the cost of having a worse overall peak definition. This will be better explored in the next experiment. The SLS combination, on the other hand, beat all the STFT-based spectrograms and their SWGM combination, having the smallest bandwidth and highest DR, although at the price of requiring a much higher computational cost.

The inclusion of a FChT-based spectrogram, via the FEMD method, resulted in a some improvement of DR and BW, when using the SWGM combination, but very similar results using the SLS method. It is worth noting that the FEMD method only includes a FChT-based spectrogram with medium window size, since it has no procedure for choosing the correct window length for each segment, as does the MRFCI. Using the FEMD method to include a FChT-based spectrogram with longer analysis window is possible, but the resulting TFR starts to suffer from too many artifacts, besides the further increase in computational burden. In the case of analysing signals with the presence of other sound sources, the number of FChTs would have to increase, making it too computationally expensive. Nevertheless, the FEMD method provides better resolution TFRs in non-stationary regions, which can be important to better capturing details in melody.

The MRFCI method, on the other hand, is a general purpose tool for generating TFRs, independently from the number of sources or nature of the signal under

analysis, and it has yielded the best results of all TFRs, except for having a slightly larger BW compared to STFT-SLS. As discussed before, the TFR computed using the MRFCI method can benefit from FChTs computed with higher resolution whenever needed.

### 6.1.2 Experiment 2: Synthesized Signals with Fixed Harmonic Relation

The previous experiment used the most natural dataset possible we found in the literature for this purpose. As mentioned, the natural variations in amplitude and energy outside the harmonic frequency peaks are present in the dataset. However, for this reason, it does not allow the analysis of some other characteristics of the methods, since only annotations regarding the $f_0$ are available. For instance, other important features to be preserved by a good TFR are the relative amplitude of the harmonics, which relates to the preservation of timbre,[3] and the energy evolution at the onsets.

For this reason, another dataset was generated by means of a bank of oscillators having as instantaneous frequency the annotated $f_0$ and its harmonics, with an arbitrary harmonic relation of amplitudes. For this experiment, it was imposed an energy loss, in dB, proportional to the harmonic index, i.e. the $h$-th harmonic is attenuated by $h$ dB. Although this attenuation factor is arbitrary, it represents a general tendency regarding the energy of harmonics produced by natural sound sources. As in the other dataset, the signals which compose this dataset still preserve all the natural frequency variations typical of melodic signals. Unvoiced excerpts are also muted. Now, with these more controlled signals, the frequency peaks can be tracked, allowing for a more detailed analysis in terms of amplitude.

#### Frequency Peak

First, the same procedure adopted on the previous experiment was applied to this dataset. Figure 6.4 depicts the new synthetic version of the same excerpt shown in Figure 6.1, also generated using the MRFCI method. As can be observed, only the first nine harmonics were generated, since the other ones are not evaluated.

Since the amplitude of each harmonic is theoretically constant, the energy smearing will be primarily caused by the incapability of the methods to correctly reproduce the fast variations in frequency, considering that all the transient parts of the signal were not taken into account. Figure 6.5 depicts the matrix obtained the same way as

[3]The timbre is a much more complex concept and it involves more than just the harmonics' amplitudes.

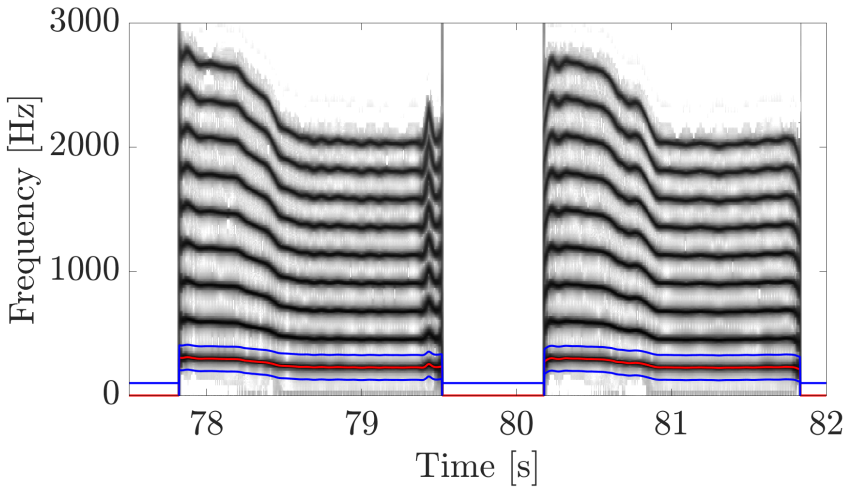



Figure 6.4: Vocal signal synthesized with arbitrary harmonic amplitudes and represented with the MRFCI method along with the annotated $f_0$ line (red) and the margins considered for the computation of the average peak (blue).

before, but for the new synthesized signal, where the energy can be observed much more consistently concentrated at the center of the vectors.

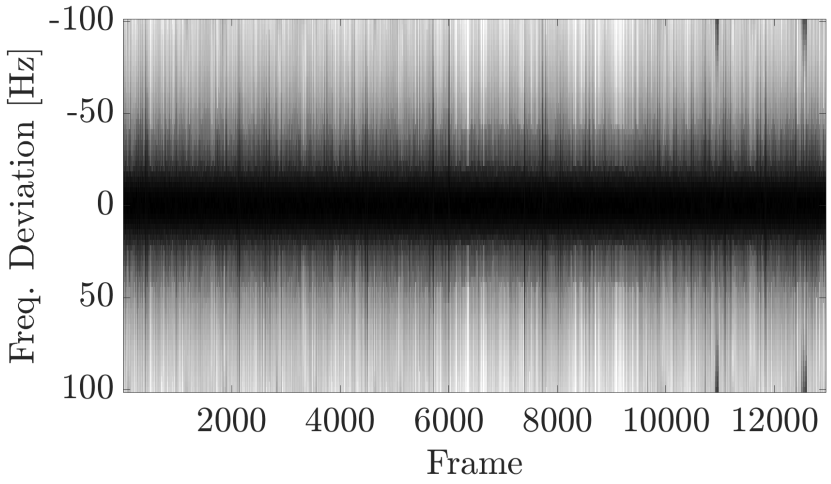



Figure 6.5: Matrix containing all the frequency peaks disposed side by side for one harmonic of a vocal signal (synthetic dataset).

Figure 6.6 depicts the average peaks computed for this dataset, using the same TFR methods, while the results are summarized in Table 6.2. As can be seen, now the resulting average curves have slightly more focused bandwidths and lower dynamic ranges. The main difference observed is regarding the FEMD method, which yielded higher improvements over the combinations of STFTs, making the average FEMD-SLS curve to be much closer to the MRFCI one. However, the overall results did not change drastically, which suggests that this analysis is considerably more dependent to the qualities of the methods and characteristics of the annotated $f_0$

than it is to the actual frequency content of the signals. Therefore, if one considers that the signals contained in this dataset are representative of what can typically be considered dominant melody sources, it seems fair to assume that the comparative performance of the methods within this pool relates to their comparative performance in real dominant melody analysis scenarios.

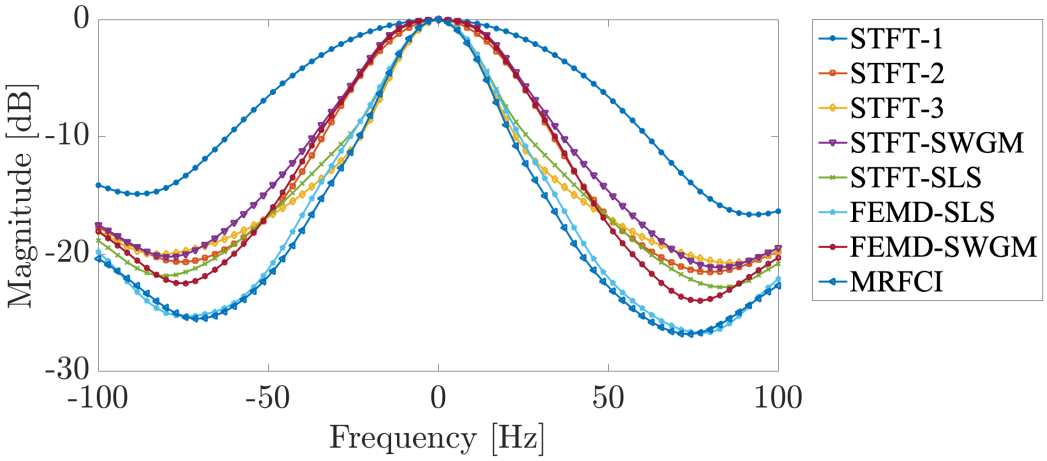



Figure 6.6: Normalized average frequency peaks, in dB, for all spectrograms (synthetic dataset).

| | BW [Hz] | DR [dB] |
|---|---|---|
| STFT-1 | 68.2 | 15.8 |
| STFT-2 | 36.1 | 21.1 |
| STFT-3 | 21.8 | 20.4 |
| STFT-SWGM | 38.6 | 20.7 |
| STFT-SLS | 23.2 | 22.4 |
| FEMD-SLS | 23.3 | 26.1 |
| FEMD-SWGM | 37.5 | 23.3 |
| MRFCI | 22.1 | 26.2 |

Table 6.2: Results for average peaks using the new synthesized dataset.

### Harmonic Relative Magnitude

This experiment is intended to assess the ability of the different methods to reproduce the correct magnitudes of each harmonic. Here, a different procedure was adopted: the peak values of each harmonic, which are the central samples of the aforementioned matrices, were stored for the whole dataset, and the attenuation previously imposed on each harmonic was compensated. Then, the median amplitude of each harmonic was computed and the distributions were normalized by the average median amplitude of each TFR method. This way, if a given representation preserves the harmonic relation imposed, the result should be similar distributions for all harmonics, with median magnitudes at 0 dB.

For a better visualization of the distributions, the boxplot standard will be adopted, as can be seen in Figure 6.7: the first quartile ($Q_1$), or lower quartile, is the middle value between the median and the smallest number in the dataset; the second quartile ($Q_2$) is the median value; the third quartile ($Q_3$), or upper quartile, is the middle value between the median and the largest number in the dataset; the central box contains the samples inside the interquartile range (IQR), which are the samples that sit between $Q_1$ and $Q_3$, IQR $= Q_3 - Q_1$, and thus concentrates 50% of the samples in the dataset. The IQR is used to define the range of feasible samples. Here, the maximum is set to the largest sample bellow $Q_3 + 1.5$IQR and the minimum is set to smallest sample above $Q_1 - 1.5$IQR. Samples outside this range are considered as outliers.

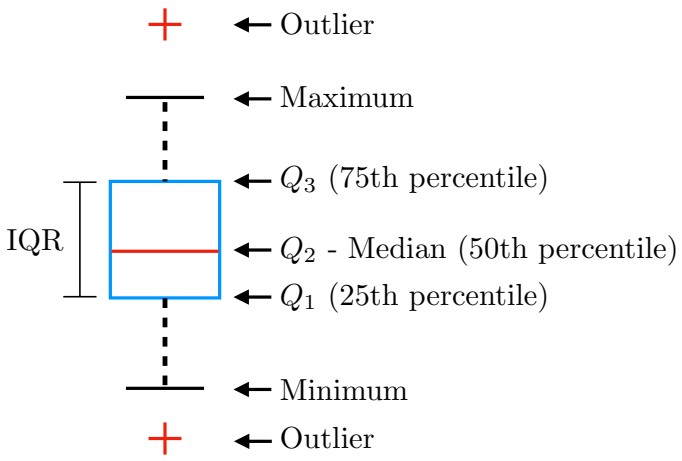



Figure 6.7: Boxplot of data.

Figure 6.8 depicts the normalized distributions of amplitudes of each TFR, in dB. Regarding the STFTs, the results show the increase in dispersion caused by increasing the length of the analysis window. One can see that the energy smearing is more pronounced at upper harmonics, since they present steeper slopes, causing this higher dispersion and tendency of amplitude loss.

Among all the combination methods, the SWGM presented the best relative amplitude preservation, also having the smallest dispersions. This is reasonable, since every time-frequency bin suffers the effect of all input representations, unlike the other methods, which tend to favour the best input representation for each bin; the loss in sharpness of the frequency peaks, i.e. large bandwidths, does not mean a lack of precision in representing amplitudes.

As for the SLS combination, using the FEMD method contributed to some reduction in the overall dispersion. The MRFCI presented a larger dispersion, which also increased with the harmonic index, and there is also a slight tendency of amplitude attenuation in upper harmonics, although much smaller than for the STFT-3. In fact, the median magnitudes of all TFRs were within a 1 dB deviation, which

shows that there is practically no tendency of systematically changing the timbre some way.

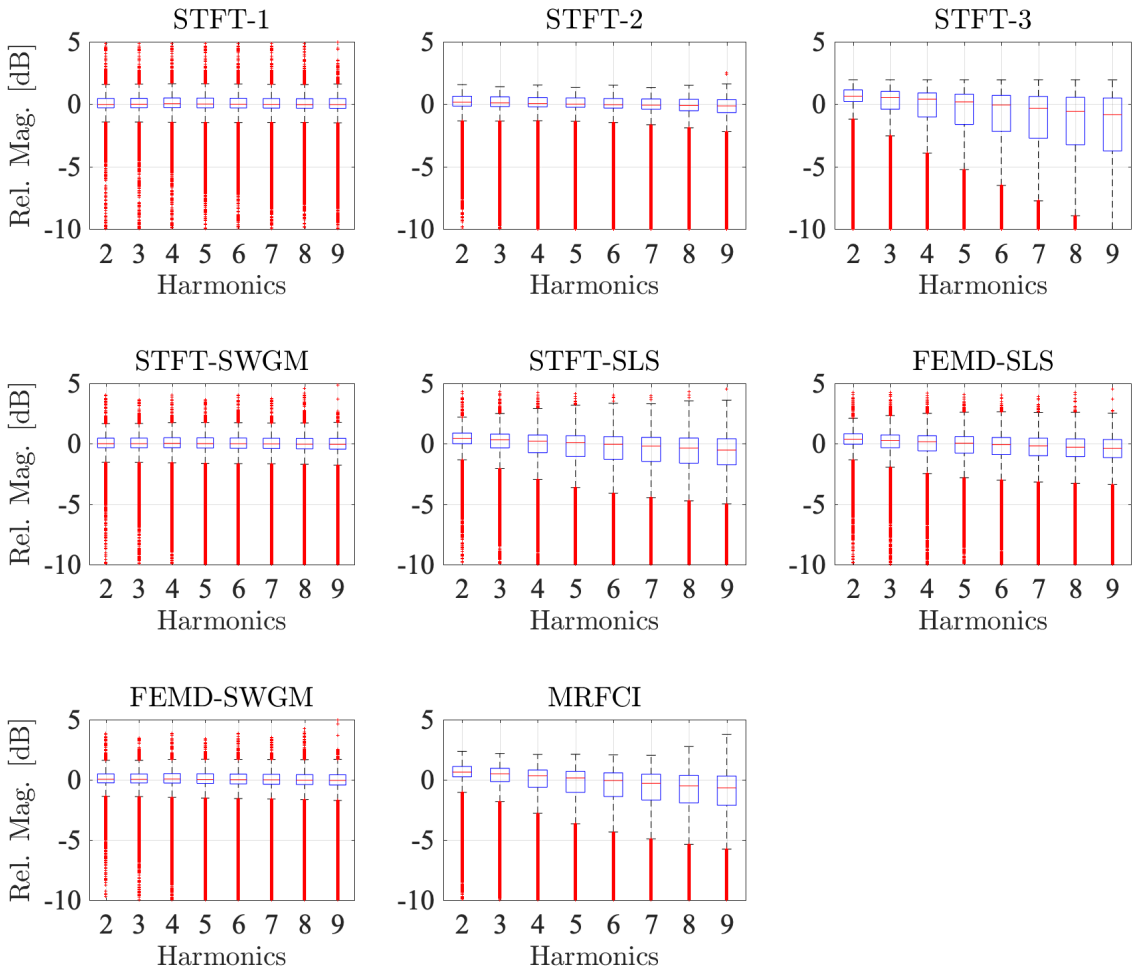



Figure 6.8: Distributions of amplitudes of each harmonic.

### Onsets

This next experiment assesses the behaviour of the methods at onsets. Two types of curves were computed centered at the onsets: the average energy (sum of all frequency bins) of the time frames; and the average indicator function (sum of all frequency bins of the differential spectrogram[4]). Since this dataset is comprised of sinusoids with fixed amplitude and the non-voiced excerpts are muted, the energy per frame should theoretically form a step at the onsets.

Figure 6.9 depicts the average energy per frame at onsets for each TFR. The results show that, as expected, the STFT-1 spectrogram is the one that best follows the aforementioned behaviour. This curve is nearly identical to the one yielded for the MRFCI, which is by far the best one among all the combination methods. The

[4]The differential spectrogram is computed by the first order differentiation time-wise.

results regarding the SLS combinations follow the same sharp slope of the MRFCI, but present an energy bump after the onset, followed by an energy stabilization towards the correct amplitude. The SWGM combinations resulted in even more abrupt slope and higher energy peak after the onset, rapidly decreasing and matching with the SLS curves.

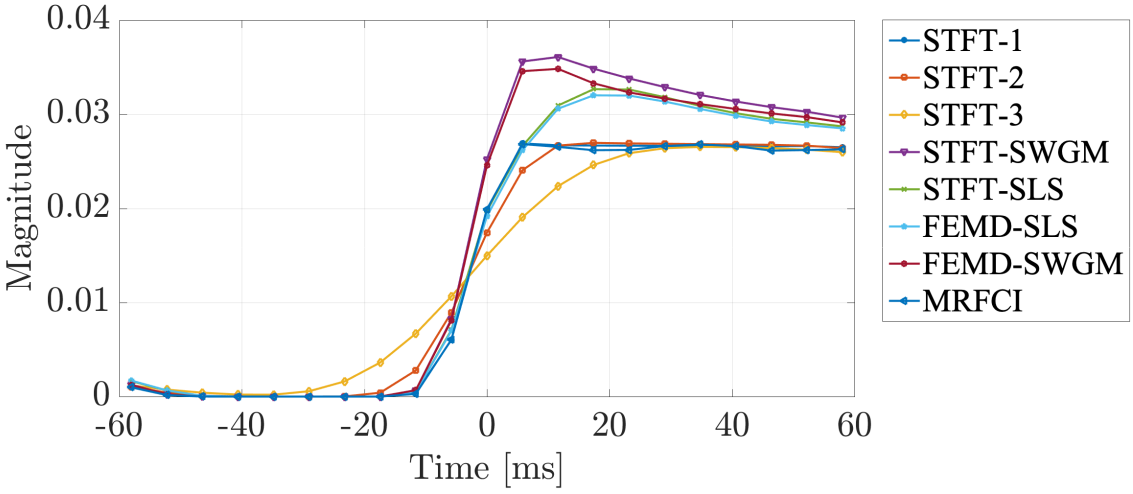



Figure 6.9: Average of the energy function at onsets for each TFR.

The normalized indicator functions, depicted in Figure 6.10, reflect a similar behaviour of the representations, showing the MRFCI curve matching the STFT-1, the SLS curves taking more frames to decay, and the SWGM curves decaying towards a more pronounced valley after the onset and then converging to 0.

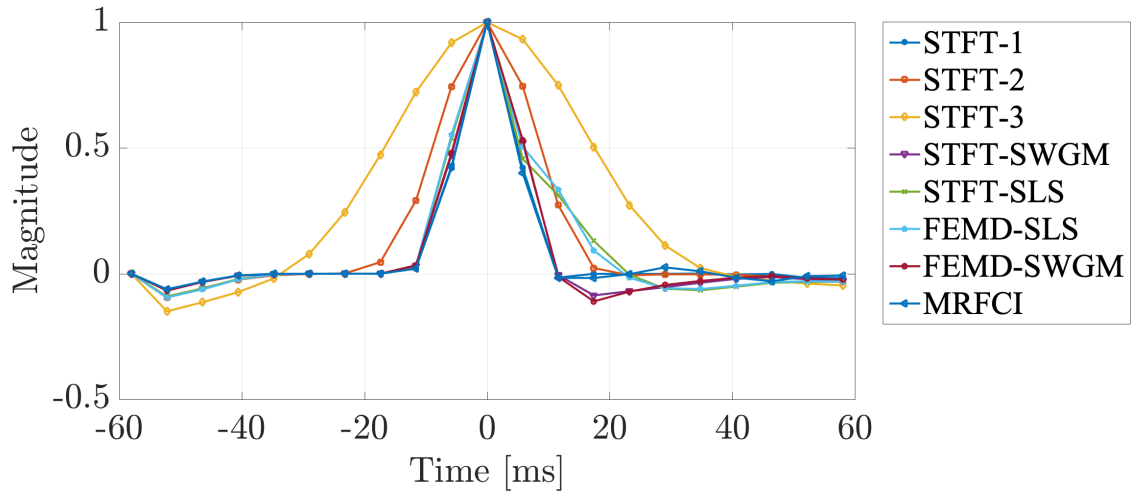



Figure 6.10: Average of the indicator function at onsets for each TFR.

From the resulting curves, one can conclude that the MRFCI tends to best preserve the energy evolution in such transient regions, for this particular dataset. Although this experiment only uses one particular type of onset, the intrinsic nature

of the MRFCI method and the results obtained with this experiment suggest that it may also provide the best representation of general transient information.

### 6.1.3 Concluding Remarks

In summary, the STFT-3, STFT-SLS, FEMD-SLS and MRFCI spectrograms have performed similarly in terms of best bandwidth, while the MRFCI and the FEMD-SLS spectrograms performed best in terms of dynamic range, considering both datasets. Regarding the preservation of peak amplitude, the SWGM combinations yielded the most concentrated distributions, while the MRFCI had the most scattered patterns, especially in upper harmonics, although no significant bias could be observed. As for the onsets, the MRFCI method showed the best results, nearly identical to the STFT-1.

Considering all these results, one can state that, among this particular pool and for these figures-of-merit, the MRFCI and the FEMD-SLS methods have performed best. The computational cost of the MRFCI method, on the other hand, is much smaller[5] than the cost of the FEMD-SLS, being comparable to the FEMD-SWGM, which makes the MRFCI the most suitable method for representing melodic signals. We believe that the combination of only the STFT-2 and the STFT-3 spectrograms would lead to better results for the SWGM method in terms of bandwidth and dynamic range, but then we would have to include as well another version of the other methods without the STFT-1, and the comparison would be too extensive. Besides, the other methods benefit from the presence of STFT-1.

It is worth commenting one more time that, depending on the circumstances, computational burden can be a bottleneck, e.g. when dealing with large datasets, hence using light combinations may be the best choice for such situations. Another limitation is that the SLS and MRFCI methods are designed to only provide representations with linear frequency resolution, while the SWGM can be applied to any set of representations having the same time-frequency dimensions. In situations where log-frequency spectrograms are required, SWGM combinations may be the only way to improve such representations. It seems possible to adapt the other methods to provide spectrograms with non-linear frequency resolution, but this work is yet to be done.

[5]This happens primarily due to the computation of the sparsity for each time-frequency bin, which envolves sorting all the samples around the bin under analysis.

## 6.2 Dominant Melody Estimation

This section contains a brief experiment for which a state-of-the-art system [40, 41] for dominant melody estimation was used. The experiment was conducted in collaboration with G. Doras (IRCAM)[6], who is the main author of [40, 41]. Here, the idea was to assess the performance of this system using different TFRs as input, maintaining the exact same architectural conditions. For this reason, we had to deal with several constraints, which limited our options of usable representations.

In [41], the *U-Net* [28, 40, 41, 43, 48–50] architecture is used in conjunction with a sequential method to train the network using ground truth data at increasing resolutions. In this architecture, the system first provides (at the descending branch of the U), via downsampling, coarser resolutions of the input information to be learned by means of convolution and pooling layers; and then (at the ascending branch of the U), upsampling is used to allow the network to learn to recreate representations at finer resolutions by means of convolution and transposed convolution layers.

As input, the original system receives HC$Q$T tensors of the audio files. Since the number of frequency bins is limited in 360, for computational reasons, it is impossible to use linear frequency spectrograms spanning the same frequency band and having usable frequency resolution.[7] Besides, it seems that only other log-frequency spectrograms are suitable for substituting the HC$Q$T in this case, for their interesting geometrical properties (see Chapter 2), which are exploited by convolutional networks. The following TFRs were tested as input of this system, all with 6 harmonic layers with indexes $i_h \in \{0.5, 1, 2, 3, 4, 5\}$:

**HCQT** – Harmonic C$Q$T set of spectrograms computed with $b = 60$ bins/octave;

**HCQT-SWGM** – three different sets of harmonic C$Q$T spectrograms computed with $b \in \{60, 40, 20\}$ bins/octave, combined using the SWGM method ($\beta = 0.5$);

**HVQT-SWGM** – three different sets of harmonic V$Q$T spectrograms [79], computed with $b = 60$ bins/octave and parameter $\kappa \in \{4, 17, 30\}$ (see Equation 2.21), combined using the SWGM method ($\beta = 0.5$).

Figure 6.11 depicts examples of the representations adopted for this experiment. As discussed in Chapter 2, the C$Q$T produces excessively smeared results for the low frequency range, due to its high frequency resolution in this region. The combination of C$Q$Ts with different resolutions provides a better overall definition of the non-stationary frequency components, at the expense of generating less pronounced

[6] `https://www.ircam.fr/person/guillaume-doras/`

[7] An attempt of using a harmonic version of the MRFCI was made, without success.

peaks. The V$Q$T produces representations which smoothly change resolution with the frequency, leaving the high frequency components unaffected (in comparison to the C$Q$T) and providing a better time definition in the low frequency range. By combining such representations using the SWGM, an even better time-frequency resolution is achieved, also at the expense of loosing some definition of frequency peaks.

To train the network, the MedleyDB [124] dataset was used, as in [41][8]. The system was retrained with each of the representations, using the 10-fold strategy. Figure 6.12 depicts the average performances of the system measured using as figure-of-merit the melody voicing recall (VR), voicing false alarm (VFA), raw pitch accuracy (RPA), raw chroma accuracy (RCA), and overall accuracy (OA) scores, as provided by the `mir_eval` toolbox (see [125] for details).

As can be seen, the differences in the results are not sufficient to attest an increase in performance and they suggest that the system was unable to take full advantage of the more detailed frequency lines provided. In fact, as discussed before, the SWGM combination provides a better overall representation at the expense of flattening the frequency peaks; for signals with a more static $f_0$, the C$Q$T should provide a more pronounced representation of the frequency peaks, and thus it might be a better TFR in such cases, while, for signals with a more dynamic $f_0$, the combinations may provide better results. This may be the explanation for the decrease in dispersion of the average results for the combined TFRs in terms of overall accuracy.

It is worth mentioning that such kind of systems may provide very different results with small architectural adjustments and that maybe this architecture is not well suited to profit from this kind of enhancement in details of frequency lines, due to its internal process tailored to deal with low-resolution versions of the input. Nevertheless, further experiments shall be conducted with different architectures or with modified versions of this same system, which may exhibit a more significant difference in performance when using TFRs with enhanced time-frequency resolution and may allow the use of linear-frequency spectrograms.

[8] In our experiments, a more precise version of the annotations, which were not available during the time the experiments in [41], are used, which explain the discrepancies in the results obtained with the HC$Q$T representations.

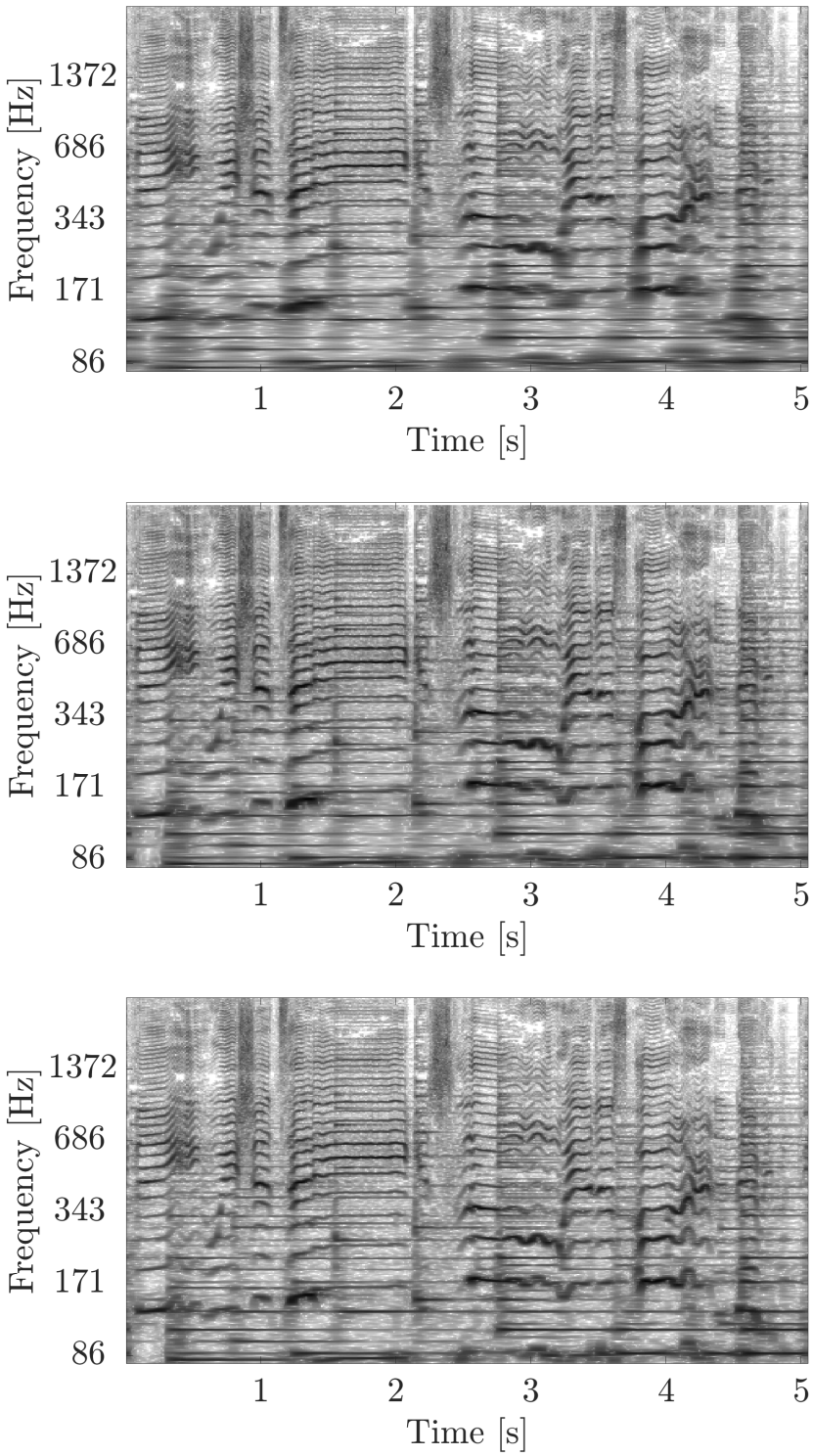



Figure 6.11: Examples of TFRs used in the experiment: C$Q$T, C$Q$T-SWGM and V$Q$T-SWGM, respectively.

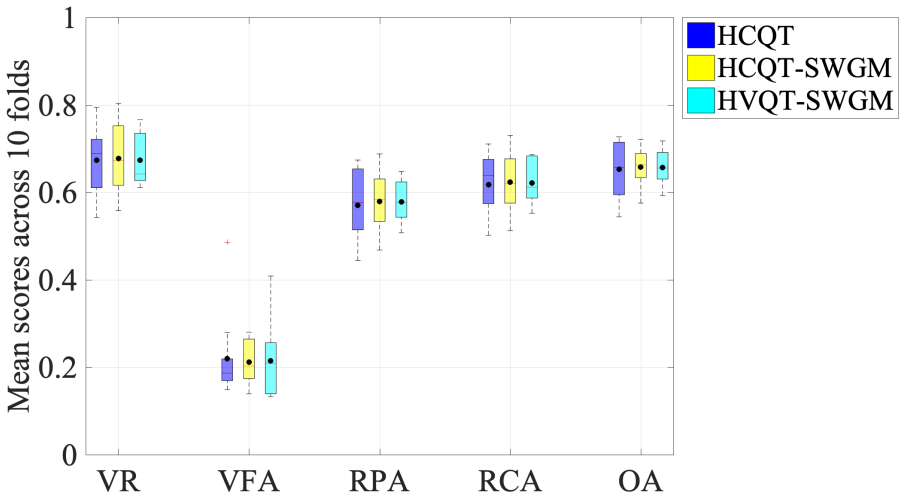



Figure 6.12: 10-folds mean scores distributions obtained with `mir_eval` using different TFRs: voicing recall (VR), voicing false alarm (VFA), raw pitch accuracy (RPA), raw chroma accuracy (RCA), and overall accuracy (OA).

# Chapter 7

# Automatic Percussion Transcription

Many percussion instruments can be played with different articulations,[1] producing a variety of tones, hence proper notation in this aspect is important in order to characterize the recorded instruments and extract useful information concerning the patterns played. To this end, it was proposed a method to automatically classify the onsets of percussion instruments, facilitating the annotation work by providing a good overall initial classification; the annotation procedure is then finalized by manually checking and correcting eventual mistakes. In this chapter, the automatic onset classification scheme will be described along with the BRID dataset [106], which comprises a set of recordings and annotations of percussion instruments playing different genres of traditional Brazilian music. As mentioned in Chapter 1, this work is related to the StaReL[2] project, and the dataset contents are available at `http://www02.smt.ufrj.br/~starel/datasets/brid.html`.

## 7.1 Onset Detection

The classification scheme uses onset annotations previously produced by means of a state-of-the-art method [126], a robust approach to onset detection designed to be applied to any type of music. Their method is based on auditory spectral features and bi-directional long short-term memory (Bi-LSTM) recurrent neural networks (RNN). The scheme is purely data driven, and yields high temporal precision as well as detection accuracy.

This method makes use of a network consisting of a concatenation of three hidden layers for each direction (6 layers in total) with 20 Bi-LSTM. The authors used two different-resolution mel-spectrograms, computed with analysis window sizes of 23.2 ms and 46.4 ms, respectively, and their first order difference as input represen-

[1] Here, in the context of percussion instruments, we consider the term articulation as being the type of hit produced, e.g. a *surdo* being struck with or without hand muffling.

[2] `www.smt.ufrj.br/~starel`

tations, in order to help the networks capture features precisely in both time and frequency. The implementations used are available in the `madmom` package.[3]

As output, a text file is generated, containing a list of time-stamps that indicate where onsets have been detected.

## 7.2 Classification Scheme

In order to perform the classification of the detected onsets, a list was made containing the number of classes (types of hit) for each kind of instrument. For instance, instruments like snare drum and *reco-reco* were considered to have only one type of articulation, while others, e.g. *pandeiro* or *tantã*, were set to have three. After this initial classification, depending on the pattern played in each recording, the number of classes may have to be manually changed. A more detailed list could also be produced to indicate the number of classes needed for each file individually.

Classification is performed in the time-frequency domain, using regions of the spectrogram as input to a non-supervised clustering procedure, namely, the $k$-means algorithm, which will be described in the next section. First, the signals under analysis are down-sampled to a sampling frequency of $F_\mathrm{s} = 11025$ samples per second. This reduces the number of samples to be considered in the spectrum and focuses only on the ones with higher energy. Since we are not interested in fine frequency resolution, the spectrograms are computed with $K = 256$ frequency samples and hop-size $h = 128$. This way, the channels of the DFT encompass relatively large frequency bands. By using this scheme, small variations in tuning and overall timbre (like those occurring when drums are muffled with the musician's hand), are neglected.

Then, the $k$-means classifier is fed with matrices containing the region of the spectrogram indicated by the onset detector. Such regions vary according to the type of instrument and include all the frequency samples, approximately from 5 ms before the onset to 200 ms after it. This time span has shown to be sufficient to allow the correct classification of the articulations. Also, a half Hamming window is applied to those matrices, so that the energy decreases in time, hence giving more weight to the earlier samples. In order to provide better synchronism, the regions are chosen so that the time frame with higher energy lays on the third position of the matrix. After that, the classification is performed and a column containing the class-index of each onset is aggregated to the list of onsets. Figure 7.1 depicts the classification procedure.

Since there is no kind of source separation procedure in this scheme, it is only useful for classifying tracks containing a single instrument, or at least with a large

[3]Madmom package version 0.16 [127].

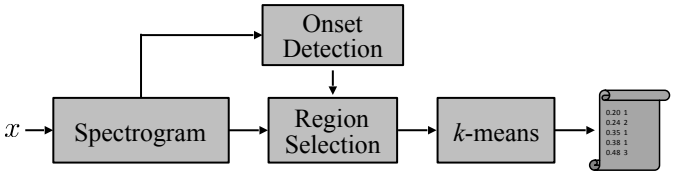



Figure 7.1: Flow-chart of the onsets classification scheme.

predominance of one instrument in terms of overall gain.

## 7.3 $k$-Means Clustering

The most ubiquitous algorithm known as the $k$-means[4] is a simple strategy of iterative refinement, based on two main steps [128]: assignment of the samples, which will be denoted as $\mathbf{s}$, to clusters; and update of the cluster centroids. These two steps are performed until the centroid updates no longer result in changes.

Let $\mathbf{u}_i^n$ be the centroid of cluster $i$ at $n$-th iteration. The algorithm[5] starts by setting $B$[6] centroids (or means) randomly in the space: $\mathbf{u}_1^1, \mathbf{u}_2^1, \mathbf{u}_3^1, \ldots, \mathbf{u}_B^1$; then, the iterations are performed as follows:

**Assignment step** – all samples in the set are assigned to its nearest centroid according to the Euclidean distance, forming clusters

$$\mathcal{U}_i^n = \{\mathbf{s}_l : \|\mathbf{s}_l - \mathbf{u}_i\| \leq \|\mathbf{s}_l - \mathbf{u}_j\| \ \forall j, 1 \leq j \leq B\}. \tag{7.1}$$

**Update step** – the centroids are updated according to the new set of clusters by averaging all its samples:

$$\mathbf{u}_i^{n+1} = \frac{1}{|\mathcal{U}_i^n|} \sum_{\mathbf{s}_l \in \mathcal{U}_i^n} \mathbf{s}_l, \tag{7.2}$$

where $|\mathcal{U}_i^n|$ denotes the number of elements in cluster $i$.

[4]It is also known as the Lloyd's algorithm, particularly in the computer science community.

[5]The Matlab implementation of this method was used in this work. For more details, see `https://www.mathworks.com/help/stats/kmeans.html`.

[6]Although originally the variable name $k$ is used to denote the number of clusters, which origins the name of the method, this same letter has been used previously in the text to denote the frequency index in spectrograms, hence the adoption of other letter in this context. Thus, strictly, we are talking about $B$-means.

## 7.4 The Brazilian Rhythmic Instruments Dataset

This section briefly describes the Brazilian Rhythmic Instruments Dataset (BRID), a copyright-free dataset for research within the MIR community, published in [106].[7] This is the dataset for which the automatic annotation procedure was designed. It is comprised of 274 solo- and 93 multiple-instrument recorded tracks of 10 different instrument classes with variations (e.g. material, size, stick) playing in 5 main rhythm classes (*samba*, *partido alto*, *samba-enredo*, *capoeira*, and *marcha*). Here, the dataset recording process and content are described. At the time of writing, annotations of beat and downbeat have been made, and the ones concerning type of articulation are still being manually checked.

## 7.5 Dataset

The BRID was originally developed in the context of sound source separation [129], but its applicability can be extended to other areas, computational rhythm analysis in particular. The dataset contains 367 short tracks of around 30 s on average, totaling 2 hrs 57 min. The tracks consist of recordings of solo or multiple instruments, playing characteristic Brazilian rhythms.

The recordings present instruments played in different Brazilian rhythmic styles. Although *samba* and two sub-genres (*samba-enredo* and *partido alto*) have been favored, BRID also features *marcha*, *capoeira*, and a few tracks of *baião* and *maxixe* styles. The number of tracks per rhythm is summarized in Tables 7.1 and 7.2.

All featured rhythms are in duple meter. *Samba* is specially known for this type of bar division and for the accentuation of the second beat [130]. Only combinations of instruments and rhythms that are traditionally seen in Brazilian music were considered, to provide a faithful representation of each rhythm.

### 7.5.1 Instruments

The recorded instruments were selected among the most representative ones in Brazilian music, more specifically in *samba* music. Ten different instrument classes were chosen: *agogô*, *caixa* (snare drum), *cuíca*, *pandeiro* (tambourine), *reco-reco*, *repique*, shaker, *surdo*, *tamborim* and *tantã*. To provide a variety of sounds, both membranophones and idiophones were featured. Also, whenever possible, instruments were varied in shape, size, material (e.g., leather or synthetic drumhead),

[7] In this thesis, the description of the dataset is strongly based on [106], which is also one of its original contributions.

Table 7.1: Tempi/number of solo tracks per rhythm.

| *Rhythm* | Tempo (bpm) | # *Tracks* |
|---|---|---|
| *Samba* (SA) | 80 | 54 |
| *Partido alto* (PA) | 100 | 55 |
| *Samba-enredo* (SE) | 130 | 60 |
| *Marcha* (MA) | 120 | 27 |
| *Capoeira* (CA) | 65 | 12 |
| *Samba - virada* (VSA) | 75 or 80 | 3 |
| *Partido alto - virada* (VPA) | 75 or 100 | 36 |
| *Samba-enredo - virada* (VSE) | 130 | 17 |
| *Marcha - virada* (VMA) | 120 | 8 |
| Other (OT) | - | 2 |

Table 7.2: Number of multi-instrument tracks per rhythm.

| *Rhythm* | # *Tracks* |
|---|---|
| *Samba* (SA) | 41 |
| *Partido alto* (PA) | 28 |
| *Samba-enredo* (SE) | 21 |
| *Marcha* (MA) | 3 |
| *Capoeira* (CA) | - |

pitch/tuning (e.g., in a *samba school*,[8] *surdos* are usually tuned in three different pitch ranges) and in the way they were struck (e.g., with the hand or with a wooden or a plastic stick), spanning a total of 32 variations. For example, the dataset features two *caixa* variations (12" in diameter with either 4 or 6 snare wires), six *pandeiro* variations (either 10", 11" or 12" in diameter with a leather or nylon drumhead) and three *tamborim* variations (one with a leather head struck with a wooden stick, and another one with a nylon head struck with either a wooden or a plastic stick[9]). Figure 7.2 shows real pictures of the instrument classes considered.

### 7.5.2 Dataset Recording

All the recordings were made in a professional recording studio in Manaus, Brazil, between October and December of 2015. The recording room has rectangular shape with dimensions of 4.3 m×3.4 m×2.3 m and is acoustically treated with a combination of wood and acoustic foam.

Both microphone model and positioning were optimized to translate the sound of each instrument as naturally as possible in the recording, considering the instrument size and the room acoustics. Most instruments were recorded with dynamic microphones within a distance of around 20 cm. The digital files were recorded with

[8] A popular association for the practice of *samba*. *Samba schools* are usually strongly connected to a specific community, where their social events take place and to whom they provide several social services. The climactic event for *samba schools* is the annual carnival parade, when imbued with communal effort they compete for the title.

[9] A leather-head *tamborim* is not played with a plastic drum stick.

(a) *Agogô* (b) *Caixa* (c) *Cuíca* (d) *Pandeiro* (e) *Reco-reco*

(f) *Repique* (g) Shaker (h) *Surdo* (i) *Tamborim* (j) *Tantã*

Figure 7.2: Instrument classes.

a sampling rate of 44.1 kHz and 16-bit resolution.

There are two groups of tracks in the dataset. The first one consists of instruments recorded solo, with the musicians performing in various Brazilian styles following a metronome track. Three musicians were recorded separately, each playing around 90 different instrument–rhythm combinations. For each instrument class, there is at least one track that consists of a *virada* of one of the main rhythms.[10] These are free improvisation patterns (still subject to the metronome track), which are very common in *rodas de samba*.[11] It is worth mentioning that the musicians brought their own instruments for the recording sessions. Although the general characteristics of each instrument are the same, e.g., size and type of material, subtle differences in construction bring additional timbre variability to the dataset.

The second set of tracks of the dataset gathers group performances, with the musicians playing together different rhythmic styles without a metronome reference. The instruments were individually captured with directional microphones, which were strategically positioned to minimize sound bleed, and two condenser microphones in omni polar pattern captured the overall sound in the room. The performances were designed to emulate typical arrangements of each style. Following this procedure, 19 recordings were made with four musicians, 29 with three musicians, and 45 with two musicians playing at a time.

[10] Except for shaker tracks.

[11] A small and informal gathering to play and dance to *samba* music. It is a common practice highly characterized by improvisation where musicians and dancers interact and learn with one another.

## 7.6 Examples

The articulation of some instruments, e.g. the *agogô*, can be very easy to classify. The *agogô* is comprised of a set of bells, whose tonal signatures are roughly invariant and very distinct from one another, which facilitates the clustering procedure. Figure 7.3 depicts a classification obtained for an *agogô* recording, where the numbers at the top and the colours are related to the cluster indexes. In this case, the *agogô* has two different bells, whose sounds were correctly identified for all the 93 onsets present in the recording. Figure 7.4 depicts one example of the time-frequency region for each articulation, or sound of each bell. As can be observed, the sound produced is strongly tonal and different frequencies are excited in each example.

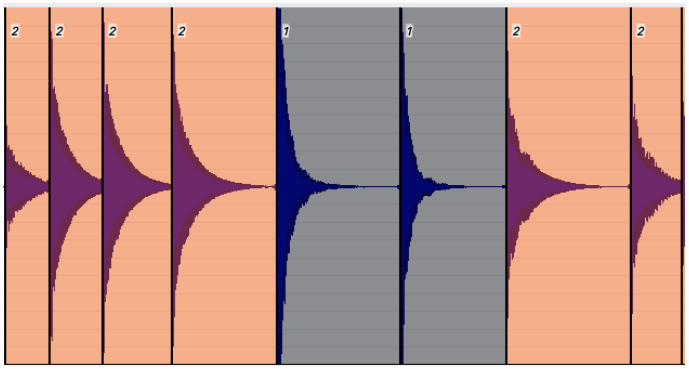



Figure 7.3: Classification of different articulations on an *agogô* recording.

However, for other instruments, identifying the correct class for each hit is not that easy, due to the variability inherent to the playing techniques and strong similarities in timbre. For instance, instruments which rely on a muffling technique to produce different articulations may produce a very wide variety of sounds, and playing consistently in terms of timbre can be very challenging. This is the case of the *pandeiro* (or tambourine), which is a complex instrument comprised of a tensioned drumhead (typically made of leather) and metal jingles. It can be played in many different ways and a very wide variety of sounds can be produced by it.

A common *pandeiro* pattern in *samba* music is played by repeating the cell of articulations $[1, 2, 3, 2]$, where such articulations are produced by: (1) hitting the drumhead with the thumb, (2) hitting the drumhead with the tip of the fingers while muffling it using the hand that is holding the *pandeiro*, and (3) hitting the drumhead with the bottom of the hand, near the fist, while muffling it. This pattern can be seen in Figure 7.5, where the recorded signal is illustrated along with the indications of each articulation, in numbers and colours.

The first articulation produces a sound comprised of the resonance of the leather and a soft jingle from the metal parts, while the articulations 2 and 3 produce very similar sounds, having a fast decay resonance of the skin and along with the shimmer

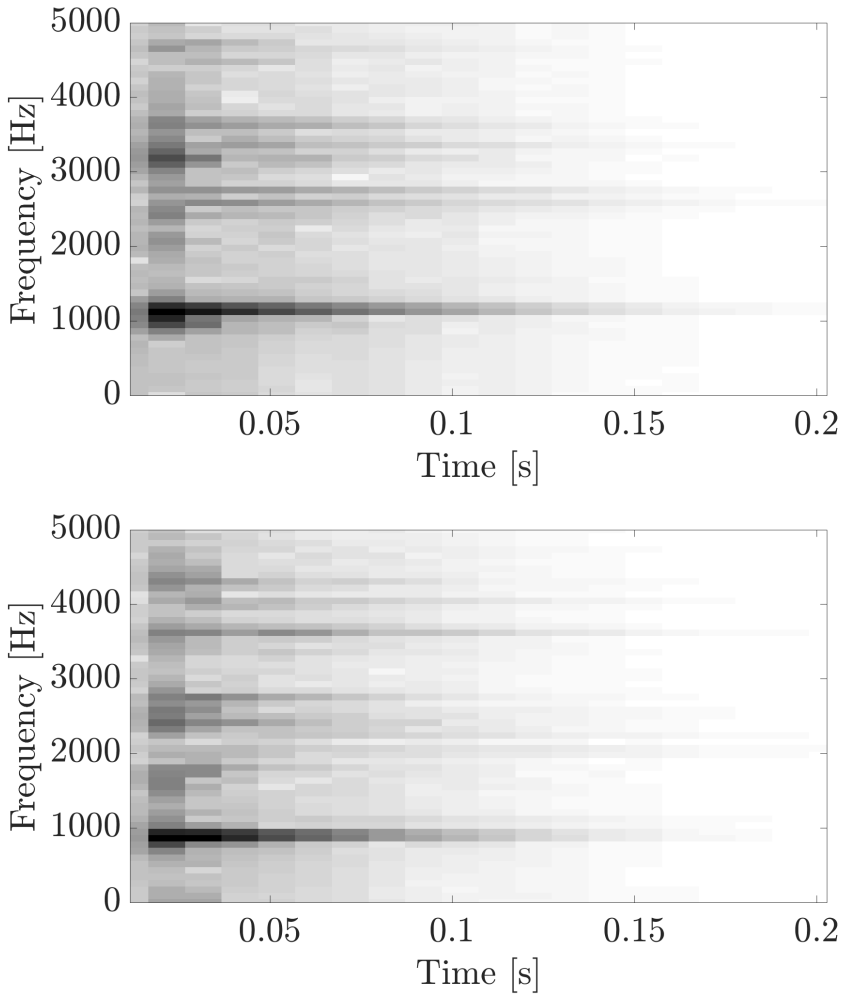



Figure 7.4: Samples of the two different articulations of an *agogô* recording.

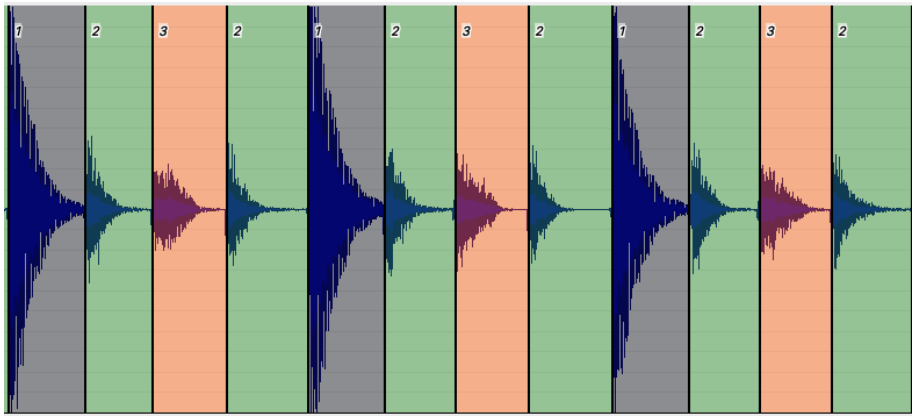



Figure 7.5: Classification of different articulations on a *pandeiro* recording.

of the metal jingles. Since the muffling changes the tuning of the drumhead, the hits using this technique may vary in terms of tonal content, and the randomness inherent to the sound of the metal jingles produces noise-like patterns in the time-frequency domain, which are also very variable. Figure 7.6 illustrates the time-frequency regions of one example of each articulation aforementioned. As can be

observed, the energy concentration in low frequencies, related to the resonance of the leather, is extended in time, and it comprises a large proportion of the overall energy; at the other articulations the energy in low frequencies lasts much shorter and there is more energy caused by the metal jingles spread through the spectrum. Since the articulations 2 and 3 are very similar, for this example, two onsets of class 2 were classified as 3, which provided an accuracy of 98.5%.

This procedure was applied to all the recordings of solo instruments, and thus good initial classifications were provided. Considering all the files and the percentage of signals which were much harder to correctly classify, the overall accuracy was lower than it was for the examples presented. The procedure provided, at least, an overall accuracy of around 75-80%. This estimate was made by considering only the corrections in terms of classes, and not inclusion or removal of onsets. As a result, the effort required to produce the final annotations was significantly reduced, although the manual checking and correcting stage was still necessary.

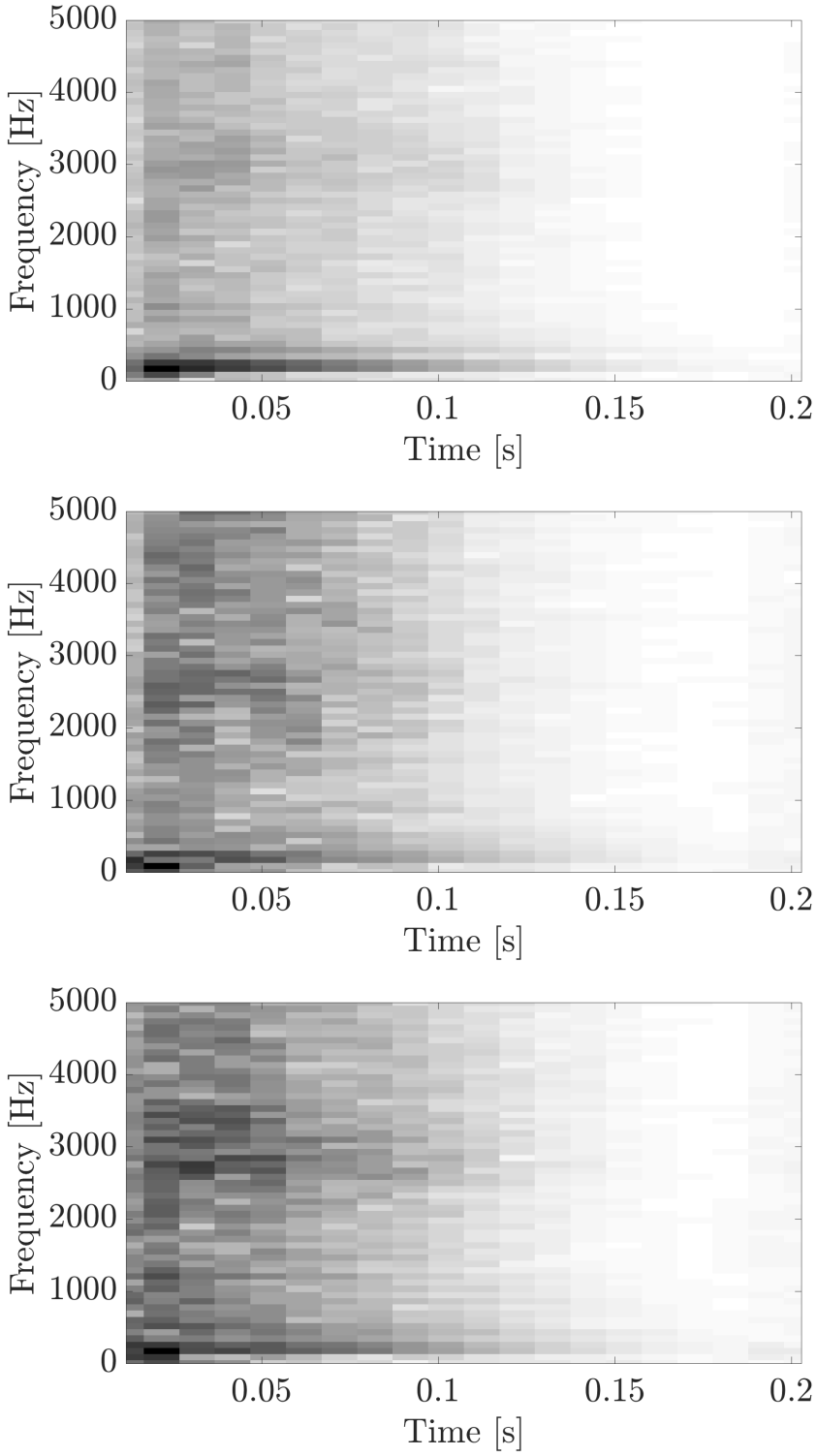



Figure 7.6: Samples of articulations 1–3 of a *pandeiro* recording, respectively.

# Chapter 8

# Conclusions and Future Work

## 8.1 Conclusions

In this thesis we have addressed the problem of generating high-resolution TFRs, with a special attention to polyphonic signals containing fast frequency variations. All the methods presented are based on the idea of combining TFRs with different resolutions to produce a TFR with high energy concentration wherever possible in the time-frequency plane. The methods were classified by their general approach, being: bin-wise combinations, combinations based on local information, and methods based on image analysis.

The combination methods were compared using synthetic and real-world signals, in an attempt to characterize their general behaviour, especially when dealing with signals of main melody. First, the methods were compared within their own class, and then more precise and extensive experiments were performed with a select group containing some of the best methods of each kind.

The preliminary results show that, among the bin-wise combination methods, the sample-weighted geometric mean (SWGM) outperforms the others, providing the best overall combination of peaks in both time and frequency, followed by the reciprocal mean. All methods of this nature present the advantage of requiring relatively low computational power and the disadvantage of having some peak distortion, which may be a problem for certain systems.

Among the combinations based on local information, the local sparsity (LS) method have shown to be a better solution than the Lukin-Todd's (LT) method in terms of time precision, due to the use of an asymmetrical window for the local energy compensation. The smoothed local sparsity (SLS) method, which is a smooth version of the LS, tends to smooth out the transitions between different representations and artifacts that are caused by such hard transitions. These methods require a very high computational power due to the calculations of the local features, but

provide significantly better results than the bin-wise combinations, as they tend to optimize the local energy concentration.

As for the methods based on image analysis, the frame-based method for estimation of main directions (FEMD) provides excellent overall results, and has the advantage of accumulating the information within the frames to compute the FChTs using the most relevant chirp directions. Since it relies on a combination method, e.g. the SLS, its computational burden and final resolution depend on it. A limitation of this method is the need for the configuration of the number of instances of FChT to be computed. The multi-resolution fan-chirp interpolation (MRFCI) method has shown to be more flexible and much less computationally heavy, when compared to the FEMD used in conjunction with the SLS, providing similar or even better results in some cases, without the necessity of configuring the system according to the signal to be analyzed.

In the experiments using main melody datasets with precise $f_0$ annotations, the STFT-SLS, FEMD-SLS and MRFCI spectrograms have shown to provide the narrowest bandwidth; the MRFCI and the FEMD-SLS spectrograms provided the largest dynamic range; the SWGM combinations yielded the most concentrated distributions regarding the preservation of peak amplitude; and the MRFCI method provided the best results in terms of onset definition.

Considering all figures of merit, the MRFCI and the FEMD-SLS have shown to be the best methods of this set. Nevertheless, since computational burden can be a bottleneck for scalability, using bin-wise combinations may be preferable in some circumstances, also due to their ability to combine representations with non-linear frequency resolution, e.g. the C$Q$T.

In the experiment using the *U-net* system for dominant melody detection, only TFRs with log-frequency scale could be used, which limited the range of methods for this task. The results have shown no relevant improvement in performance when using the representations combined with the SWGM. It seems like this system was unable to profit from the increase in definition of the frequency lines, maybe due to the characteristics of the architecture adopted or to the fact that the SWGM combination flattens the frequency peaks.

The method for automatic classification of types of hit proved very handy, facilitating the manual effort of annotating the rhythmic patterns contained in the BRID dataset. The precise annotations regarding onset and types of hit will allow for studies of microtiming and pattern characterization of Brazilian traditional rhythms.

## 8.2 Future Work

The future works have been duly mentioned in the text, in their own contexts. Here, they are summarized in the following. This work focused on providing magnitude spectrograms of different natures for audio signal analysis. Though, in the context of audio manipulation, one must be able to reconstruct the signal back to time domain. Future research can be conducted in order to provide the methods with such capability. Also, a lighter version of the LS and SLS methods may be viable by downsampling the region around the time-frequency bins in such a way that the sparsity is preserved. This could considerably reduce the computational burden of those methods, making them more suitable for practical applications.

Regarding the experiment with dominant melody detection, some different neural-network architectures may be tested or developed to be able to use high-resolution TFRs such as the ones presented in this work. Furthermore, procedures such as the structure tensor, the combinations based on local information, and the fan-chirp transform might could be modified to work properly with log-frequency spectrograms, which would allow the whole set of proposed methods to be used with a much larger set of methods for MIR.

# References


[1] SCHOENHERR, S. "Recording Technology History". `http://www.aes-media.org/historical/html/recording.technology.history/notes.html`, 2005. Accessed: 2020-02-25.

[2] BITTNER, R. M. *Data-Driven Fundamental Frequency Estimation.* PHD Thesis, New York University, New York, United States of America, 2018.

[3] NISHIKIMI, R., NAKAMURA, E., GOTO, M., et al. "Scale- and Rhythm-Aware Musical Note Estimation for Vocal F0 Trajectories Based on a Semi-Tatum-Synchronous Hierarchical Hidden Semi-Markov Model". In: *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, October 2017.

[4] MÜLLER, M. *Fundamentals of Music Processing.* Berlin, Germany, Springer Verlag, 2015.

[5] DIXON, S. "Onset Detection Revisited". In: *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx)*, Montreal, Canada, September 2006.

[6] OJIMA, Y., ITOYAMA, K., YOSHII, K. "A Hierarchical Bayesian Model of Chords , Pitches , and Spectrograms for Multipitch Analysis". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York, United States of America, August 2016.

[7] SCHLÜTER, J., BÖCK, S. "Improved Musical Onset Detection With Convolutional Neural Networks". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[8] ROCAMORA, M., CANCELA, P. "Pitch tracking in Polyphonic Audio by Clustering Local Fundamental Frequency Estimates". In: *Anais do 9o Congresso de Engenharia de Áudio da AES-Brasil*, São Paulo, Brasil, May 2011.

[9] BARANIUK, R. G., FLANDRIN, P., JANSSEN, A. J. E. M., et al. "Measuring time-frequency information content using the Renyi entropies", *IEEE Transactions on Information Theory*, v. 47, n. 4, pp. 1391–1409, May 2001.

[10] MAUCH, M., DIXON, S. "Simultaneous Estimation of Chords and Musical Context From Audio", *IEEE Trans. on Audio, Speech, and Language Processing*, v. 18, n. 6, pp. 1280–1289, August 2010.

[11] OZEROV, A., VINCENT, E., BIMBOT, F. "A general flexible framework for the handling of prior information in audio source separation", *IEEE Transactions on Audio, Speech, and Language Processing*, v. 20, n. 4, pp. 1118–1133, May 2012.

[12] LIDY, T., JR., C. N. S., CORNELIS, O., et al. "On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-Western and ethnic music collections", *Signal Processing*, v. 90, n. 4, pp. 1032–1048, April 2010.

[13] ABESSER, J., CANO, E., FRIELER, K., et al. "Score-Informed Analysis of Intonation and Pitch Modulation in Jazz Solos". In: *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, October 2015.

[14] BÖCK, S., KREBS, F., WIDMER, G. "A Multi-model Approach to Beat Tracking Considering Heterogeneous Music Styles". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, October 2014.

[15] EWERT, S., MÜLLER, M. "Estimating note intensities in music recordings". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czechia, May 2011.

[16] ROCAMORA, M., JURE, L., BISCAINHO, L. W. P. "Tools for detection and classification of piano drum patterns from Candombe recordings". In: *Proceedings of the 9th Conference on Interdisciplinary Musicology (CIM)*, Berlin, Germany, December 2014.

[17] HAMANAKA, M., HIRATA, K., TOJO, S. "Musical Structural Analysis Database Based on {GTTM}". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, October 2014.

[18] BÖCK, S., KREBS, F., WIDMER, G. "Joint Beat and Downbeat Tracking with Recurrent Neural Networks". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York, United States of America, August 2016.

[19] GOUYON, F., DIXON, S., PAMPALK, E., et al. "Evaluating rhythmic descriptors for musical genre classification". In: *Proceedings of the 25th AES International Conference*, London, United Kingdom, June 2004.

[20] SERRA, X. "A multicultural approach in music information research". In: *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, United States of America, October 2011.

[21] PAULUS, J., KLAPURI, A. "Measuring the Similarity of Rhythmic Patterns". In: *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, Paris, France, October 2002.

[22] VELARDE, G., WEYDE, T., CHACÓN, C. C., et al. "Composer Recognition Based on 2D-Filtered Piano-Rolls". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York, United States of America, August 2016.

[23] DITTMAR, C., CANO, E., ABESSER, J., et al. "Music information retrieval meets music education", *Multimodal Music Processing*, v. 3, n. 1, pp. 95–120, 2012.

[24] DURAND, S., BELLO, J. P., DAVID, B., et al. "Robust Downbeat Tracking Using an Ensemble of Convolutional Networks", *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, v. 25, n. 1, pp. 76–89, January 2017.

[25] MADDAGE, N. C. "Automatic Structure Detection for Popular Music", *IEEE MultiMedia*, v. 13, n. 1, pp. 65–77, January 2006.

[26] DIXON, S. "Extraction of musical performance parameters from audio data". In: *Proceedings of the First IEEE Pacific Rim International Symposium on Dependable Computing*, Los Angeles, United States of America, December 2000.

[27] SCHEIRER, E. D. "Using musical knowledge to extract expressive performance information from audio recordings". In: *Proceedings of the IJCAI Workshop on Computational Auditory Scene*, Montreal, Canada, August 1995.

[28] ZIH-SING, F., SU, L. "Hierarchical Classification Networks for Singing Voice Segmentation and Transcription". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[29] ROMÁN, M. A., PERTUSA, A., CALVO-ZARAGOZA, J. "A Holistic Approach to Polyphonic Music Transcription with Neural Networks". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[30] JIANG, J., CHEN, K., LI, W., et al. "Large-vocabulary Chord Transcription Via Chord Structure Decomposition". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[31] CHOI, K., CHO, K. "Deep Unsupervised Drum Transcription". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[32] YCART, A., MCLEOD, A., BENETOS, E., et al. "Blending Acoustic and Language Model Predictions for Automatic Music Transcription". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[33] HOLZAPFEL, A., BENETOS, E. "Automatic Music Transcription and Ethnomusicology: a User Study". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[34] BENETOS, E., DIXON, S., DUAN, Z., et al. "Automatic Music Transcription: An overview", *IEEE Signal Processing Magazine*, v. 36, n. 1, pp. 20–30, January 2019.

[35] COGLIATI, A., DUAN, Z., WOHLBERG, B. "Piano Transcription with Convolutional Sparse Lateral Inhibition", *IEEE Signal Processing Letters*, v. 24, n. 4, pp. 392–396, April 2017.

[36] GOWRISHANKAR, B. S., BHAJANTRI, N. U. "An exhaustive review of automatic music transcription techniques: Survey of music transcription techniques". In: *Proceedings of the International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, Paralakhemundi, India, June 2016.

[37] BENETOS, E., DIXON, S., GIANNOULIS, D., et al. "Automatic music transcription: challenges and future directions", *Journal of Intelligent Information Systems*, v. 41, n. 3, pp. 407–434, December 2013.

[38] DESSEIN, A., CONT, A., LEMAITRE, G. "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence". In: *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, August 2010.

[39] KLAPURI, A. "Automatic transcription of music (2003)". In: *Proceedings of the Stockholm Music Acoustics Conference*, Stockholm, Sweden, August 2003.

[40] DORAS, G., PEETERS, G. "Cover Detection Using Dominant Melody Embeddings". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[41] DORAS, G., ESLING, P., PEETERS, G. "On the use of U-Net for Dominant Melody Estimation in Polyphonic Music". In: *Proceedings of the 2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*, Milano, Italy, January 2019.

[42] BOSCH, J. J., BITTNER, R. M., SALAMON, J., et al. "A comparison of melody extraction methods based on source-filter modelling". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York City, United States of America, August 2016.

[43] RIGAUD, F., RADENEN, M. "Singing Voice Melody Transcription using Deep Neural Networks". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York, United States of America, August 2016.

[44] BITTNER, R. M., SALAMON, J., ESSID, S., et al. "Melody extraction by contour classification". In: *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, October 2015.

[45] SALAMON, J., GOMEZ, E. "Melody extraction from polyphonic music signals: approaches, applications, and challenges", *IEEE Signal Processing Magazine*, v. 31, n. 2, pp. 118–134, March 2014.

[46] RAFII, Z., DUAN, Z., PARDO, B. "Combining rhythm-based and pitch-based methods for background and melody separation", *IEEE/ACM Transactions on Audio Speech and Language Processing*, v. 22, n. 12, pp. 1884–1893, December 2014.

[47] SALAMON, J., GÓMEZ, E. "Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics", *IEEE Transactions on Audio, Speech, and Language Processing*, v. 20, n. 6, pp. 1759–1770, Aug 2012.

[48] MESEGUER-BROCAL, G., PEETERS, G. "Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[49] JANSSON, A., BITTNER, R. M., EWERT, S., et al. "Joint Singing Voice Separation and F0 Estimation with Deep U-Net Architectures". In: *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, Coruña, Spain, September 2019.

[50] JANSSON, A., HUMPHREY, E., MONTECCHIO, N., et al. "Singing Voice Separation With Deep U-Net Convolutional Networks". In: *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, October 2017.

[51] YU, S., ZHANG, H., DUAN, Z. "Singing voice separation by low-rank and sparse spectrogram decomposition with pre-learned dictionaries", *Journal of the Audio Engineering Society*, v. 65, n. 5, pp. 377–388, May 2017.

[52] CHAN, T. S. T., YANG, Y. H. "Informed group-sparse representation for singing voice separation", *IEEE Signal Processing Letters*, v. 24, n. 2, pp. 156–160, January 2017.

[53] MIMILAKIS, S. I., DROSSOS, K., VIRTANEN, T., et al. "A Recurrent Encoder-Decoder Approach with Skip-Filtering Connections for Monaural Singing Voice Separation". In: *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Tokyo, Japan, September 2017.

[54] FÜG, R., NIEDERMEIER, A., DRIEDGER, J., et al. "Harmonic-Percussive-Residual Sound Separation Using the Structure Tensor on Spectrograms". In: *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 2016.

[55] IKEMIYA, Y., ITOYAMA, K., YOSHII, K. "Singing voice separation and vocal f0 estimation based on mutual combination of robust principal component analysis and subharmonic summation", *IEEE/ACM Transactions on Audio Speech and Language Processing*, v. 24, n. 11, pp. 2084–2095, November 2016.

[56] FAN, Z. C., JANG, J. S. R., LU, C. L. "Singing voice separation and pitch extraction from monaural polyphonic audio music via DNN and adaptive pitch tracking". In: *Proceedings of the IEEE 2nd International Conference on Multimedia Big Data (BigMM)*, Taipei, Taiwan, April 2016.

[57] IKEMIYA, Y., YOSHII, K., ITOYAMA, K. "Singing voice analysis and editing based on mutually dependent F0 estimation and source separation". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, South Brisbane, Australia, April 2015.

[58] COHEN, L. *Time-Frequency Analysis*. Englewood Cliffs, United States of America, Prentice Hall, 1995.

[59] AARABI, H. F., PEETERS, G. "Deep-Rhythm for Global Tempo Estimation in Music". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[60] BÖCK, S., DAVIES, M. E. P., KNEES, P. "Multi-Task Learning of Tempo and Beat: Learning One to Improve the Other". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[61] LEMAIRE, Q., HOLZAPFEL, A. "Temporal Convolutional Networks for Speech and Music Detection in Radio Broadcast". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[62] FUENTES, M., MAIA, L., ROCAMORA, M., et al. "Tracking Beats and Microtiming in Afro-Latin American Music Using Conditional Random Fields and Deep Learning". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[63] WIGGINS, A., KIM, Y. "Guitar Tablature Estimation with a Convolutional Neural Network". In: *Proceedings of the 20th International Society for*

*Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[64] CHOI, J., LEE, J., PARK, J., et al. "Zero-shot learning for audio-based music classification and tagging". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[65] PAUWELS, J., O'HANLON, K., GÓMEZ, E., et al. "20 Years of automatic chord recognition from audio". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[66] HUMPHREY, E. J., REDDY, S., SEETHARAMAN, P., et al. "An Introduction to Signal Processing for Singing-Voice Analysis: High Notes in the Effort to Automate the Understanding of Vocals in Music", *IEEE Signal Processing Magazine*, v. 36, n. 1, pp. 82–94, January 2019.

[67] BONADA, J. "A pitch salience function derived from harmonic frequency deviations for polyphonic music analysis". In: *Proceedings of the 11th Conference on Digital Audio Effects (DAFx)*, Espoo, Finland, September 2008.

[68] CANO, E., BEVERIDGE:, S. "Microtiming Analysis in Traditional Shetland Fiddle Music". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[69] GOUYON, F. "Microtiming in "Samba de Roda" — Preliminary experiments with polyphonic audio". In: *Proceedings of the 11th Brazilian Symposium on Computer Music (SBCM)*, São Paulo, Brazil, September 2007.

[70] APOLINÁRIO, I. F., BISCAINHO, L. W. P. "Fan-chirp transform with a timbre-independent salience applied to polyphonic music analysis". In: *Anais do XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, Santarém, Brasil, September 2016.

[71] IKEMIYA, Y., ITOYAMA, K., OKUNO, H. G. "Transcribing vocal expression from polyphonic music". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[72] DA COSTA, M. V. M., APOLINÁRIO, I. F., BISCAINHO, L. W. P. "Sparse Time-Frequency Representations for Polyphonic Audio Based on Com-

bined Efficient Fan-Chirp Transforms", *Journal of the Audio Engineering Society*, v. 67, n. 11, pp. 894–905, November 2019.

[73] DA COSTA, M. V. M., BISCAINHO, L. W. P. "High-Definition Time-Frequency Representation Based on Adaptive Combination of Fan-Chirp Transforms via Structure Tensor". In: *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx)*, Birmingham, United Kingdom, September 2019.

[74] DA COSTA, M. V. M., BISCAINHO, L. W. P. "Combining Time-Frequency Representations via Local Sparsity Criterion". In: *Proceedings of the 2nd AES Latin American Congress of Audio Engineering*, Montevideo, Uruguay, September 2018.

[75] DA COSTA, M. V. M., BISCAINHO, L. W. P. "Combining Time-Frequency Representations for Music Information Retrieval". In: *Anais do 15o Congresso de Engenharia de Áudio da AES-Brasil*, Florianópolis, Brazil, October 2017.

[76] LIN, R., CHUNHUI DU, LUO, S., et al. "Performance on a combined representation for time-frequency analysis". In: *Proceedings of the 2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, China, June 2017.

[77] APOLINÁRIO, I. F., BISCAINHO, L. W. P., ROCAMORA, M., et al. "Fan-chirp transform with nonlinear time warping". In: *Anais do 13o Congresso de Engenharia de Áudio da AES-Brasil*, São Paulo, Brasil, May 2015.

[78] HANSSON-SANDSTEN, M. "Evaluation of non-linear combinations of rescaled reassigned spectrograms". In: *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, September 2014.

[79] SCHÖRKHUBER, C., KLAPURI, A., HOLIGHAUS, N., et al. "A Matlab Toolbox for Efficient Perfect Reconstruction Time-Frequency Transforms with Log-Frequency Resolution". In: *Proceedings of the 53rd AES International Conference*, London, United Kingdom, January 2014.

[80] VAN DER SEIJS, M. *Improvements on Time-Frequency Analysis using Time-Warping and Timbre Techniques*. Master Dissertation, Delft University of Technology, Delft, Netherlands, 2011.

[81] SEJDIĆ, E., DJUROVIĆ, I., JIANG, J. "Time-frequency Feature Representation Using Energy Concentration: An Overview of Recent Advances",

*Digital Signal Processing: A Review Journal*, v. 19, n. 1, pp. 153–183, January 2009.

[82] XIAO, J., FLANDRIN, P. "Multitaper Time-Frequency Reassignment for Nonstationary Spectrum Estimation and Chirp Enhancement", *IEEE Transactions on Signal Processing*, v. 55, n. 6, pp. 2851–2860, June 2007.

[83] LUKIN, A., TODD, J. "Adaptive Time-Frequency Resolution for Analysis and Processing of Audio". In: *Proceedings of the 27th AES International Conference*, Paris, France, May 2006.

[84] SANG, T., WILLIAMS, W., O'NEILL, J. "An algorithm for positive time-frequency distributions". In: *Proceedings of the 3rd International Symposium on Time-Frequency and Time-Scale Analysis (TFTS)*, Paris, France, June 1996.

[85] AUGER, F., FLANDRIN, P. "Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method", *IEEE Transactions on Signal Processing*, v. 43, n. 5, pp. 1068–1089, May 1995.

[86] LOUGHLIN, P. J., PITTON, J. W., HANNAFORD, B. "Fast approximations to positive time-frequency distributions, with applications". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, United States of America, May 1995.

[87] LOUGHLIN, P., PITTON, J., HANNAFORD, B. "Approximating Time-Frequency Density Functions via Optimal Combinations of Spectrograms", *IEEE Signal Processing Letters*, v. 1, n. 12, pp. 199–202, December 1994.

[88] FONOLLOSA, J., NIKIAS, C. "A New Positive Time-frequency Distribution". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Adelaide, Australia, August 1994.

[89] PITTON, J. W., ATLAS, L. E., LOUGHLIN, P. J. "Applications of positive time-frequency distributions to speech processing", *IEEE Transactions on Speech and Audio Processing*, v. 2, n. 4, pp. 554–566, October 1994.

[90] FRAZER, G., BOASHASH, B. "Multiple View Time-Frequency Distributions". In: *Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, United States of America, August 1993.

[91] COHEN, L. "Positive Time-Frequency Distribution Functions", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 33, n. 1, pp. 31–38, February 1985.

[92] KIM, J. W., SALAMON, J., LI, P., et al. "CREPE: A Convolutional Representation for Pitch Estimation". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, April 2018.

[93] BITTNER, R. M., MCFEE, B., SALAMON, J., et al. "Deep Salience Representations for F0 Estimation in Polyphonic Music". In: *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, October 2017.

[94] SALAMON, J., BITTNER, R. M., BONADA, J., et al. "An Analysis/Synthesis Framework for Automatic F0 Annotation of Multitrack Datasets". In: *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, October 2017.

[95] DEMETRIOU, A., JANSSON, A., KUMAR, A., et al. "Vocals in Music Matter: the Relevance of Vocals in the Minds of Listeners". In: *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, September 2018.

[96] SCHRAMM, R., MCLEOD, A., STEEDMAN, M., et al. "Multi-Pitch Detection and Voice Assignment for a Cappella Recording of Multiple Singers". In: *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, October 2017.

[97] BALKE, S., DITTMAR, C., ABESSER, J., et al. "Data-driven solo voice enhancement for jazz music retrieval". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, United States of America, March 2017.

[98] NISHIKIMI, R. "Musical Note Estimation for F0 Trajectories of Singing Voices Based on a Bayesian Semi-Beat-Synchronous Hmm". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York, United States of America, August 2016.

[99] SCHLÜTER, J., GRILL, T. "Exploring Data Augmentation for Improved Singing Voice Detection With Neural Networks". In: *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, October 2015.

[100] MADHUSUDHAN, S. T., CHOWDHARY, G. "DeepSRGM - Sequence Classification and Ranking in Indian Classical Music Via Deep Learning". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019.

[101] DURAND, S., ESSID, S. "Downbeat Detection With Conditional Random Fields And Deep Learned Features". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York, United States of America, August 2016.

[102] DURAND, S., BELLO, J. P., DAVID, B., et al. "Downbeat tracking with multiple features and deep neural networks". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015.

[103] BÖCK, S., ARZT, A., KREBS, F., et al. "Online realtime onset detection with recurrent neural networks". In: *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx)*, York, United Kingdom, September 2012.

[104] KREBS, F., BÖCK, S., DORFER, M., et al. "Downbeat Tracking Using Beat Synchronous Features with Recurrent Neural Networks". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York, United States of America, August 2016.

[105] KASHIMA, K. L., MONT-REYNAUD, B. *The bounded-Q approach to time-varying spectral analysis*. Technical Report STAN-M-28, Stanford University, 1985.

[106] MAIA, L. S., DE TOMAZ JÚNIOR, P. D., FUENTES, M., et al. "A Novel Dataset of Brazilian Rhythmic Instruments and Some Experiments in Computational Rhythm Analysis". In: *Proceedings of the 2nd AES Latin American Congress of Audio Engineering*, Montevideo, Uruguay, September 2018.

[107] DINIZ, P. S. R., DA SILVA, E. A. B., NETTO, S. L. *Digital Signal Processing: System Analysis and Design*, v. 1. 2 ed. Cambridge, Cambridge University Press, 2010.

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

[109] CANCELA, P., LÓPEZ, E., ROCAMORA, M. "Fan Chirp Transformation for Music Representation". In: *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, September 2010.

[110] APOLINÁRIO, I. F., DA COSTA, M. V. M., BISCAINHO, L. W. P. "Structure Tensor Applied to Parameter Estimation in the Fan-Chirp Transform". In: *Proceedings of the 2nd AES Latin American Congress of Audio Engineering*, Montevideo, Uruguay, September 2018.

[111] BIGUN, J., GRANLUND, G. H. "Optimal orientation detection of linear symmetry". In: *Proceedings of the IEEE First International Conference on Computer Vision*, London, United Kingdom, June 1987.

[112] KNUTSSON, H. "Representing local structure using tensors". In: *Proceedings of the 6th Scandinavian Conference on Image Analysis*, Oulu, Finland, June 1989.

[113] BROWN, J. C. "Calculation of a constant Q-spectral transform", *Journal of the Acoustical Society of America*, v. 89, n. 1, pp. 425–434, January 1991.

[114] BROWN, J. C., PUCKETTE, M. S. "An efficient algorithm for the calculation of a constant-Q transform", *Journal of the Acoustical Society of America*, v. 92, n. 5, pp. 2698–2701, June 1992.

[115] SCHÖRKHUBER, C., KLAPURI, A. "Constant-Q Transform Toolbox for Music Processing". In: *Proceedings of the 7th Sound and Music Computing Conference*, Barcelona, Spain, July 2010.

[116] CANCELA, P., ROCAMORA, M., LÓPES, E. "An Efficient Multi-Resolution Spectral Transform for Music Analysis". In: *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, October 2009.

[117] DRESSLER, K. "Sinusoidal Extraction Using and Efficient Implementation of a Multi-Resolution FFT". In: *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx)*, Montreal, Canada, September 2006.

[118] DETKA, C., LOUGHLIN, P., EL-JAROUDI, A. "On Combining Evolutionary Spectral Estimates". In: *Proceedings of the IEEE 7th Signal Processing Workshop on Statistical Signal and Array Processing*, Quebec, Canada, June 1994.

[119] SHORE, J. E. "Minimum Cross-Entropy Spectral Analysis", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 29, n. 2, pp. 230–237, April 1981.

[120] CHEUNG, S., LIM, J. S. "Combined Multiresolution (Wide-Band/Narrow-Band) Spectrogram", *IEEE Transactions on Signal Processing*, v. 40, n. 4, pp. 975–977, April 1992.

[121] HURLEY, N., RICKARD, S. "Comparing measures of sparsity", *IEEE Transactions on Information Theory*, v. 55, n. 10, pp. 4723–4741, October 2009.

[122] BOSI, M., GOLDBERG, R. E. *Introduction to Digital Audio Coding and Standards*. New York, United States of America, Kluwer Academic Publishers (Springer), 2003.

[123] SOBEL, I., FELDMAN, G. "A 3x3 Isotropic Gradient Operator for Image Processing", presented at the Stanford Artificial Intelligence Project (SAIL), 1968.

[124] BITTNER, R., SALAMON, J., TIERNEY, M., et al. "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, October 2014.

[125] RAFFEL, C., MCFEE, B., HUMPHREY, E. J., et al. "mir_eval: A transparent implementation of common mir metrics". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, October 2014.

[126] EYBEN, F., BÖCK, S., SCHULLER, B., et al. "Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks". In: *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, August 2010.

[127] BÖCK, S., KORZENIOWSKI, F., SCHLÜTER, J., et al. "madmom: a new Python Audio and Music Signal Processing Library". In: *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, Netherlands, October 2016.

[128] MACQUEEN, J. B. "Some Methods for Classification and Analysis of Multivariate Observations". In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, United States of America, January 1967.

[129] DE TOMAZ JÚNIOR, P. D. *Separação Automática de Instrumentos de Percussão Brasileira a partir de Mistura Pré-Gravada.* Master Dissertation, Federal University of Amazonas, Manaus, Brazil, 2016.

[130] GONÇALVES, G., COSTA, O. *The Carioca Groove: The Rio de Janeiro's Samba Schools Drum Sections.* Rio de Janeiro, Brazil, Groove, 2000.