COPPE UFRJ
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia

RHYTHMIC ANALYSIS AND MODELING OF EXPRESSIVE MUSIC PERFORMANCES — SAMBA AS A CASE STUDY

Lucas Simões Maia

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
Junho de 2024

RHYTHMIC ANALYSIS AND MODELING OF EXPRESSIVE MUSIC PERFORMANCES — SAMBA AS A CASE STUDY

Lucas Simões Maia

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) 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, D.Sc.
Prof. João Baptista de Oliveira e Souza Filho, D.Sc.
Prof. José Gabriel Rodríguez Carneiro Gomes, Ph.D.
Prof. Hugo Tremonte de Carvalho, D.Sc.
Prof. Diego Barreto Haddad, D.Sc.
Prof. Paulo Antonio Andrade Esquef, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2024

Maia, Lucas Simões

Rhythmic Analysis and Modeling of Expressive Music Performances — Samba as a case study/Lucas Simões Maia. – Rio de Janeiro: UFRJ/COPPE, 2024.

XXV, 303 p.: il.; 29, 7cm.

Orientador: Luiz Wagner Pereira Biscainho

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2024.

Referências Bibliográficas: p. 245 – 278.

1. Music information retrieval. 2. Automatic drum transcription. 3. Rhythmic patterns. 4. Meter tracking. 5. Microtiming tracking. 6. Samba. I. Biscainho, Luiz Wagner Pereira. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Para meu avô.

Acknowledgements

Agradeço, de modo primeiro, ao Bom Deus, a quem insistentemente pedi e peço a inspiração para minhas ações e a ajuda para realizá-las, e cuja presença consoladora não deixei de perceber um só momento, mesmo quando eu esmorecia.

Agradeço a meus pais pelo apoio inestimável, pela educação que me deram, e por jamais terem medido esforços por mim ou por meus irmãos. Agradeço também a minhas irmãs pelos momentos descontraídos no corredor da casa, e por terem sido sempre carinhosas quando minha cabeça precisava espairecer.

Agradeço a minha noiva (em breve esposa), pela confiança inabalável, e por partilhar comigo este pequeno mas frutuoso oásis de fé, amor, e cumplicidade.

Agradeço a meu orientador, Luiz Wagner, por ser para mim mais que um exemplo, por presentear-me sempre com sua pedagogia infalível, e pela administração incansável de sua paciência para com minhas inúmeras trapalhadas "lucasianas". Após catorze anos trabalhando juntos (doze mais proximamente), já não sei dizer onde começa o que aprendi de você e onde termina o que você me ensinou. Obrigado pela liberdade com que me deixou conduzir este trabalho, sem deixar de me guiar para o caminho mais preciso.

Agradeço a meus amigos de faculdade e de laboratório, que, mesmo passados muitos anos de novas carreiras, novas vidas de fato, persistem ao meu lado (e eu ao lado deles).

Agradeço também a meus amigos uruguaios que por bem pouco não são brasileiros, Magdalena e Martín. Esta tese carrega um pouco de vocês também, como de mim e do Luiz.

Agradeço aos membros da banca, por todas as suas contribuições e pelo cuidado que tiveram com este trabalho longo (no tempo e no espaço).

Merci à Slim Essid, qui m'a accueilli chez Télécom, et merci aussi aux autres professeurs et collègues, pour les bonnes discussions pendant les déjeuners.

Je dois beaucoup remercier ma deuxième famille : la chorale de Sainte-Anne de la Butte-aux-Cailles. Vous êtes toujours dans mes pensées. Merci m'avoir accueilli tant que j'étais un étranger. Grâce à vous j'ai rétrouvé ma voix que je croyais avoir perdue. Il faut surtout remercier nos grands-parents — Radu et Jeanine — pour les doux câlins, pour les histoires, pour les apéritifs, et pour le minutal à la Matius !

Agradeço, por fim, ao CNPq e à CAPES, pelo suporte concedido na forma de bolsa de estudo e de doutorado sanduíche. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Mestre:
A música é uma ciência que ensina a bem modular. Você concorda?
Aluno:
Talvez, se eu puder ver com clareza em quê consiste a modulação.

(Sto. Agostinho, Sobre a Música)

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

ANÁLISE E MODELAGEM RÍTMICA DE INTERPRETAÇÕES MUSICAIS EXPRESSIVAS — SAMBA COMO ESTUDO DE CASO

Lucas Simões Maia

Junho/2024

Orientador: Luiz Wagner Pereira Biscainho

Programa: Engenharia Elétrica

A transcrição musical automática pode ser definida, de maneira geral, como o processo inverso à execução musical, visando à representação de seus elementos principais (e.g., aspectos temporais, timbres) em alguma forma de notação. Esta tese propõe o desenvolvimento de metodologias de processamento de sinais e aprendizagem de máquina que auxiliem na transcrição automática, ampliando o escopo cultural desta tarefa ao tratar o samba como objeto de estudo. Para tanto, nós curamos e anotamos dois conjuntos de dados de samba: BRID e SAMBASET.

A primeira parte do trabalho é dedicada à classificação de sons percussivos. Nosso foco é o reconhecimento das articulações produzidas ao percutir a membrana ou o corpo do instrumento. Sugerimos um descritor de modulação que, juntamente com descritores temporais tradicionais, apresentou o melhor resultado na classificação das articulações de repique e tantã. Também investigamos a classificação de toques arquetípicos nesses instrumentos, obtendo uma medida F de 89%.

A segunda parte trata da tarefa de descrição rítmica. Discutimos alguns descritores da literatura e a percepção que fornecem sobre os padrões encontrados na BRID. Constatamos a incapacidade de generalização para o samba dos sistemas do estado da arte em rastreamento de métrica musical. Por isso, propomos uma metodologia para a adaptação destes modelos, a partir de um pequeno esforço de anotação, a conjuntos de dados não vistos durante o treinamento, considerando a homogeneidade destes conjuntos. Mostramos que, anotando-se menos de 1,5 minuto, pode-se obter resultados comparáveis ao de estratégias tradicionais de treinamento. Também descrevemos uma técnica para determinar quais amostras são informativas para se anotar. Finalmente, apresentamos um sistema para a inferência de pulsos musicais e microtempo, que corresponde a desvios temporais em pequena escala.

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

RHYTHMIC ANALYSIS AND MODELING OF EXPRESSIVE MUSIC PERFORMANCES — SAMBA AS A CASE STUDY

Lucas Simões Maia

June/2024

Advisor: Luiz Wagner Pereira Biscainho

Department: Electrical Engineering

Automatic music transcription can be broadly defined as the “inverse” of music performance, aiming to represent its defining elements (e.g., timing, timbres) in some form of notation. This thesis proposes developments to signal processing and machine learning methodologies that facilitate the task of automatic transcription, widening the cultural scope of the task by investigating *samba* as a subject. For this reason, we curate and annotate two datasets of *samba* music: BRID and SAMBASET.

The first part of this work is dedicated to the classification of drum sounds, focusing on the recognition of the articulations produced by striking the drumhead or shell of two instruments: *repique* and *tantã*. We propose a modulation-based descriptor that, when combined with traditional temporal descriptors, displays the best performance in classifying articulations from each instrument. Additionally, we explore the classification of archetypal strokes on both instruments, obtaining an F-measure of 89% in this task.

The second part deals with the task of rhythmic description. We discuss a few descriptors from the literature and the insights they provide on patterns found in BRID. We verify that state-of-the-art meter tracking models are unable to generalize to *samba* music. Therefore, we propose a methodology for adapting these models, with a small annotation effort, to work on out-of-corpus datasets given their homogeneity. We show that by annotating less than 1.5 min, an end user can train a model with this data and achieve results comparable to those using traditional training schemes. We also describe a pipeline for determining which informative samples to annotate. Lastly, we present a system for the inference of beats and microtiming, which corresponds to small-scale temporal deviations.

Contents

List of Figures

List of Tables

List of Symbols

$D_x(c)$	Scale transform of $x(t)$, p. 72
$X(\mathrm{e}^{\mathrm{j}\Omega})$	Discrete-time Fourier transform of $x[n]$, p. 63
$X(\mathrm{j}\omega)$	Fourier transform of $x(t)$, p. 62
$X_{\mathrm{CQ}}[k,m]$	Constant-Q transform of $x[n]$, p. 67
$X[k,m]$	Short-time Fourier transform of $x[n]$, p. 65
$X[k]$	Discrete Fourier transform of $x[n]$, p. 64
Ω	Angular frequency of a discrete-time signal, p. 63
$\angle z$	Phase (argument) of z, p. 71
$*$	Linear convolution operator, p. 63
$\cdot$	Dot product operator, p. 149
$\circledast$	Circular convolution operator, p. 64
j	Imaginary unit, p. 62
$\Vert\cdot\Vert$	L_2-norm, p. 131
$\mathbb{N}$	Set of natural numbers, p. 65
$\mathbb{N}^*$	Set of non-zero natural numbers, p. 65
$\mathbb{R}$	Set of real numbers, p. 72
$\mathbb{R}^+$	Set of positive real numbers, p. 72
$\mathbb{Z}$	Set of integers, p. 63
$\mathbb{Z}^*$	Set of non-zero integers, p. 166
$\mathbf{X}_{N\times M}$	$N \times M$ matrix, p. 110

$\mathbf{x}$	D-dimensional random variable, $[x_1, \ldots, x_D]$, p. 164
$\mathbf{x}_{1:K}$	Random sequence, $\{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, p. 164
$\mathcal{H}\{\cdot\}$	Discrete Hilbert transform, p. 76
$\mathcal{S}_l\{\cdot\}$	l-th order scattering transform, p. 95
$\mathcal{U}_l\{\cdot\}$	l-th order wavelet modulus transform, p. 95
ω	Angular frequency of a continuous-time signal, p. 62
$\mathrm{sgn}(\cdot)$	Sign function, p. 80
$\|z\|$	Magnitude of z, p. 65
$d_{\cos}(\cdot, \cdot)$	Cosine distance function, p. 149
$s_{\cos}(\cdot, \cdot)$	Cosine similarity function, p. 149
$x(t)$	Continuous-time signal, p. 62
$x[n]$	Discrete-time signal, p. 63
z^*	Complex conjugate of z, p. 67
bpm	Beats per minute, p. 51

List of Abbreviations

t-SNE	*t*-distributed Stochastic Neighbor Embedding, p. 153
ACF	Autocorrelation Function, p. 142
ADT	Automatic Drum Transcription, p. 5
AI	Artificial Intelligence, p. 4
APF	Auxiliary Particle Filter, p. 234
BFCC	Bark-Frequency Cepstral Coefficients, p. 89
BGRU	Bi-Directional Gated Recurrent Unit, p. 164
BLSTM	Bi-Directional Long Short-Term Memory, p. 163
CD	Complex Domain, p. 121
CFCC	Cent Filter-bank Cepstral Coefficients, p. 116
CNN	Convolutional Neural Network, p. 92
COG	Spectral Center-of-Gravity Demodulation, p. 72
CQFP	Constant-*Q* Fluctuation Patterns, p. 146
CQT	Constant-*Q* Transform, p. 66
CRF	Conditional Random Field, p. 162
CRNN	Convolutional Recurrent Neural Network, p. 163
DBN	Dynamic Bayesian Network, p. 162
DCT	Discrete Cosine Transform, p. 86
DFT	Discrete Fourier Transform, p. 64
DIV	Diversity Sampling, p. 212

DPW	Dynamic Periodicity Warping, p. 145
DP	Dynamic Programming, p. 161
DST	Direct Scale Transform, p. 73
DTFT	Discrete-time Fourier Transform, p. 63
DTW	Dynamic Time Warping, p. 110
DWT	Discrete Wavelet Transform, p. 67
ERB	Equivalent Rectangular Bandwidth, p. 88
E	Short-Time Energy ODF, p. 121
FFT	Fast Fourier Transform, p. 67
FIR	Finite Impulse Response, p. 69
FP	Fluctuation Patterns, p. 146
FSA	From-Scratch Training with Data Augmentation, p. 201
FS	From-Scratch Training, p. 201
FTA	Fine-Tuning Training with Data Augmentation, p. 201
FT	Fine-Tuning Training, p. 201
FT	Fourier Transform, p. 62
GFCC	Gammatone-Frequency Cepstral Coefficients, p. 88
GMM	Gaussian Mixture Model, p. 162
HFC	High Frequency Content, p. 121
HMM	Hidden Markov Model, p. 113
HSMM	Hidden Semi-Markov Model, p. 173
IAI	Inter-Annotation Interval, p. 180
IBI	Inter-Beat Interval, p. 149
IF	Instantaneous Frequency, p. 72
IOIH	Inter-Onset Interval Histogram, p. 144

IOI	Inter-Onset Interval, p. 110
IQR	Interquartile Range, p. 123
IT	Information Technology, p. 3
KLT	Karhunen-Loève Transform, p. 85
KNN	k-Nearest Neighbors, p. 107
LLACF	Log-Lag Autocorrelation Function, p. 177
LOGOCV	Leave-One-Group-Out Cross-Validation, p. 131
LSF	Line Spectral Frequencies, p. 109
LSTM	Long Short-Term Memory, p. 163
MAE	Mean Absolute Error, p. 127
MAP	Maximum a posteriori, p. 167
MED	K-Medoids Sampling, p. 212
MFCC	Mel-Frequency Cepstral Coefficients, p. 86
MFL	Maximum Facility Location Sampling, p. 212
MGP	Mean Ground Truth Performance, p. 189
MIR	Music Information Retrieval, p. 3
MMA	Mean Mutual Agreement, p. 190
NMF	Non-Negative Matrix Factorization, p. 110
OC	Onset Coefficients, p. 146
ODF	Onset Detection Function, p. 120
OPH	Onset Patterns Histogram, p. 148
OP	Onset Patterns, p. 146
OSS	Onset Strength Signal, p. 145
OVO	One-Vs-One Multiclass Classification Strategy, p. 131
OVR	One-Vs-Rest Multiclass Classification Strategy, p. 153

PCA	Principal Component Analysis, p. 113
PD	Phase Deviation, p. 121
PFNMF	Partially-Fixed Non-Negative Matrix Factorization, p. 110
PF	Particle Filter, p. 162
PMF	Probability Mass Function, p. 80
RBF	Radial Basis Function, p. 131
RCD	Rectified Complex Domain, p. 121
RH	Rhythm Histogram, p. 146
RMS	Root Mean Square, p. 76
RND	Random Sampling, p. 213
RNN	Recurrent Neural Network, p. 163
ROC-AUC	Area Under the Receiver Operating Characteristic Curve, p. 153
SD	Spectral Difference, p. 121
SF	Spectral Flux, p. 121
SISR	Sequential Importance Sampling/Resampling, p. 234
SIS	Sequential Importance Sampling, p. 233
SPL	Sound Pressure Level, p. 111
STFT	Short-Time Fourier Transform, p. 65
STM	Scale Transform Magnitudes, p. 145
ST	Scale Transform, p. 72
SVM	Support Vector Machine, p. 107
TAE	True Amplitude Envelope, p. 76
TCN	Temporal Convolutional Network, p. 163
UMAP	Uniform Manifold Approximation and Projection, p. 153
VQT	Variable-Q Transform, p. 92

VTK	Fast Vote-*k* Sampling, p. 211
WPD	Weighted Phase Deviation, p. 121
WST	Wavelet Scattering Transform, p. 92
ZCR	Zero-Crossing Rate, p. 80

Chapter 1

Introduction

"Long before time, before hours and minutes and seconds, on the continent of Africa, the rhythm of the earth beat for the first people" — it is with these words that Daddy Wes begins telling his story to Mat and Martha in the book "To be a drum"[1] by COLEMAN [1]. Wes teaches his children about the importance and symbolisms surrounding the drum and its ties to their African roots. He explains that, even when their ancestors were enslaved by men who "could not listen to the rhythm of the earth", even when their families were torn apart, even when they had the drums taken from them: "We were the earth's people, we were the living drums, we would always be free".

Percussion instruments are among the first instrument ever built by human hands. They could be seen as a prolongation of men themselves, who had to live by their ability to strike for food or for survival [2]. Percussion continued to carry an important role in many aspects of life in the ancient civilizations. They were used for music, for communication and war, and for religious practices. There were drums everywhere, drums of every shape and size. From Sumeria (Figure 1.1), Mesopotamia, Egypt, China, India... In the Western world, we can cite the tambourine (*tympanon*) of Ancient Greece [3], which was initially used in rites for the gods, and later incorporated in theatrical music [2]. In the Middle Ages, the combination of tabor — a double-headed drum with a single snare on the batter head [3] — and pipe, played by a single musician, was widely used as accompaniment to dancing performances.

In Africa, the popularity of drums cannot be understated. After all, music and drums are, alongside dance, the elements of unity beneath all the diversity seen in the continent [4].[2] In dance music practices of Ewe people, for instance, we have: the dancers; the clapping; the *gangoki* (double clapperless bell) that establishes the

[1] An audio version of this book can be appreciated, in the powerful voice of James Earl Jones, here: `https://www.youtube.com/watch?v=7BVBBe56MUg`.

[2] We note that there are rare cases of tribes that do not use drums in dance music practices [4].

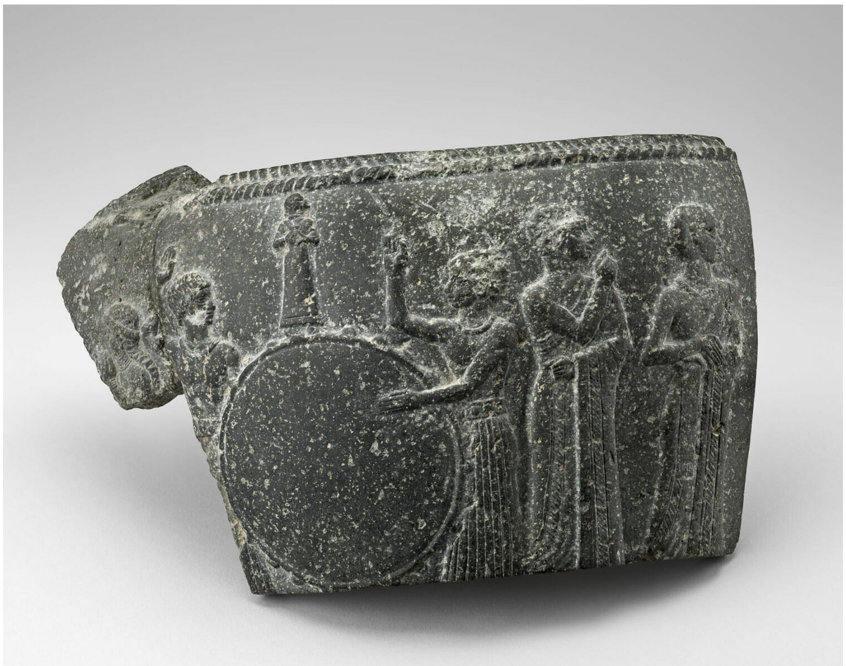


Figure 1.1: *Vase aux musiciens* (AO 5682), vessel fragment (Neo-Sumerian, 2250–2000 BC). Photograph by RMN-Grand Palais (Musée du Louvre)/Mathieu Rabeau.

rhythmic matrix; and the drums, the most important part of the orchestra [4]. In Africa, the drums do indeed "speak"! Not only the *dùndún*, the variable-pitch talking drum of the Yoruba people, which is used for long-distance communication [2]: African drummers can produce differently-sounding notes by hitting the drumhead in certain locations, with the hand or the drumstick [4].

Despite the diversity in the musical expressions across Africa, some common traits can be identified in most of them [3]: (1) full spectra, with rattles, snares, and jingles adding to this broadband characteristic; (2) cyclic form in rhythm, harmony, and melody; (3) polyrhythms, where two or more contrasting rhythms are superimposed (e.g., three over two); and (4) offbeat phrasing, when the space in-between beats is accentuated. All of these elements of African drumming can also be found, in some form or another, in the music of the diaspora. This is particularly true in the rhythms used in Afro-Brazilian religious (e.g., *candomblé*) and music practices (e.g., *coco*). Of course, this is also holds for *samba*, which is recognized as part of Brazil's intangible cultural heritage.

In this thesis, we approach the melo-rhythmic properties of *samba*'s percussion from a computational perspective and attempt to contribute to the automatic analysis of music from within this selected musical context. Since the research in computational music analysis has lacked in producing multicultural or generalizable approaches, except for more recent efforts, our aim is that the methodologies and models developed in this work apply to other underrepresented music genres or, at least, that it provides valuable insights for the promotion of a larger diversity in the field. In this chapter, we set the research context for this thesis and present its motivation. We also discuss our objectives in more detail and provide a summary of the organization of this manuscript.

1.1 Research Context

The amount of data that is produced and shared through the Internet and especially in social platforms grows more and more rapidly each year. For example, it is estimated that, in 2022 YouTube had surpassed the mark of 500 hours of video content uploaded each minute, against 60 hours per minute in 2012. It is clear that information technology (IT) has to advance in order to be capable of matching this pace — helping users not only to create, but also to organize, visualize, and interpret this increasing volume of data. Since we live in a plural world, designers of IT tools must take care in allowing for the capture of multicultural contexts.

Music has always been one of the most important expressions of human culture. It is a physical, sociocultural, emotional, and artistic phenomenon that has always shaped the way we connect to one another and has created group identities throughout the world. Thus, it is no surprise that music material is also currently being produced and consumed at an increasing rate, which was facilitated by the development of recording and storage techniques in the 20th century. We now have large music collections available in digital format on the internet and simple ready-to-install *apps* on mobile phones allow the user to easily interact with music, creating samples and even complex musical structures, among other things.

Music Information Retrieval (MIR) is a relatively new research field that is thriving in this paradigm. It encompasses many tasks within music technology and has several industrial and academic applications, such as genre recognition, tempo and beat tracking, chord recognition, music transcription, and music structure analysis, just to name a few [5]. MIR is known for its interdisciplinarity; it combines efforts from music theory and musicology, signal processing, machine learning, statistics, information theory, etc. The end users of MIR technologies go from the researchers themselves (e.g., programmers, musicologists) to sound engineers and music teachers, and, finally, to anyone who is interested in or even simply listens to music — recommender systems like the ones used by Last.fm[3] and Spotify[4] apply MIR techniques to analyze audio content and predict the users' preferences.

We have recently reached an understanding that information technology data and models in general present a strong cultural bias due to years of short-focused research on aspects and problems of the Western world [6]. In the case of MIR re-

[3] `http://www.last.fm/`

[4] `http://www.spotify.com/`

search, the usage of Western corpora has conditioned both the problems we work on and the solutions that are found [6]. Even though there are still developments being made within the research with this kind of music, novel works have arisen with the objective of modeling musical expressions from different cultural traditions under a fresh perspective, since current MIR models may not be appropriate to analyze this data [7]. These efforts have put the computational approaches in contact with ethnomusicological and antropological practices. A recent article by HUANG et al. [8] recognizes data diversification as a first but insufficient step, calling MIR to reevaluate its ontological, epistemological, methodological, and axiological assumptions. This goes in tandem with current trends putting into question technical, legal, and ethical aspects of artificial intelligence (AI) systems.

1.2 Motivation

Rhythm is one of the innate and fundamental aspects of music. It is usually understood as a kind of "macro" descriptor that incorporates all aspects of movement in music with respect to its organization in time [3, 9]. In a specific sense, rhythm can also indicate the succession of notes in a established pattern, which might be constrained by meter and tempo [3]. An automatic system that attempts a full description of rhythm has to deal with all of these aspects as they occur in specific musical realizations [10]. Moreover, it also has to consider the cognition of rhythm. We all subconsciously use temporal and structural regularities in music as cues to learn what to expect from succeeding musical events. For example, if we listen to a recording of a Classical piece with lively tempo or a modern pop song, we can readily "synchronize" to the music and start tapping to it. If many people repeat this same experiment, there is usually a consensus as to where tap positions are, which indicates that they are related to events in the music (even if they do not coincide with physical events — notes or other sounds — in the recording) [11].

In the seminal work "A Generative Theory of Tonal Music" (GTTM), LERDAHL and JACKENDOFF [12] borrowed elements from linguistics, Schenkerian analysis, and Gestalt psychology, to formally describe the musical "intuitions" that a listener has when exposed to a musical "idiom" he is versed in. They observe that: "when hearing a piece, the listener naturally organizes the sound signals into units such as motives, themes, phrases, periods, theme-groups, sections and the piece itself" [12]. Thus, the cognition of musical rhythm can be regarded as a multilevel process. At the surface level, we have the organization in time of duration patterns; the meter is deployed at deeper levels and involve our perception and anticipation of a hierarchical structure inferred from the music surface [13]. Models used in music cognition to discern the metrical structure are generally based on rule systems, which describe the possible structural descriptions (interpretations) and establish the criteria for the evaluation of such possibilities [11, 12]. These rules usually presume an evenly spaced (isochronous) structure at all levels [14]. This is, of course, not feasible in actual performances; musicians are bound to both physical/technical limitations and implicit stylistic/personal requirements, i.e., mistakes and timing fluctuations (enforced or not) occur naturally. Even in those cases, a certain amount of regularity is to be expected. However, some musical traditions are characterized by non-isochronous rhythms — this is the case of the music from Southeast Europe and from Africa and the African diaspora [14], for example. This non-isochrony is usually regarded as genre-defining temporal deviations at a very small scale [15].

As stated before, another important aspect of African and Afro-diasporic music is the fact that, when in full play, the rhythmic patterns executed on the drums have almost a melodic component to them. Musicians do in fact control the sound production by changing the placement of the attack (e.g., on the center of the membrane, near the rim, on the shell) and its velocity, by muting the sound (e.g., leaving the drumstick or hand on the skin), among others.

The area of automatic music transcription (and more specifically, the subarea of automatic drum transcription — ADT) encompasses all of the aforementioned problems. It looks for the "inverse" of the performance, i.e., from the acoustic realization, retrieve the pitches, timings, sources, or more simply all of the constituents of the produced sound [16]. Despite the name, ADT research usually deals with both membranophones and idiophones.[5] Most works are dedicated to the transcription of the components of modern drum kits — snare and kick drums, tom toms, hi-hats and other cymbals —, with a few exceptions analyzing instruments from non-Western cultures (e.g., Indian *mridangam* and *tabla* drums; Chinese drums, cymbals, and gongs). An application of ADT techniques to African drumming has to deal with more than simple instrument recognition and onset detection, it also has to be able to identify how the sound was produced, e.g., where the membrane was struck, to be culturally-aware of the nuances in sound production of this kind of music.

Another large subtask of transcription is rhythmic description. This involves discovering the underlying meter and being able to automatically synchronize with the music signal as humans are capable of. As we mentioned above, African and Afro-diasporic music have very particular rhythmic characteristics. The implicit Western bias in MIR research has led to the elaboration of datasets that mostly exhibit stable tempo, and in which drums clearly indicate the positions of the temporal grid [17]. Since state-of-the-art rhythmic description algorithms have been trained on such datasets, they have a hard time in estimating meter of music with

[5] We will also use the word "drum" to refer to instruments from both families.

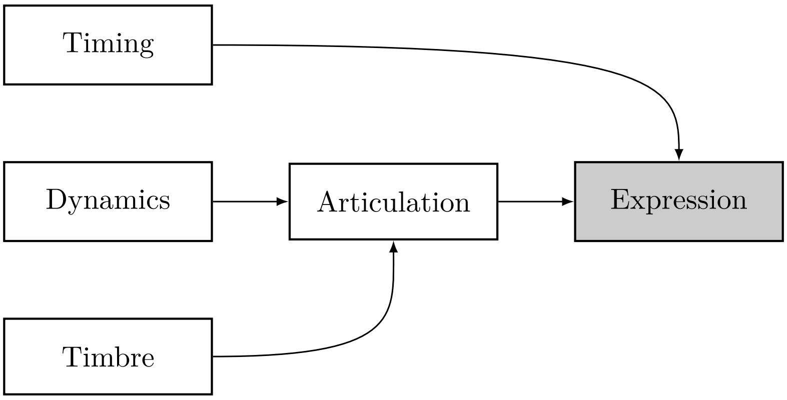


Figure 1.2: Components of rhythmic expression in percussive music that are investigated in this work. Adapted from [18].

very different characteristics [7], unless further refined on these contexts.

Figure 1.2 presents the main elements of expression in percussive music that are of interest to this work.

1.3 Scope, Objectives, and Main Contributions

The primary aim of this thesis is to improve and develop signal processing and machine learning tools that facilitate the automatic transcription of music from audio recordings, widening the cultural scope of the MIR field by investigating *samba* as a subject in this task. In particular, we emphasize the analysis of musical expressiveness within the context of drum articulation recognition and rhythm description, which includes both meter and microtiming tracking. It is in our interest that these techniques be generalizable to other musical contexts with few modifications motivated by domain-specific knowledge. Moreover, we hope that our work can lead to a better understanding of the musical phenomenon of *samba* (and other underrepresented genres in MIR) by enabling and supporting new findings from experts musicologists and allowing culturally-aware analyses even in face of a paucity of data.

Our objectives can be described as follows:

- To curate music datasets of *samba* and related genres that cover a wide range of years and instruments, enabling this work and future research.
- To annotate the curated datasets with usual MIR targets (e.g., beats, downbeats, onsets), storing relevant metadata.
- To investigate features for the automatic classification of note articulations in this cultural context.

- To assess the suitability of state-of-the-art methods for tracking meter, beat and downbeat in these datasets.
- To develop novel methods for the description of rhythmic expression at a fine timescale (microtiming).

Our main contributions can be summarized as follows:

- Two datasets representing *samba* music in various contexts — from solo performances to full commercial recordings — along with annotations for beat, downbeat, and onset positions.
- An adaptation for a state-of-the-art onset detection function and a proposed modulation-based feature for classifying drum note articulations.
- An adaptation for the training of state-of-the-art beat and downbeat tracking models in small data scenarios and a methodology for the selection of informative samples in this situation.
- A model for joint estimation of beats and microtiming.

1.4 Publications

We list here all the publications that were produced within the context of this thesis.

1.4.1 Papers accepted in peer-reviewed conferences

[19] 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: *Proc. 2018 Latin American Congr. Audio Eng. Soc. (AES-LAC)*, pp. 53–60, Montevideo, Uruguay, Sep. 2018.

[20] MAIA, L. S., FUENTES, M., BISCAINHO, L. W. P., et al. "SAMBASET: A Dataset of Historical Samba de Enredo Recordings for Computational Music Analysis". In: *Proc. 20th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 628–635, Delft, The Netherlands, Nov. 2019.

[21] FUENTES, M., MAIA, L. S., ROCAMORA, M., et al. "Tracking Beats and Microtiming in Afro-Latin American Music Using Conditional Random Fields and Deep Learning". In: *Proc. 20th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 251–258, Delft, The Netherlands, Nov. 2019.

[22] MAIA, L. S., ROCAMORA, M., BISCAINHO, L. W. P., et al. “Adapting Meter Tracking Models to Latin American Music”. In: *Proc. 23rd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 361–368, Bengaluru, India, Dec. 2022.

1.4.2 Articles published in peer-reviewed journals

[23] MAIA, L. S., ROCAMORA, M., BISCAINHO, L. W. P., et al. “Selective Annotation of Few Data for Beat Tracking of Latin American Music Using Rhythmic Features”, *Trans. Int. Soc. Music Inform. Retr.*, v. 7, n. 1, pp. 99–112, Mar. 2024.

1.5 Organization and Outline

This thesis is structured in three parts as follows.

In Part I, we lay the foundations for this work. With this intent, in Chapter 2 we briefly describe the history of *samba*, its the main instruments and rhythmic elements. Chapter 3 presents the datasets that were curated and annotated for this work. The basic transforms that are applied to audio signals in this thesis are discussed in Chapter 4.

In Part II, we approach the problem of drum sound classification. We start with Chapter 5 reviewing the main features used in the literature on this topic, and introducing our proposed features. In Chapter 6, we skim over this literature, highlighting the main works that deal with the classification of percussion instruments and their articulations. In Chapter 7, we present an investigation on the classification of articulations, going through the entire pipeline: from the detection and segmentation of note onsets to the feature extraction and classification processes. We apply this pipeline and assess how the available and proposed features deal with the nuances of sound production in two *samba* instruments: *tantã* and *repique*.

In Part III, we tackle the problem of rhythmic description from a few perspectives. First, in Chapter 8, we discuss this topic in the light of two features for compactly representing the temporal organization of music, allowing for the direct comparison of rhythmic patterns. Chapter 9 reviews the literature on two subtasks: meter and microtiming tracking. In the latter case, we give special attention to works that perform analyses on *samba* and other genres from the African diaspora. Chapter 10 contains an investigation of the beat and downbeat tracking performances of state-of-the-art algorithms in the datasets compiled for this work. In particular, we show the limitations in the direct application of these tracking systems to music for which they were not developed (or on which they were not trained). This means that different assumptions may be needed for these systems to work effectively with such music. In Chapter 11, we propose an approach for adapting deep-learning-based beat and downbeat trackers with few data. This is presented as an attempt to alleviate the burden of annotation when state-of-the-art models fail to accurately track "out-of-corpus" music. With the suggested methodology, under certain homogeneity constraints, the end user would be able to annotate few minutes of data, train the neural network, and obtain tracking results on the remainder of the dataset that are on par with those of more traditional training schemes (e.g., separating the dataset in folds). We also present a scheme for selecting an informative portion of the dataset for annotation. Chapter 12 first assesses the microtiming profile of *samba-enredo* as expressed by the *tamborim*, and then presents our model for automatically tracking these systematic deviations. We perform inference in the model using exact and approximate methods, and show that approximate can be used to obtain estimates for the evolution of small-scale deviations during a performance.

Chapter 13 discusses our main contributions and presents the perspectives of future work.

Additionally, we have included two appendices. Appendix A contains a detailed description of all files in our purely instrumental *samba* dataset. Appendix B clarifies part of the terminology used for describing instruments by succinctly presenting the history of organology and of the taxonomy of musical instruments.

Part I

Preliminary Aspects

Chapter 2

Samba Carioca

Samba plays a huge role in Brazil's imaginary overseas, right alongside soccer (and soccer players like Pelé and Ronaldinho) and the country's rich fauna and flora, commonly associated with the Amazon rainforest. Every year, Brazil receives hundreds of thousands of international tourists for the activities of *carnaval* in the cities of Salvador, São Paulo and Recife, for example. The domestic tourism also increases during this period, as many Brazilians go see the "Greatest Show on Earth" — how the *desfiles* (parades) held in Rio de Janeiro's Sambadrome are usually called by media outlets. Happiness, dance, music, the feast where "anything goes" — when reflecting upon these modern images of *carnaval*, it might be perplexing to learn that *samba* was disesteemed and faced repression during the late nineteenth and early twentieth centuries. In order to understand why today's *samba* can be regarded as Brazil's quintessential rhythm and how the festivities of *carnaval* acquired their considerable importance in the country's urban centers, we first need to peer at the circumstances that led to their formation as sociocultural activities.

Similarly to what occurred in most of the former European colonies in the Americas, the development of Brazil during its Colonial, Royal, and later Imperial rules relied heavily on extractive activities (e.g., *pau-brasil* in the sixteenth century, gold and diamonds in the mid-eighteenth century), on the plantation economy (at first the sugarcane, which was then succeeded by coffee, but also other agricultural products like tobacco and cotton) and on the slavery system. In the first years that followed the "discovery" of Brazil (1500), the Portuguese colonizers used indigenous people as labor, mainly for the extraction of *pau-brasil* (brazilwood). These workers were then rewarded with unimportant manufactured objects such as mirrors, combs, knives, etc., in a barter economy. Shortly after, however, a system of capture and enslavement of indigenous people was established, although, due to many factors — including legal impediments (that only permitted the imprisonment in situations of "just war"), the epidemics brought by Europeans (for which the natives' immunological systems were not prepared), and captivity escapes (Native Americans had a

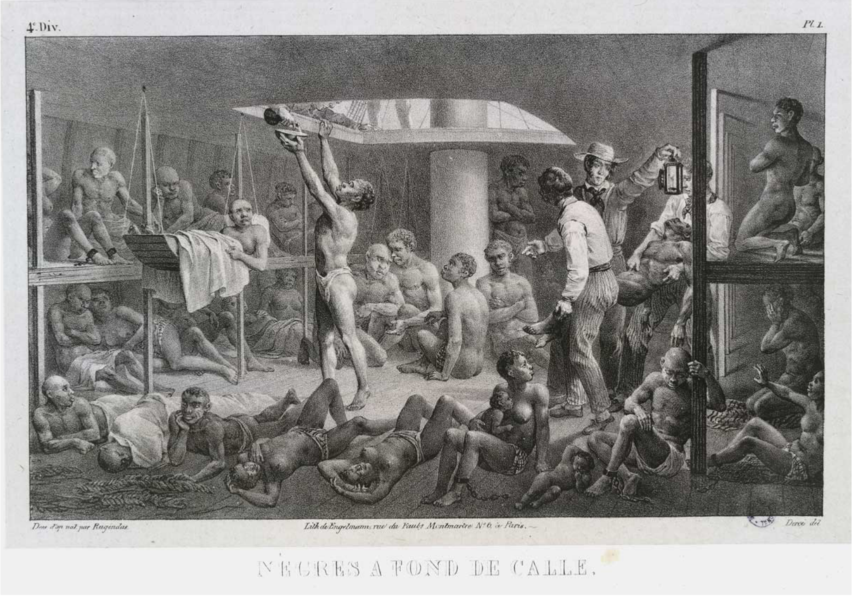

4e Div.
Pl. 1
Dess. d'ap. nat. par Rugendas
Lith de Engelmann, rue du Faubg Montmartre N° 6, à Paris.
Deroy del
NÈGRES A FOND DE CALLE.

Figure 2.1: *Nègres a fond de calle*, lithographic print (Rugendas, 1827 [25]). Engraved by Laurent Deroy after a drawing by Johann Moritz Rugendas. It displays a group of Africans in the cargo hold of a slave ship headed to Brazil. The conditions were very poor and many slaves died before reaching the Americas. The poet Antônio Frederico de Castro Alves (1847–1871) described the sufferings of these enslaved men and women in the famous poem "*O Navio Negreiro*" ("The Slave Ship"): "Yesterday, the Sierra Leone / The war, the lion hunting, / The sleep sleeping peacefully... / Under the tents of amplitude... / Today... the dark basement, deep / Infected, crowded, filthy, / Housing the plague instead of a jaguar... / And sleep always interrupted / By the sudden pull of a deceased / And the crashing of a body into the sea..." [26].

good knowledge of their land) —, it was somewhat short-lived. In the mid-sixteenth century, indigenous labor started to decline, and was substituted by the African slave trade (cf. Figure 2.1). This exploitation system, which was used throughout American colonies (not without nuances in the British, Portuguese, and Spanish controlled territories), allowed the consolidation of Brazil's economy in the Atlantic, since its production was mostly intended for the northern hemisphere. For three centuries, enslaved Africans (and their descendants) were the main source of manpower for the agricultural production. Practices associated with Africa in the country were thus assessed as being "manifestations of slaves" [24] and were typically depreciated.

It was at the then capital, Rio de Janeiro, that following many decades of alienation the religious and festive practices of Afro-Brazilian populations found fertile grounds on which to grow. The *samba* that was developed after the Abolition of Slavery (1888) allowed black and mulatto people — "lower class citizens" —, who then fought repression in expanding urban centers [27, 28], to exert a new cultural power. As many anthropologists put, it was mainly through *samba* and the communities built around it that a set of "symbolic negotiations" [24] aiming for equality began to take place. When Carmen Miranda — the Portuguese-born "Brazilian bombshell" — hit Hollywood in the 1940s, back in Brazil *carnaval* was already a "national ritual": a ritual of inversion, in which rules and social hierarchies are temporarily suspended [29]. In this ritual, the *samba urbano* (urban *samba*), also known as *samba carioca*[1] refers to both the dance and the music of *carnaval*, but it is not accessory neither delegated to the background like a soundtrack; instead it is the very means by which such inversions seem to occur [29].

This chapter is dedicated to presenting a brief history of *samba* as a dance and music genre, its subgenres and its relation to *carnaval*. In Section 2.1, we examine the meaning of the word "*samba*". Sections 2.2 and 2.3 trace *samba* and *carnaval* back to their origins and provide insights on how they became a part of Brazil's identity. Finally, Sections 2.4 and 2.5 discuss the idea of *ritmo* (rhythm) in a *bateria* (drum ensemble), and describe the instrumentation and a few rhythmic patterns that are used in *samba* music. We also refer the interested readers to corresponding musicological and anthropological literature.

2.1 Meaning of "*Samba*"

Despite its importance as cultural practice, there is little consensus on what the word *samba* really means. Scholars agree that, due to its origins among circle dances of people of African (Bantu) descent, "*samba*" would have been derived from a word of one of the Bantu languages. It could, therefore, have come from the Kimbundu verb "*semba*", meaning "to court"/"to please" [28], or even the homographic Kimbundu noun that defines a choreographic navel gesture of certain dance practices (in Portuguese, *embigada* or *umbigada*) [30, 31]. It could also have come from the Kikongo word "*sàmba*", designating another dance practice where dancers touch their chests together with force [28]. Another hypothesis is that it comes from the Chokwe, again as verb, "*samba*", and meaning "to caper"/"to play like a goatling" [28].

Whichever its origin may be, *samba* is most definitely a very broad term. For a long time in Brazil, it was synonymous with *batuque* — from the verb "*bater*", meaning "to hit, to beat" (a drum) — as the collective of folkloric practices of Afro-Brazilian people particularly in enslaved communities (see Figure 2.2). These festive activities were usually expressed in the form of circle dancing, accompanied by singing and instrument playing, and characterized by the *umbigada*.[2] They

[1] *Carioca* is the demonym used to designate anything related to the city of Rio de Janeiro.

[2] In fact, scholars point out three different historic *batuque* areas in Brazil, with some superposition [32]: the *coco* zone (Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas); the *samba* zone (Maranhão, Bahia, Rio de Janeiro, São Paulo, Minas Gerais); and the *jongo* zone (Rio de Janeiro, São Paulo, Minas Gerais, Goiás). This is not to say, however, that *batuques* in a given area are homogeneous in terms of form. Instead, these varieties exhibit a combination of different characteristics. They can have the typical shape of a dance of *umbigada* (as seen in the

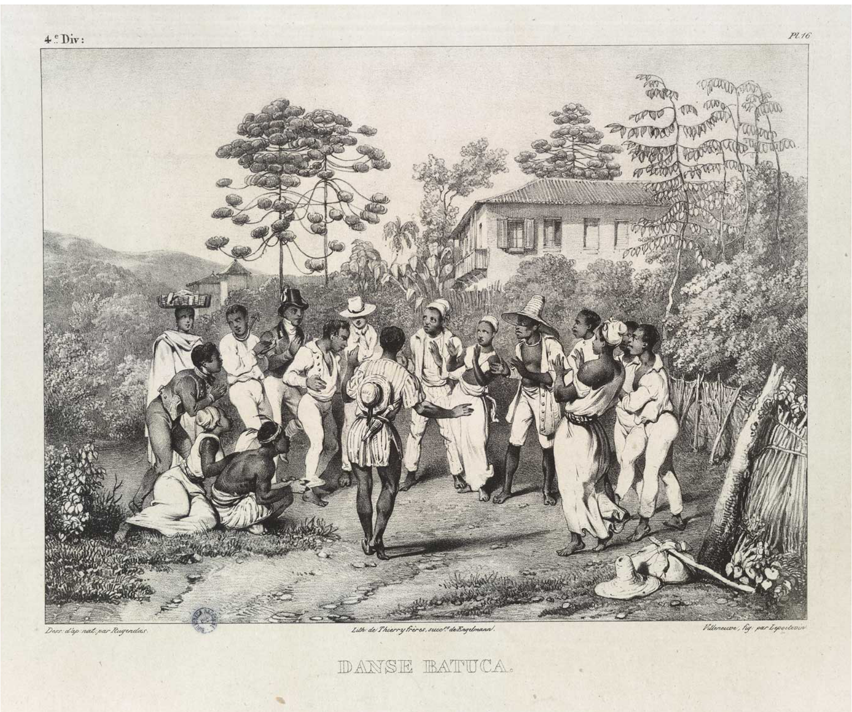

4e Div:
Pl.16
Dess: d'ap: nat: par Rugendas.
Lith: de Thierry frères, succrs de Engelmann.
Villeneuve, fig: par Lepoitevin
DANSE BATUCA.

Figure 2.2: *Danse Batuca*, lithographic print (Rugendas, 1827 [25]). Engraved by Louis Villeneuve after a drawing by Johann Moritz Rugendas. It depicts typical dance practices for people of African descent that Rugendas witnessed during his 1822–1825 stay in Brazil.

spread throughout the country and acquired different flavors depending not only on the region they were developed but also on the origin of the enslaved populations. Examples of such practices include, but are not limited to [30, 33, 34]: *tambor-de-crioula*, in the state of Maranhão; *coco*, in Ceará and Paraíba; *coco-de-zambê* or *bambelô*, in Rio Grande do Norte; *coco-de-parelha*, *coco-travado*, *coco-de-roda*, in Pernambuco; *samba-de-parelha* in Sergipe; *samba-de-roda*, *bate-baú*, *samba-de-chave*, and *samba-corrido*, in Bahia; *jongo* in Espírito Santo, Rio de Janeiro and Minas Gerais; *caxambú*, *partido-alto*, and *samba-duro* (*batucada* or *pernada*) in Rio de Janeiro; *samba-lenço* or *samba-de-bumbo* in São Paulo.

Later, "*samba*" would come to be used in hybrid music genres primarily aimed at the *carnaval* season: *samba-tango carnavalesco*, *samba-jongo*, *samba-choro*, *samba-rumba*, *samba-canção*, *samba-macumbeiro* [24]. Nowadays, many different rhythm variants can be found, such as: *samba-de-breque*, *samba-exaltação*, *samba-de-terreiro*,

Luanda region), which we describe in more detail in Section 2.2, but suffice it to say here that it is a somewhat free dance started by a male or female in the center of a circle and ended when this dancer chooses a substitute through the sign of the *umbigada*. They can also be pair dances, when it is a couple that starts the dance together. Finally, they can be circle dances in which the choreography is fixed (varied only in rhythm) or even line dances. In all these cases, the *umbigada* sign is executed (in full or in a simulated manner) or not observed at all [32]. From the examples given in the text, we highlight: *samba-lenço* (*samba* zone, line/pair dance with *umbigada*); *partido-alto* (*samba* zone, dance of umbigada/circle dance with *umbigada*); and *jongo* (pair/circle dance where the *umbigada* is replaced by a bowing movement).

samba-enredo, *sambalanço*, *samba-de-quadra*, *sambalada*, *samba-chulado*, *samba-raiado*, *samba-coco*, *samba-choro*, *samba-canção*, *samba-batido*, *samba-de-partido-alto*, *samba de gafieira* [33]. The *bossa-nova* (famously known due to Tom Jobim's "*Garota de Ipanema*" — "The Girl from Ipanema") and the *pagode* are also counted among *samba*'s subgenres. The term "*pagode*" itself was used, from the early twentieth century, to designate a kind of festive gathering, which took place in the houses of the *tias baianas* [24].

As one can see, the definition of the "music genre" of *samba* is not straightforward, being at least as convoluted as its connection to generic dance practices. To illustrate this, we can look at the history of the genre's phonographic recordings. The first *samba* to be recorded is usually considered to be "*Pelo Telefone*", which dates from 1917. Before it, however, many *sambas* had already been recorded with the same genre indication (the first one possibly in 1908 [31]). As CABRAL [34] points out, also before "*Pelo Telefone*", there were recordings that did not carry the "*samba*" name, despite being *sambas* while others showed "*samba*" on their covers even though they were clearly examples of other genres. Still, many critics declare that only the music produced starting in the late 1920s can receive the title "*samba*" [31], and previous recordings (and probably also the music matrices they meant to register) were more akin to other rhythms (e.g., *maxixe*). In contrast, Pixinguinha (1897–1973), instrument virtuoso and one of the greatest Brazilian popular music composers, once went on record saying that the "true" *samba* and the "true" *sambistas* were much older than "*Pelo Telefone*" [34].

We end this section with a short anecdote showing how convoluted this nomenclature problem can be even for insiders. In 1995, *samba* singer/songwriter Zeca Pagodinho was interviewed by comedian Jô Soares in his talk show. We transcribe in the following a symbolic excerpt from this encounter:

> Jô – Zeca, which *samba* do you like the most: *partido-alto*, *samba-de-roda*, or *pagode*? And what are the differences — from *pagode* to *partido-alto*, for example?
>
> Zeca – (clears throat) Well, *pagode* is this thing... *Samba*... *Samba* is that *batucada*, but *pagode* is the same thing. For example, we could make a *samba* here, right? But we wouldn't make a *samba-enredo. Samba-enredo* is that which a layman think about when someone says "*samba*", that *batucada* and the *Avenida* [of the Sambadrome]. But it isn't that. There are several types of *samba.* There's *pagode*, there's *partido-alto*, which is improvised...
>
> Jô – *Samba-enredo...*
>
> Zeca – But *samba-enredo... Samba-enredo* is February...
>
> Jô – (sarcastically) February...
>
> Zeca – And *pagode* is the whole year.
>
> Jô – So *pagode* is not very different from *partido-alto*, is it?

> Zeca – (hesitates) ...
>
> Jô – Ok, I see that it is...
>
> Zeca – [To make *pagode*] you must know... Know how to *versar* [to make verses], to improvise.
>
> Jô – And *partido-alto*?
>
> Zeca – *Também* [same thing]!
>
> Jô – So, Zeca, which one do you prefer?
>
> Zeca – You know, Jô, it's hard to answer this question. I've been asked this for the last eight years and I still can't answer. So I say "yes, it is...", "no, it isn't...", "I don't know...". You ask me "What's the difference?". There's no difference, it's the same thing, and yet it's different.

All the confusion aside, throughout this thesis the term "*samba*" will be used in specific reference to the *samba urbano carioca*, i.e., the *samba* that was developed in the city of Rio de Janeiro at the turn of the twentieth century, but also to *partido-alto* and *samba-enredo*, unless otherwise noted. These genres' ties are clearly and abundantly described in the musicological literature [28]. This delimitation of the term brings, of course, many "imperfections" since we will not be considering many historical and cultural aspects that have had their fair share in the sound formation of today's *samba*; it is done, therefore, given the broadness of the word, in a conscious effort to limit the study subject to a coherent "minimum". This is also why we found it important to give, in the following, a brief description of the transformations and mixes that led to the emergence of *samba* in the twentieth century, and the metamorphosis it went through shortly after.

2.2 History and Evolution

As we mentioned before, the roots of *samba* dance and music practices can be traced back all the way to the *umbigada* family of dances. More specifically, scholars usually recognize the *lundu*, a dance of *umbigada* that was brought to Brazil by enslaved Bantus from the Congo–Angola region in the late eighteenth century [31] (cf. Figure 2.4), as the "grandfather of *samba*" [35]. In this kind of *batuque* — the main generic (and somewhat derogatory) term used to refer to black community dance practices throughout Colonial and Imperial periods in Brazil —, all participants (musicians inclusive) are gathered in a *roda* (circle). They clap, stamp their feet, play the guitar, and sing choruses in response to a soloist's improvised verses, while a dance takes place, one pair at a time [31]. The *lundu* was a common style both in Brazil and in Africa, despite its consideration as an insinuating and lascivious dance form [27]. In any given dance of *umbigada*, one participant stands out in the center of the circle and starts to dance, usually performing moves with agility and rhythm.

If desired, he can choose a partner of the opposite sex; the soloist couple then shares a choreography,[3] and the first dancer returns to the circle [32]. The term *umbigada* denotes the bumping of navels (in Portuguese, "*umbigo*") by the soloist with the person chosen to replace them [32].

In spite of its African origins, the *lundu* dance would not be restricted to the black communities in rural areas. In fact, it was adopted in urban centres and modified by white and *pardo* groups. The *lundu* is thought to have been eventually exported to the metropolis by Domingos Caldas Barbosa (ca. 1740–1800), palace minstrel and catholic priest, where it gained popularity, specially in the Portuguese court. The dance was further "Iberianized and Atlanticized" [36], incorporating elements from the *fandango*, such as the snapping of fingers, the raising of arms or the resting of hands at the hips [35]. By the time of Brazil's Independence, the *lundu* had been transformed into what SANDRONI [31] calls "*lundu-canção*" (literally, *lundu*-song), a salon music genre closely related to the *modinha* and usually containing humorous allusions to the sexual interplay between lord and slave [31]. In Figure 2.3, two lithographic prints based on drawings by Johann Moritz Rugendas display the "*landu*" (sic) in different circles, the top one showing distinct Iberian elements such as raised arms and castanets.

Around the 1830s, Brazil's standpoint on slavery started to shift, as the country had its hand forced by one of its main commercial partners, the United Kingdom. The UK outlawed international slave trade in the 1807 Slave Trade Act. The practice would not be entirely stopped, however and, in 1833, the country expanded on its previous act by means of an Abolition Act, making both slave purchase and ownership illegal. The British, once major participants in the slave trade, were now very eager to replicate the abolition laws worldwide. Brazil achieved its sovereignty in 1822, with the Independence from Portugal, but was indebted to and relied greatly on the economic trade agreements made with the British. Their external pressure to end slavery was intensified in the late 1860s, during the Paraguayan war, when Brazil's debt rose, yet the UK's demands where met with great inertia from the part of the Brazilian government. The two countries had signed a treaty in November 1826 (with effects starting on March 1827), which gave Brazil three years to declare extinct the slave trade [37]. A law was passed in 1831 to bring the treaty's requirement into action, but it wasn't rigorously applied and local judges generally absolved the illegal traffickers. This is possibly the origin of the idiomatic expression "*para inglês ver*" (lit. "for the Englishmen to see"), meaning "for the sake of appearances, without validity". A few bills were enacted the following years and actually put into effect — notably the Eusébio de Queirós Law in 1850 (prohibiting

[3] Differently from partner dances such as the waltz and the polka, in *lundu* the couple performs choreographic elements while dancing separately.

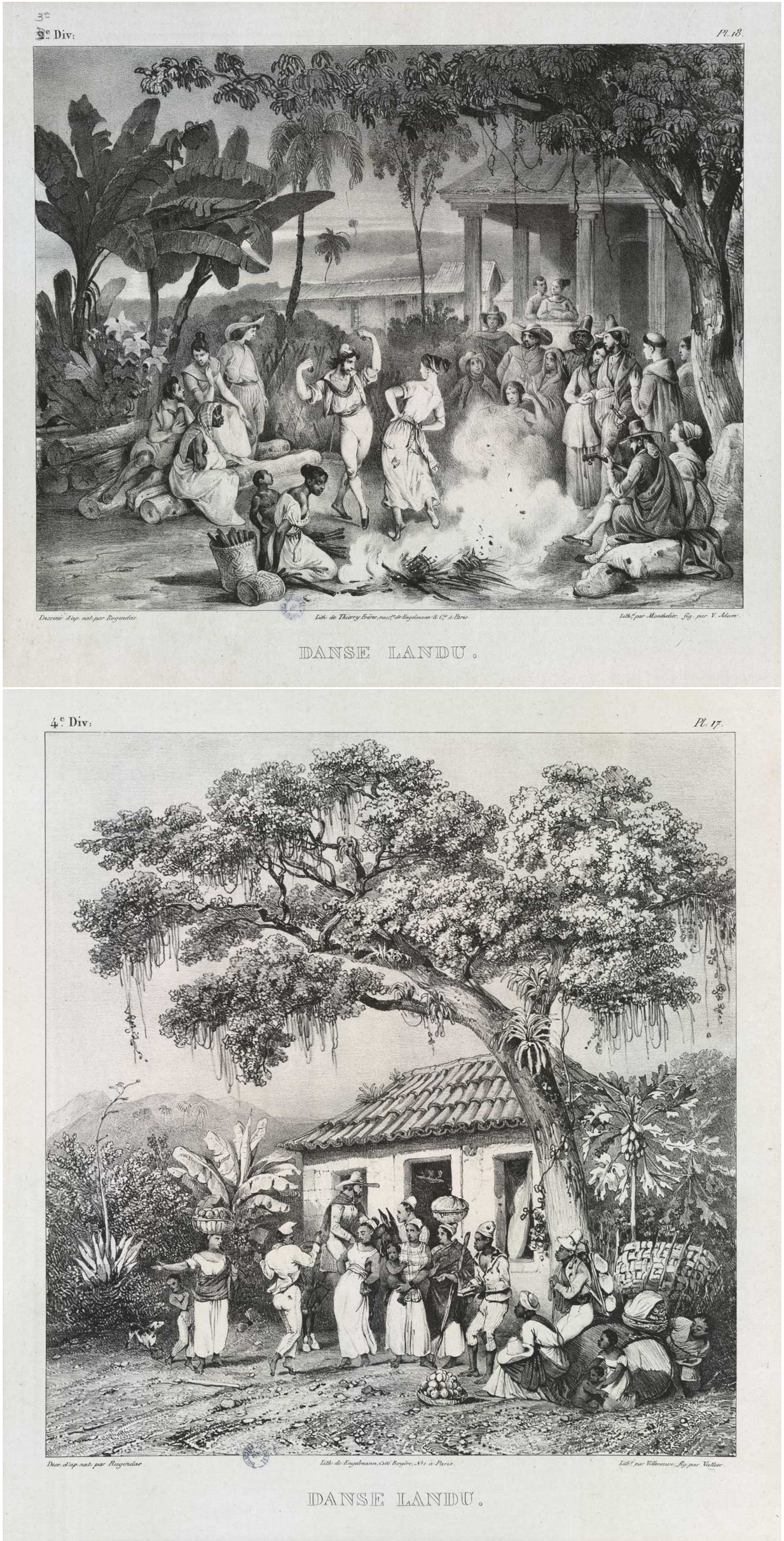

3e Div:
Pl. 18.
Dessiné d'ap. nat. par Rugendas.
Lith. de Thierry Frères, succrs de Engelmann & Cie à Paris.
Lithé par Monthelier, fig. par V. Adam.
DANSE LANDU.
4e Div:
Pl. 17.
Dess. d'ap. nat. par Rugendas.
Lith. de Engelmann, Cité Bergère, No 1 à Paris.
Lithé par Villeneuve, fig. par Vallier.
DANSE LANDU.

Figure 2.3: *Danse Landu*, lithographic prints (Rugendas, 1827 [25]). Engraved by Jules Monthelier (above) and Louis Villeneuve (below) after drawings by Johann Moritz Rugendas. In the early nineteenth century, both *batuque* (see Figure 2.2) and *lundu* where recognized by foreigners as Brazilian national dances [36].

international slave trade), the Rio Branco Law in 1871 (known as the "Free Womb Law"), and the Saraiva–Cotegipe Law in 1885 (the "Sexagenarian Law", granting freedom to slaves who reached the age of 60). Another law, dating of 1866, also granted freedom to slaves who fought in the Paraguayan war [37]. Despite all this legislative work, the country would only completely abolish slavery in 1888, when Isabel, Princess Imperial of Brazil, signed the *Lei Áurea* acting as regent to Emperor Pedro II, who was in Europe at the time. Figure 2.4 depicts the four main slave trade routes used to bring people to Brazil from the African continent from the 1550s until the Eusébio de Queirós Law.

During the same period, in the beginning of the nineteenth century, the weakening of Brazil's sugar economy in face of external competitors shifted the focus of the country's commercial endeavors to the production of coffee [27]. The grain, which was not native to the Americas, had arrived in the country at the start of the eighteenth century, but it would only have a big economic "boom" later, to meet an increasing demand on the international market. As a result, the labor force was abruptly transferred from the sugar farms in the north to coffee plantations down south in the provinces of Rio de Janeiro, São Paulo and Minas Gerais. This movement initially consisted primarily of enslaved people (when internal trafficking was still not abolished), whereas later governors would try to import workforce from an overpopulated Europe [27].

Many emancipated slaves were attracted by the life in the capital (at the time, Rio de Janeiro). In the pre-Abolition era, they would form guilds that helped buy freedom for other slaves (e.g., their own relatives) [28], and established themselves in city center *bairros* (neighborhoods) like Gamboa and Saúde, then at the outskirts of the city. The flux of migrants towards Rio de Janeiro was intensified after the Abolition as people were hopeful to find opportunities there. A diaspora from Bahia settled in the capital's port region and later in the Cidade Nova neighborhood — areas that would come to be known as *Pequena África* (lit. "Little Africa") [27]. The Abolition also marked the end of the Brazilian monarchical regime. With each constitutional law towards the end of the slavery system, large-scale farmers grew dissatisfied at the monarchy, from whom eventually they withdrew all support. This was one of the tipping points that allowed a military coup to declare Brazil a republic in November, 1889. Despite its constituent ideals, the post-Abolition newly-born republic did not see an improvement in the quality of life of black and mulatto people. Instead, the choice of immigrant labor force in the plantations and the lack of opportunities in other areas of the economy left these communities in a delicate situation [37]. They were also met with prejudice and seen with distrust, which fed and was fed by this aforementioned social inequality; their descendants still face these challenges today in Brazil and, as some authors argue, in yet greater

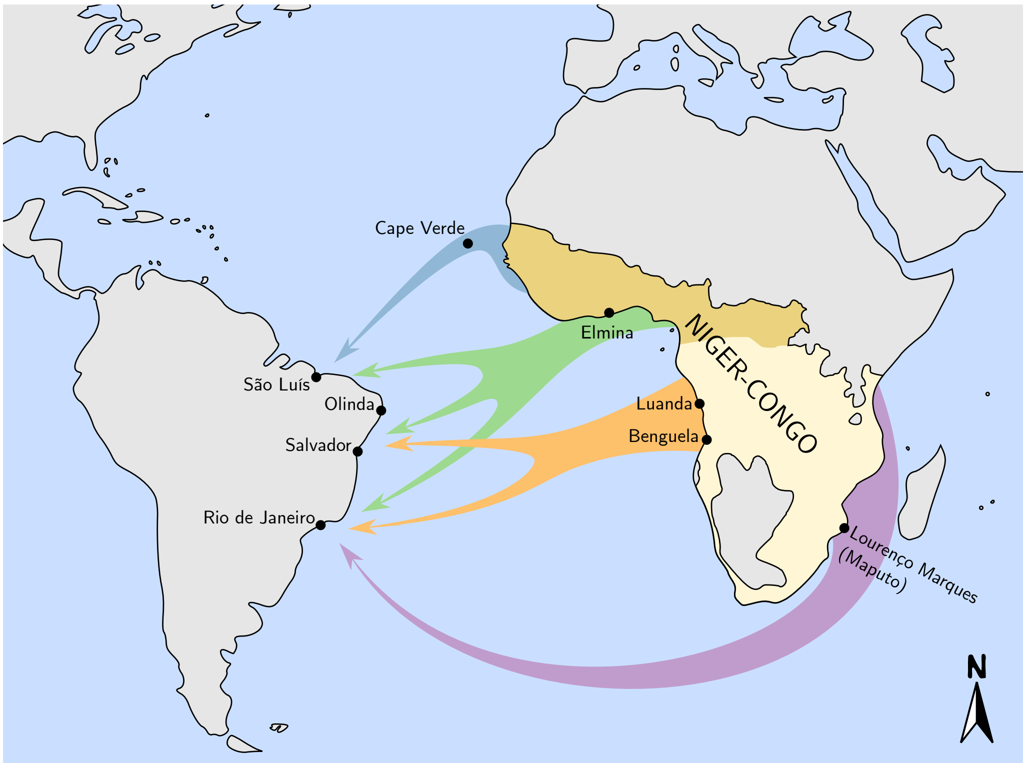


Figure 2.4: The four main slave trade routes from Africa to Brazil and ports (sixteenth through nineteenth centuries). It is estimated that over four million enslaved Africans, the majority young males, were brought to Brazil between 1550 and 1855 [37]. Almost a quarter of this trade took place in the span of only twenty years, from 1811 to 1831, with the destination port being the Valongo Wharf in Rio de Janeiro. The map also shows their main ethnolinguistic group of origin — Niger-Congo (from which Bantu, in lighter color, is a subgroup) —, but enslaved Africans actually came from many different kingdoms and tribes, such as: Fulani, Fanti, Ashanti, Bakongo, Ambundu, Yoruba and also Hausa (Afro-Asiatic ethnic group, not represented). The region of provenance for enslaved Africans in Brazil was dependent on the time period — the trade organization and local conditions in the continent — and, to a lesser extent, on slave owners' preferences [37]. In the sixteenth century, first the Guinean (blue, going through Cape Verde) and then the Mina routes (green, departing mostly from Elmina) were the main sources of slaves [37]. Starting at the seventeenth century, slaves were exported to Brazil primarily from the more southern Congo-Angola region, departing from the ports of Luanda and Benguela (orange), for example [37]. The Mozambique route (purple) was mostly unexplored until the nineteenth century when, due to the United Kingdom's influence in Abolition fights, Portugal abolished all slave trade north of the Equator and Brazil signed the 1826 treaty on the slave trade ban. An illegal trade would continue until a bit after the Eusébio de Queirós Law (1850), when the central government ratified further sanctions on local authorities that ignored this legislation. Internal traffic, however, was not banned until 1888: from 1850 until the Abolition, it is estimated that between 100 000 and 200 000 slaves where displaced in interprovincial trades motivated by the coffee "boom" [37]. Information on slave trade routes from Africa to other parts of South America, especially the Platine region, can be seen in [38].

proportions [24]. Figure 2.5 displays a map of the capital in 1885, few years prior to the Abolition and the Proclamation of the Republic. This map indicates the main historic entry points for Africans in the port area of Rio de Janeiro, the *Alfândega* and the Valongo region, and also the *Pequena África*, where the black community would settle and grow.

With this influx of people, the city of Rio suffered an unorganized growth. Prices were high due to real estate speculation practices; newcomers — black people, multiracial and poor whites —, had to house themselves in the highly compartmentalized *cortiços* (tenements), where epidemics were not infrequent. In the first years of the twentieth century, the city's mayor, Francisco Pereira Passos, initiated a process of "embellishment and sanitation" in European/*Belle Époque* molds that would, among other things, end up demolishing most of the *cortiços*.[4] The "lower class" was faced with no other option than to move — many searched for habitation on Cidade Nova and on the city outskirts or went up the *morros* (hills), forming the first *favelas* (slums) [27]. Figure 2.5 also shows the *Morro da Providência* (lit. "Providence Hill"), which would become Rio's first *favela*. A group of emancipated *baianos*,[5] their descendants and aggregates formed an "elite" amongst those expelled from the city center, becoming a reference in this heterogenous community [27]. Most of them came from Yoruba nations and remained unified under traditional religious practices. This group is also credited as being the one who brought the *samba-de-roda baiano* to the capital and who modified it in this urban environment.

As the city grew and specially after the reforms in Pereira Passos tenure, Cidade Nova became Rio's most populous neighborhood, while also being deemed the city's place for infamous amusements [31]. In the *gafieiras*[6] of this neighborhood, the black communities' musical traditions would again be mixed with "Western" popular rhythms. This effervescent environment, secretly in the beginning, popularized a new dance style — the *maxixe* [31]. The style's precise origin is uncertain, but it is believed to have been greatly influenced by polka and waltz, dance forms that arrived

[4] This embellishment project became known as "*bota-abaixo*" (lit. "knockdown"), referring to the forced evictions and subsequent demolitions.

[5] *Baiano* is the gentilic used to designate people from the state of Bahia.

[6] The *gafieiras* were dance clubs that started appearing in Rio at the end of the nineteenth century. They were frequented mostly by the lower classes, migrants from rural parts of the country (in its majority, formerly enslaved people), who procured new forms of amusement in the capital [39]. The *gafieiras* provided them a middle ground between folkloric festivities and high-society balls of the urban environment [39]. The first clubs would open in the city center, especially in Cidade Nova, examples being *Aristocratas da Cidade Nova* and *Kananga do Japão* (where Sinhô would work for over a decade) [39]. The desire for social mobility was such that the balls had entry fees and owners would generally impose very strict rules and dress codes — these would not prevent beginning dancers to make real gaffes (which, legend has it, is the origin of the name "*gafieira*", "the place where many gaffes are committed"). At first, the music played in *gafieiras* consisted of mainly piano pieces (waltzes, polkas and *maxixes*); later, *sambas*, fox-trots and even jazz music would become crowd favorites [39].

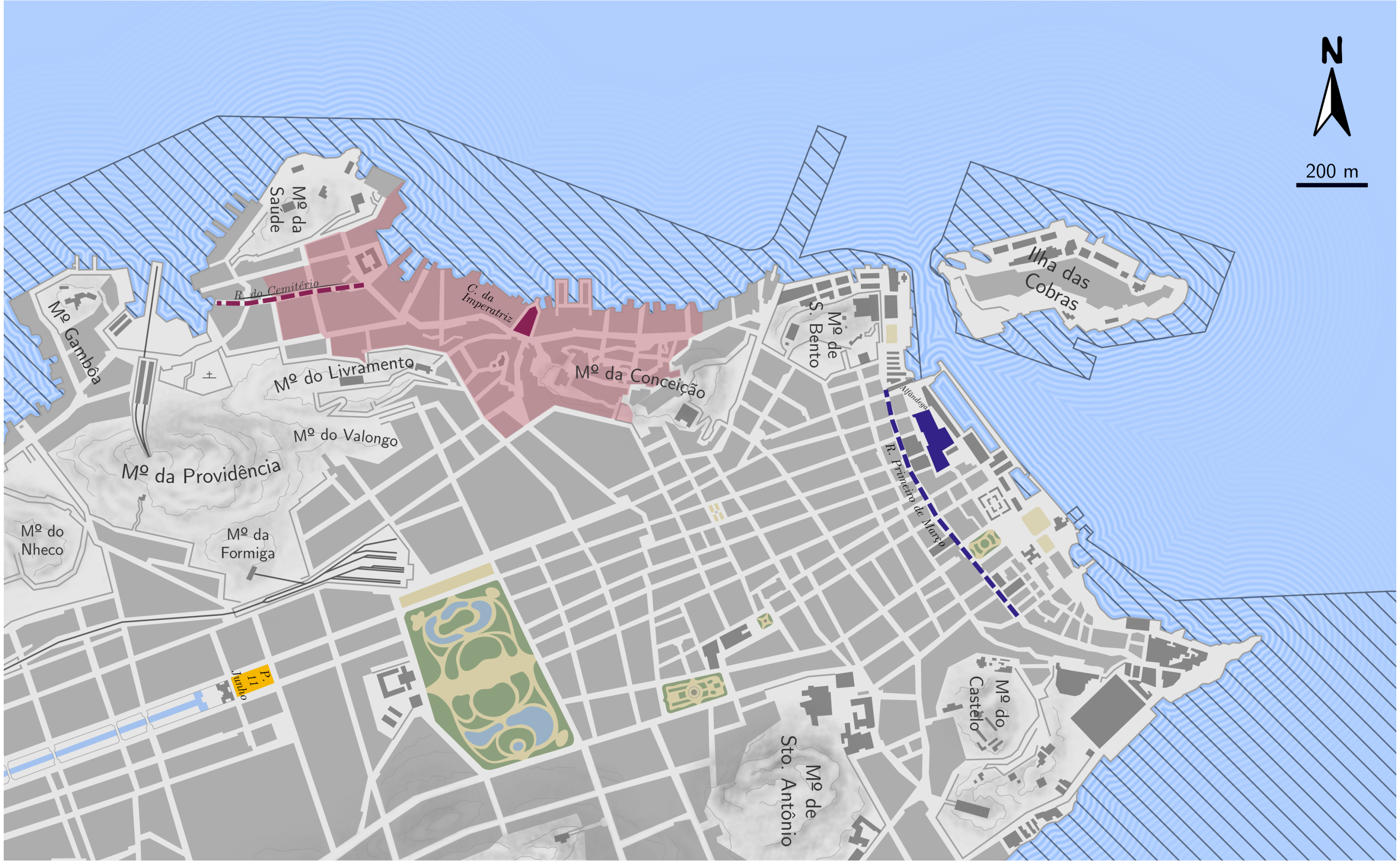


Figure 2.5: Map of Rio de Janeiro's city center in 1885 (adapted from [40]). (Continued on the following page.)

Figure 2.5: (Continued from previous page.) We indicate (in blue, top) the *Alfândega* (custom house) and the *Rua Primeiro de Março*, formerly known as *Rua Direita*, by then the capital's main road (dashed blue line). In purple (left), we show the *Cais da Imperatriz* (lit. "Empress Wharf", former *Cais do Valongo* and later renamed as *Praça Municipal*) and the *Rua da Harmonia* (lit., "Harmony Street", former *Rua do Cemitério*, "Cemetery Street", where enslaved men and women were buried). We also highlight (in a pinkish hue, left) the region of Gambôa and Saúde in which the poor black migrants first settled; it eventually expanded and reached the *Praça 11 de Junho* (yellow, bottom), in Cidade Nova, where *maxixe* and *samba* would be born. The hatched area corresponds to the current city limits. Before 1779, all enslaved men captured in Africa entered Rio through the *Alfândega* and were sold to slave owners mostly from the markets in the *Rua Direita* [41]. A law was introduced in 1758 prohibiting the commercialization of slaves inside the city in view of the many epidemics it was facing [41]. This law was mostly ignored, possibly due to traffickers' pressure, up until the rule of D. Luís de Almeida, the Marquess of Lavradio, as Viceroy in the Portuguese colony (1769–1779). Landing and commerce were transferred to the Valongo region, in the city outskirts. By 1831, almost a million slaves had entered the city via this port and, with the first official ban on slave trade, the *Cais do Valongo* became obsolete. In 1843, the *Cais* was buried and gave place to the *Cais da Imperatriz* — it was through this new construction that D. Teresa Cristina de Bourbon, Empress consort of Emperor D. Pedro II, arrived. The site would be buried again during the revitalization of the port in the tenure of Mayor Pereira Passos (1902–1906), accompanied by many demolitions of low-income houses in the area, only to be rediscovered through excavations held in 2011.

in Brazil in the 1840s [31]. Regardless of its origin, the fact is that, at the start of the twentieth century, *maxixe* would take *lundu*'s place as Brazil's "national dance".[7] It was considered unsophisticated and sensual, due to the intense and abundant thigh and hip movements of the performers, but, unlike *lundu*, in *maxixe*, the couple dances together and simultaneously with other pairs. Also, the music accompaniment is purely instrumental (i.e., there is no singing) with the piano serving as base, instead of the *viola* or the guitar. With these characteristics, *maxixe* was able to conquer the variety entertainment theaters and *clubes carnavalescos* (*carnaval* clubs). At that time, these clubs were the city's main associations for *carnaval* practices, managed not by the poorest, but by medical students, civil servants and merchants [31].

The mix between European ballroom dances and Brazil's characteristic rhythm lead to hybrid genres such as *polka-lundu*, *polka-maxixe*, and *tango-maxixe*, which could arguably be grouped under the general label of "*maxixe*" music [31]. As mentioned above, this genre was often associated with vulgarity, which could explain why many composers that took interest in it (e.g., Chiquinha Gonzaga, Ernesto Nazareth) preferred to tag their pieces as *tangos* or *tangos brasileiros* (Brazilian

[7] Around the same time the word "*samba*" would replace "*batuque*" as the placeholder name for black community folk dances and festivities.

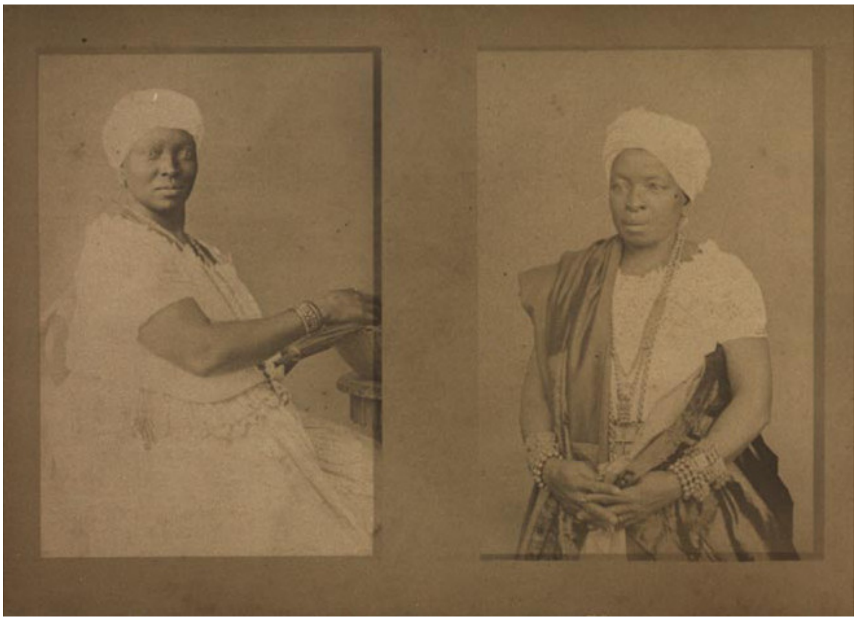


Figure 2.6: Black women in traditional *baiana* attire. Photographs by Marc Ferrez (ca. 1870–1899) [43].

tangos). The Platine rhythm was a contemporary of *maxixe* and they both bore rhythmic and choreographic similarities, at least until the 1920s when the *tango* went through a few transformations [31]. For a meticulous analysis of the appearance of the name "*maxixe*" and the uprising and persecution of this dance in the Abolition's aftermath, the reader is referred to a research by EFEGÊ [42] published in "*Maxixe – A Dança Excomungada*" (lit. "*Maxixe* – the Excomunicated Dance"). For the purpose of this work it suffices to mention that *maxixe*'s popularity slowly faded away after the "birth" of *samba* in 1917.

As *lundu* and *maxixe*, *samba* will walk into the clash between "popular" and "elite" cultures at the turn of the century, suffering a good deal of prejudice and discrimination [27, 31]. Initially considered an alien in the urban scenery — "*samba*" was associated with the country's Northeast region (specially with the state of Bahia), with the rural environment, and, of course, with "people of color" and dance of *umbigada* performances —, the genre will ferment in parties thrown at the houses of *tias baianas*;[8] slowly breaching through society's barriers and eventually invading the processions of *carnaval*, *samba* will spread from the capital throughout the entire nation. It is necessary to distinguish between two *samba* traditions in the

[8] *Tias baianas* (lit. aunts from Bahia) were black women that came to Rio de Janeiro from the state of Bahia. They were influencers and great supporters of the poor communities settled in the *Pequena África* region. The character of the *tia baiana* is so important to the foundation of *carnaval* and its traditions in Rio de Janeiro that still today they are annually honored by all the competing *escolas de samba* in the *ala das baianas*. This *ala* (wing) is composed of older women dressed in traditional *baiana* attires; in the first years of the competition, it also counted with the participation of men (in the same costume). Figure 2.6 presents two women in the traditional *baiana* attire.

first decade of the twentieth century: the *samba baiano* or *samba-de-roda* — older, folkloric and rural —, dance of *umbigada* whose first name will be used as a replacement for "*batuque*", as mentioned before; and the urban *samba*, the *samba carioca* — popular pair dance without *umbigada* (i.e., in *samba* couples dance together) —, which is *maxixe*'s successor in Rio's cultural scene after 1917. This is not to say that these two *sambas* are unrelated. In fact, as MOURA [44] points out, the *roda* should be credited as the "physical origin" of the first-generation *samba carioca*, the locus where the fusion of its "aesthetic roots" (polka, *maxixe*, *lundu*, etc.) will be unfolded. To put it another way, the *samba-de-roda* that came from the Bahian diaspora will be stylized in the capital by virtue of its cultural contact with dances that were popular in the city's nightlife.

Tia Ciata (1854–1924), born Hilária Batista de Almeida, was arguably the most famous *tia baiana*, whose life offers us a perspective on the genesis of *samba*. At the age of 22, Ciata left Bahia for Rio, bringing a young daughter, and soon settled with the Bahian colony in the Saúde neighborhood. Later she moved to Cidade Nova (near Praça 11 de Junho, see Figure 2.7) to a house that is considered by many researchers as the "capital" of *Pequena África* and birthplace of the *samba carioca* [27, 28, 31, 44]. The *tias baianas* were known to host many famed and intimate house parties — simply referred to as "*sambas*" — with food, drinks, music, and dance, either to celebrate a specific event or to congregate family and friends. In Tia Ciata's house, these *soirées* were split into three environments, with different atmospheres and entertainments, as listed here in an ascending order of intimacy [31, 44]: the living room was where *choro* was played and "balls" were held, with European-inspired dances where couples performed together; the dining room was the place for *samba-de-partido-alto* (*partido-alto*), which by then strongly resembled a dance of *umbigada*, according to written records [31]; finally, a *batucada* or *pernada* — a variant form of *capoeira* — took place in the *terreiro* (i.e., the outside yard behind the house), due to its violent component. It was also in the *terreiro* that spiritual ceremonies sometimes occurred. Ciata and her husband, João Batista da Silva (?–1910), had amicable relations with the elite — he was a public servant at the *Alfândega*; she was a sweets vendor and seamstress (and had several other *baianas* working for her in these two occupations), *mãe-de-santo*,[9] and main acolyte to a major *candomblé*[10] leader at the time. Later, through her healing works as *mãe-de-santo* for the president of the Republic, Ciata would get her husband a job at the office of the chief of police [27]. For these reasons, amongst her parties' invitees, one could find *sambistas*, intellectuals and journalists [44], officers

[9] *Mãe-de-santo* or *ialorixá* is a priestess of *candomblé* (see Footnote 10).

[10] Afro-Brazilian religion, derived from traditional religious practices first brought to Bahia by enslaved Africans. While spreading throughout country, these practices suffered many adaptations, including the adoption of syncretic elements from Catholicism.

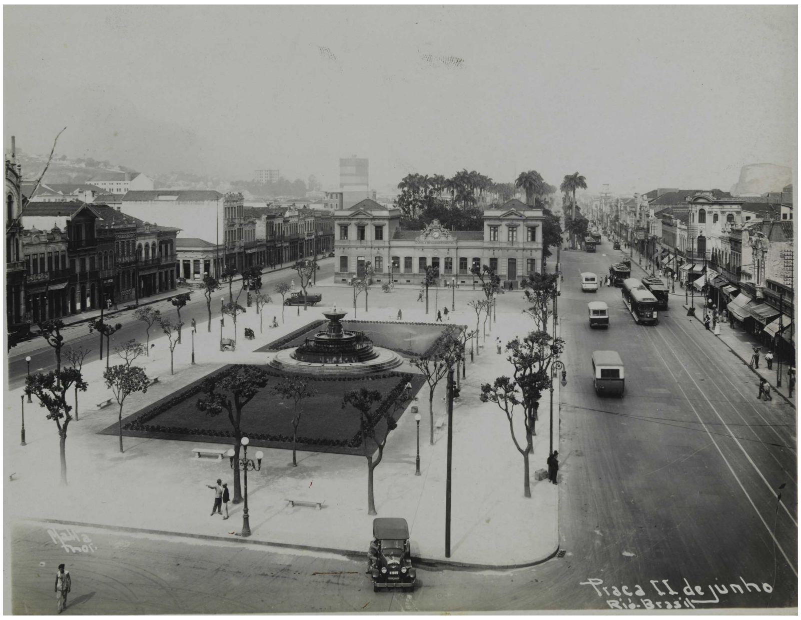


Figure 2.7: *Praça 11 de Junho - Rio.* Seen here in a photograph by Augusto Malta (ca. 1922) [45], this *praça* (square) was at the center of Cidade Nova neighborhood. It was named after the Battle of the Riachuelo (June 11, 1865), one of the major victories of Brazil in the Paraguayan war. Cidade Nova was first urbanized in the early nineteenth century, over a wetland, and by the 1850s, many factories would be built there. With the increase in population density throughout the city, many rich dwellers would abandon the region in search of more "salubrious" places [35]. At the turn of the twentieth century, Cidade Nova was a "busy" place: many *gafieiras* were established in the region (including *Kananga do Japão*). The square also became a very important place for the *samba carioca.* One of the streets around it would house tia Ciata, from 1899 until her death, and it would be used as a reference point for the *concentração*[11] of many *ranchos* and later *escolas de samba.* The Sambadrome is located about 300 m to the east of where the square once stood.

and prestigious people [31].

The familiar environment provided by Ciata's parties will be reproduced by *rodas de samba* in other houses and sometimes on the streets. The police watched closely these community gatherings, which were viewed by the dominant classes with much suspicion — not only the religious component of *candomblé*, or the dance–fight *capoeira* which was deemed violent, but also the *samba* and other music practices with an "African flavor". As we mentioned above, considering that Ciata and her husband had a good relationship with the elites, their parties stood in a very privileged position. However, in many other cases, the police will be called to action: dispersing *rodas*, reprehending and even arresting musicians (a few anecdotal examples can be seen in [34]). Consequently, *samba*'s penetration power in society will depend on the very same "upper class" and on the connections such as the ones es-

[11] *Concentração* is how insiders refer to the warm-up gathering that takes place immediately before a parade.

tablished in Ciata's house. The first *sambas* would be immortalized in the voices of acclaimed white singers/radio hosts [35]; and businessmen, observant to the trends, would seek the opportunity to finance incipient *samba* groups [31]).

In 1916, Donga (Ernesto Joaquim Maria dos Santos) registered "Pelo Telefone". Erroneously considered the first recording labeled as "*samba*",[12] it was without doubt the first "*samba*" hit, the one that put the genre on the map in the *carnaval* of 1917 [31, 34, 44]. The song's authorship was contested by Ciata, Sinhô and others; and Donga, who was a regular in Tia Ciata's house, would later admit that lyrics and melody were "developed from" the ones sung at those joyful meetings [31]. It was recorded by Casa Edison in the voice of Baiano,[13] yet this arrangement of "Pelo Telefone" more closely resembles those of *maxixe*'s instrumental music. As SANDRONI [31] observes, even though "Pelo Telefone" and the succeeding "*sambas*" recorded before the 1930s present more syncopated events, specially between two adjacent measures, when contrasted with the music sheets of older Brazilian genres (*lundu*, for example), they still exhibit a remarkable similarity with respect to the subjacent rhythmic structure. Its rhythmic cell was in fact so ubiquitous in the Brazilian music of the nineteenth and twentieth centuries that Mário de Andrade (1893–1945) — modernist writer and musicologist — commonly addressed it as the "characteristic syncopation" of the national pieces. In Section 2.4, we delve a little bit further into this structure, which is shown in Figure 2.12.

At the end of the 1920s, however, *samba* will witness the inauguration of a new rhythmic standard, which presented more syncopation or contrametric structures. The transformation into this second-generation *samba* was mainly lead by Ismael Silva, Nilton Bastos, Alcebíades Barcelos (Bide), Sílvio Fernandes (Brancura) and others, in the Estácio de Sá neighborhood, hence it is usually referred to as the "Estácio paradigm" [31]. This "new" *samba* would rapidly be absorbed by composers from other parts of the city, gaining terrain from the "old" *samba* represented by Tia Ciata's group. The confrontation between "old" and "new" can be synthesized in an emblematic inquiry conducted by CABRAL [34] in the 1960s. Cabral asked Donga and Ismael — exponents from each group — what was the true *samba*. The following discussion is reported below, in a free translation:

> Donga – Well, *samba* has long been this: "*O chefe da polícia / Pelo telefone / Mandou me avisar / Que na Carioca / Tem uma roleta para se jogar*" [lyrics to "Pelo telefone"].
>
> Ismael Silva – This is maxixe.

[12] NETO [35] shows that, from 1908 to 1915, at least twenty other songs were sold in records and labeled as "*sambas*". However, according to today's standards, given their arrangements, all these songs would be classified in other styles: *jongo*, *embolada*, *toada*, *choro* and — as it is the case of "Pelo Telefone" — *maxixe*.

[13] Interestingly enough, the first Brazilian recording of a song — the *double entendre lundu* "Isto é Bom" — was also produced by Casa Edison and voiced by Baiano, in 1902.

> Donga – What is *samba* then?
>
> Ismael Silva – "*Se você jurar / Que me tem amor / Eu posso me regenerar / Mas se é / Para fingir, mulher / A orgia assim não vou deixar*" [lyrics to "Se você jurar", a samba from 1931].
>
> Donga – This is not *samba*, it is *marcha*.

The average opinion among researchers and music critics is that Ismael was right: the pre-1930s *samba* is a "*samba amaxixado*" (i.e., that borrowed many elements of *maxixe* with regard to its rhythmic properties) or even, in a more extreme opinion, a false *samba* [31]. Sandroni gives an in-depth look at this discussion, contrasting the slight differences in the specialists' viewpoints (e.g., some argue that Sinhô's compositions in the early 1920s can already be considered *samba*). In another interview, Ismael told Cabral that this detachment from *maxixe*'s rhythm and influences was planned [34]:

> Ismael – When I started, *samba* wasn't good for groups of *carnaval* to walk the streets, like we see nowadays. [...] *Samba* was like this: *tan tantan tan tantan.* [...] Then, we started to play *samba* like this: *bum bum paticumbumprugurundum.*

We investigate this new paradigm in more detail in Section 2.4.

As we can see, the effect of the alterations introduced by this new style was thus not limited to its rhythm or even to its faster pace, allegedly modified for the ease of marching and dancing at the same time [35]. Additionally and more importantly, the second-generation *samba* allowed black and low-income communities to gain terrain, giving them voice in their fight for social justice. *Samba*'s venue had changed from the *roda* to the streets, particularly to *botequins* (bars) and *blocos carnavalescos*.[14] The sociability component — observable at the houses of the *tias baianas* — is still there, but it was perhaps now even stronger, after all, the street is a public space and everyone is invited to participate. Thus, the *roda* cedes its place to the *botequim*, the ball is replaced by the *desfile* (parade). This opposition between the *house* and the *street* is in the heart of Roberto DaMatta's anthropological studies of the Brazilian society. We refer the reader to the works by DAMATTA [29] and MOURA [44] for more in-depth studies of *carnaval* and *samba*, respectively, through the lens of this "Brazilian cultural dichotomy".

The Estácio group made many other contributions to *samba*, for its popularization and commercialization. For example, it is credited to this group of *sambistas* the invention of three typical *samba* instruments, *surdo*, *cuíca* [34] and *tamborim* [35],

[14] *Blocos* are a popular expression of *carnaval* in Brazil, and are usually organized by a street band that parades while performing *samba* or *marcha* music. Countless revelers are attracted each year to these block processions, which occur before, during, and after the *carnaval* days.

though this inventorship was never proved.[15] Most notably, they created "Deixa Falar", a *carnaval* association historically considered the first *escola de samba* [34]. Deixa Falar paraded from 1929 to 1931 and it paved the way for the modern *escolas de samba. Samba* continued to be transformed in the twentieth century, but the rhythmic framework established by the Estácio group subsists, so that the recent transformations are not defining of the genre. Thus, after a bit of ear training, any listener should be able to easily identify the majority of the *sambas* composed from 1930 to the present day as pertaining to the one and same genre, while in turn classifying the collection of pre-1930 *sambas* in a separate group.

2.3 *Carnaval*

Brazilian *carnaval* practices originated in colonial times with the *entrudo*, shown in Figure 2.8. From the latin *introitus*, meaning "entrance", the *entrudo* lasted about four days, from Saturday to Ash Wednesday, and served as prelude to the Lenten season.[16] Brought by the Portuguese from Europe probably in the mid-seventeenth century, it was a playful (sometimes violent) period, a sort of war, in which revelers would throw water, flour, *limões-de-cheiro* (wax or plastic balls filled with scented water), and sometimes even bodily fluids [34] at one another. Needless to say that the *entrudo* was strongly combated by civil authorities. In some cases, the punishments given to revelers depended on their social status, and enslaved people usually had the worst of it [34].

By the nineteenth century, after a series of prohibitions on the *entrudo*, several other *carnaval* activities were being developed. This transformation of *carnaval* in Brazil was first led by the upper classes, which, in search of a more civilized way of enjoying the *carnaval* period, started mimicking the festivities that were common in Europe, especially in France and Italy. Thus, starting in 1840, *Arlecchinos*, *Colombinas* and other *commedia dell'arte* characters flooded the masked balls that

[15] Conversely, the existence of large bass drums similar to *surdos* in African cultures was reported by European explorers of the eighteenth century [35]. Likewise, friction drums were not uncommon in the music of Angola, and a prototypical *tamborim* could already be found in northeast Brazilian folklore [35].

[16] Lent, in Christian tradition, is the penitential period of preparation for Easter. During approximately 40 days, a symbolic number in itself, the faithful are invited to join in and contemplate the mystery of Jesus' journey into the desert by partaking in ascetic and repentance practices. These include prayer, fasting, and more distinctively the abstention from the consumption of meat. This period is bounded by different celebrations: its beginning is the *dies cinerum*, the Ash Wednesday, when the priest marks the forehead of each faithful with blessed ashes ("thou art dust and unto dust thou shalt return"); and it leads to the paschal *Triduum*, the most solemn time and center of the ecclesiastical year, where the institution of the Eucharist, the Crucifixion and Death, and the Resurrection of Christ are celebrated. Shrovetide/"Carnival" practices such as the *entrudo* are commonly celebrated just before Lent, as a "natural" and permissive counterpoint to this period of penance.

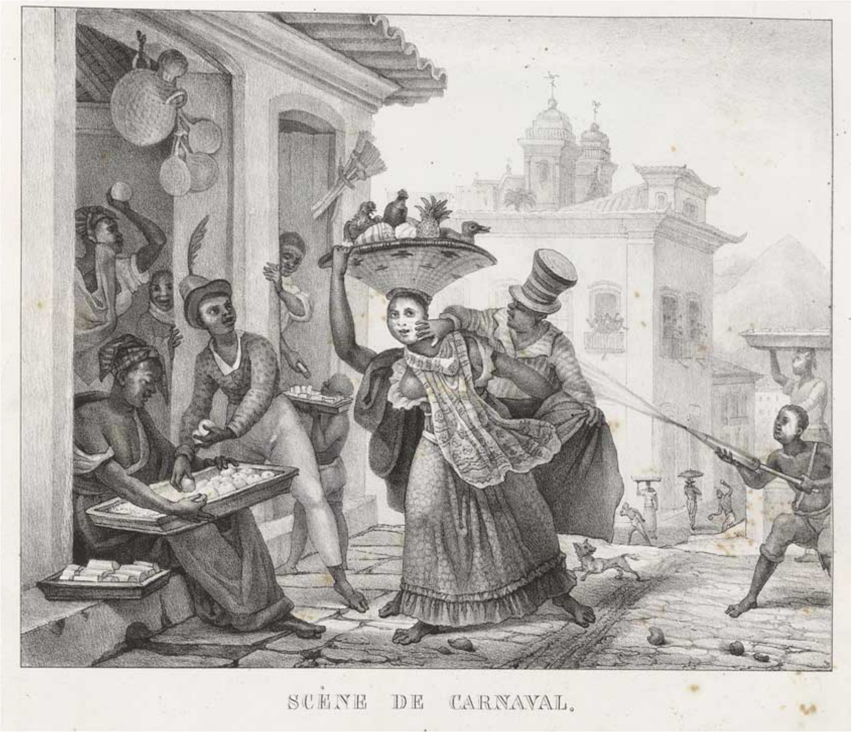


Figure 2.8: *Scène de Carnaval* (*Dia d'entrudo*), lithographic print (Debret, 1835 [46]). Engraved by Thierry Frères after a drawing by Jean-Baptiste Debret. Debret describes the scene typically seen in Rio during the days of *carnaval.* Here, a young black woman is seen balancing a basket of provisions on her head, while being attacked by a young man and a *moleké* (sic), who throw *polvilho* (starch) and water at her. Other characters are shown to the left, also marked by dirt and other matters; some of them are readying a retaliation. A vendor of *limões-de-cheiro* and *polvilho* is sitting on the step of a spices boutique.

were thrown in the large halls of hotels or in theaters, in the capital and later in other urban centers. Around the same time, the *carioca* elite would also create the first *clubes carnavalescos*, alternatively called *sociedades carnavalescas* (*carnaval* societies) or *Grandes Sociedades* (lit. Great Societies), which promoted this kind of parties but also parades leading to them. These parades would soon become an event in itself, either in a "simple" format with open cars (*corsos*), or a more complex one, involving the use of music and allegorical floats. The press remarked how Rio de Janeiro's *carnaval* rivaled the festivities that took place in Nice, Venice and Rome (qtd. in [34]). The "victory", in the *carioca* society, of these practices over the popular *entrudo* would be furthered by the Haussmannian renovations of Pereira Passos (who was, in turn, honored in a few processions) [35], although traces of the latter would survive until today. Hence, instead of liquids of dubious provenance, the throwing of *confete* (confetti) and *serpentina* (serpentine), frequently seen at European masked balls of the time, became common in *carnaval* activities, although subject to the characteristic playfulness of *entrudo* customs.

Entrudo would fade out faster as other popular manifestations appeared in the late-nineteenth-century Rio de Janeiro. These groupings would give people from underprivileged neighborhoods a way to express themselves during *carnaval*; but similarly to what happened with the *entrudo*, these popular practices faced repression and were closely surveilled by the police, specially in the *Belle Époque* period. One of these traditions, the *cordões*, were usually composed by masked revelers dressed in different archetypal costumes (elders, clowns, kings, queens, etc. [34]), who danced to percussive music while following a master [32]. The revelers of *cordões* were known for disturbing the public order, their identities protected by the masks, and even for attacking other *cordões* in an attempt to steal their banners [35]. The creation and development of this movement was concomitant with that of *ranchos*, and in fact at the beginning both were very similar in form; but eventually all *cordões* would either disappear, converge with the latter (which became better organized and "more civilized" with respect to singing, dancing and their creative aspects [47]), or change names, becoming *blocos* [48].

Ranchos would pass through a series of adaptations in order to survive the changing times, reaching tremendous popularity and dominating Rio's *carnaval* for over 30 years [34]. Many of these innovations would end up as significant contributions to the *carnaval* as it is known and played today. They introduced, for example, the use of string and woodwind instruments in the orchestra that accompanied their activities. In 1908, one of the most famous *ranchos*, Ameno Resedá, would inaugurate the structure of a complex parade: it involved an *enredo* (plot), which in turn required plot-related music and costumes [47]. The following year, the first contest between *ranchos* would be promoted by Jornal do Brasil, a traditional newspaper and publisher [47]. As a result of their aesthetic renovations, they were held by society and authorities at a higher "moral ground" when compared to *cordões* — and unlike those, many *ranchos* were given by the police official licenses for parading [47]. *Cordões*, *ranchos*, *blocos* would customarily perform at Praça 11 and many, while parading, would pass in front of Tia Ciata's house (and other *baianas*') to pay their respects to these important cultural characters of the time [34].

We now take a step back and look at the music being played at all these events. In the case of the *entrudo* and other simple cultural manifestations, there was usually no music at all or just a single musician playing his drum. At the end of the nineteenth century, the elite was dancing to waltz, polka, and other European rhythms at the masked balls, but the avantgardist *Grandes Sociedades* would introduce *marchas* and even *maxixes* in their parties and street parades. The *marcha* was also the predominant rhythm in *cordões*, *blocos*, and *ranchos*,[17] although *maxixes* were rather common as well. As we mentioned before, *samba* would only be incorporated in parades and later take over as a fundamental rhythm in *carnaval* with the transformation promoted by the Estácio group and Deixa Falar.

Deixa Falar was founded in the late 1920s under the name *Escola de Samba Deixa Falar*. Despite its status as the first *escola de samba*, it paraded as a *bloco carnavalesco* and, later, as a *rancho*, in 1932 [34]. The meaning behind this title — "*escola de samba*" — is related to the founders' intention of creating a new standard for *carnaval* festivities. Ismael Silva suggested in an interview [34] that the name might have been his idea, inspired by the normal school that was located near the association's headquarters. Instead of training high school teachers, the musicians from *Estácio* were themselves "*professores*" (teachers) and "*mestres*" (masters) of *samba* [34], thus deserving of their own school. The term became widespread and many *escolas* were created the following years in *morros* and outskirts of Rio. Examples include *escolas*: *Estação Primeira* (from Morro da Mangueira); *Azul e Branco*, *Depois Eu Digo*, and *Unidos do Salgueiro* (Morro do Salgueiro); *Prazer da Serrinha* (Morro da Serrinha), from which stemmed Império Serrano; and *Vai Como Pode* (from Oswaldo Cruz neighborhood), later Portela. These popular associations for the practice of *samba* were then and still are strongly connected to a specific community where their social events take place and to whom they provide several services.

The first *Desfile das Escolas de Samba* — the organized competition between *escolas de samba* — took place in 1932, promoted by journalist Mário Rodrigues Filho [34]. Deixa Falar had already been "upgraded" to a *rancho*, and did not take part in the competition [34]. During the parade, each *escola* presented two

[17] In the case of *ranchos*, due to, among other things, the different instrumentation, the rhythm would be specifically called *marcha-rancho* [48].

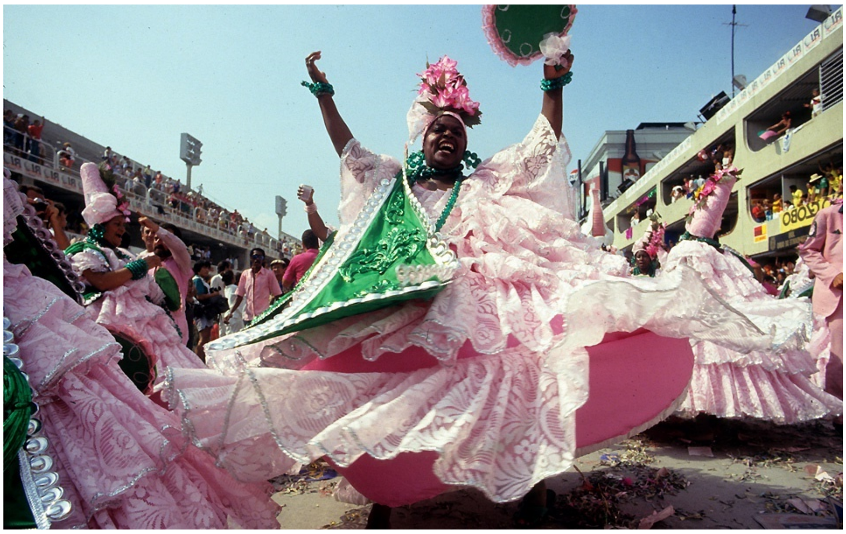


Figure 2.9: *Baianas* parading for Mangueira (1987). *Escolas de samba* pay homage to the matriarchs of *samba* in this obligatory *ala*. Compare the costumes shown here (in green and pink, the *escola*'s colors) with the traditional attire of Figure 2.6. Photo by Otávio Magalhães, on Flickr (CC-BY-SA).

to three *sambas*, structured with a fixed refrain and improvised verses [31, 34]. These *sambas* were not necessarily composed to the parade, and thus, in general, they were not related to its plot. The *samba-enredo* — a *samba* composed to be performed at the parade, and subject to the *enredo* (plot) — was only introduced in the late-1940s, characterizing a new *samba* subgenre. As we already mentioned, the *Desfile* is the high point of today's *carnaval* in Rio de Janeiro, and arguably in the entire country. Currently, it consists of a two-day parade (*carnaval* Sunday and Monday) when the *escolas* that make up the *Grupo Especial*[18] (lit. Special Group) parade with their components, organized in *alas*. Examples of *alas* are the *ala das baianas* (which pays homage to the matriarchs of *samba*) and the *ala da bateria* (drum ensemble); these are respectively illustrated in Figures 2.9 and 2.10. There are around three thousand members per *escola* in a parade, each dressed according to the *ala* they're in. Most members in a given *ala* will be dressed in the same costume and parade on the *Avenida* (Avenue), while singing the *samba-enredo* of the *escola* and dancing a choreography (also *ala*-specific); however, some members, called *destaques*, dress in more expensive and personalized costumes and are carried in *carros alegóricos* (allegorical floats), illustrated in Figure 2.11. The music, the choreographic elements, and each set of costumes and floats help telling the *enredo* (plot) chosen by each *escola*. During the *desfile*, each *escola* is subject to a series of constraints (e.g., parade duration,[19] number of members per *ala* and in total,

[18] The "*Grupo Especial*" is the first division of *escolas de samba*. Other *escolas* are arranged in further divisions, which are commonly known as *Grupos de Acesso* (lit. Access Groups). The second division also parades at the Sambadrome, but on a different date.

[19] From 60 to 70 minutes, in 2020.

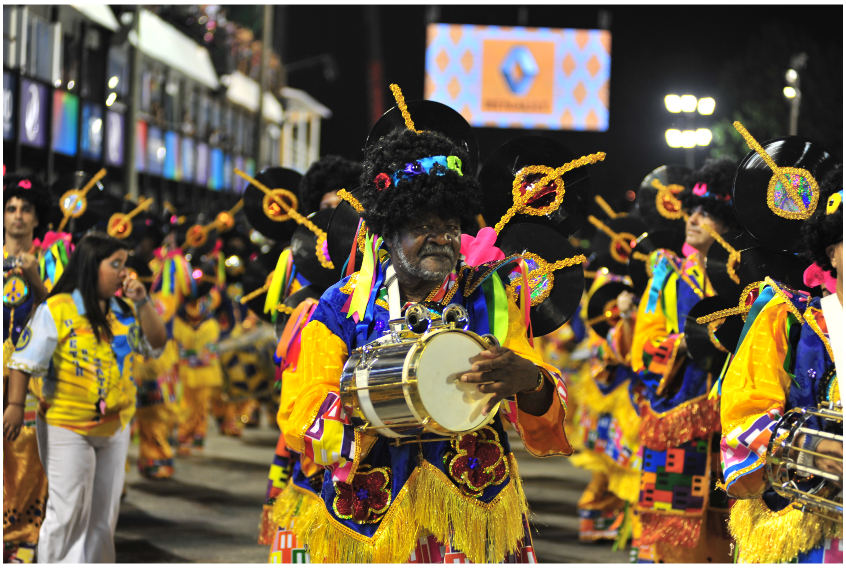


Figure 2.10: *Cuícas* from Paraíso do Tuiuti 2013 parade. All *alas* are dressed in costumes tied to the parade's *enredo*. Photograph by Circuito Fora do Eixo, on Flickr (CC-BY-SA).

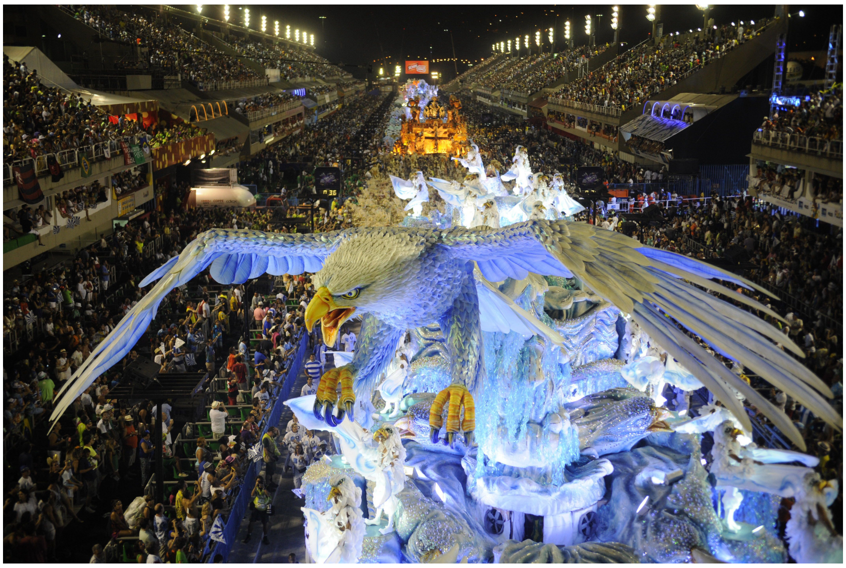


Figure 2.11: *Águia da Portela.* The symbol of this *escola* is the eagle ("*águia*", in Portuguese), which is always represented in the *abre-alas*, the first float. Photograph by Fernando Frazão/Agência Brasil (CC-BY-BR).

number of floats) and is evaluated by a committee with respect to a few criteria (e.g., *bateria*, *enredo* and *samba-enredo*, costumes, singing, dancing) in a strict point system. Evaluations are collected and the results are presented on Ash Wednesday — the winning *escola* and a few runner-ups then parade again the following weekend at the *Desfile das Campeãs* (lit. winners parade).

2.4 Rhythm

"*Música, o samba caracteriza-se pelo constante emprego da síncopa*" [Music, *samba* is characterized by the constant use of the syncope] [49]. The opening statement of *Carta do Samba* (lit. the *Samba* Letter) — the final document of the *I Congresso Nacional do Samba* (First National *Samba* Conference) of 1962 — is very clear about the main characteristic of this rhythm: the syncope. In fact, as we previously mentioned, the syncopation is one of the most prominent aspects of the music developed this side of the Atlantic, fruit of the miscegenation of different — e.g., European, African — rhythms. In the particular case of Brazil, this element of musical rhythm has been broadly studied, in the popular genres from the eighteenth to the twentieth century, by scholars such as Mário de Andrade, who carefully studied this notion of a "characteristic" syncope (qtd. in [31]) and its origins. This motif, composed of sixteenth note–eighth note–sixteenth note,[20] can be seen in Figure 2.12.

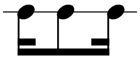
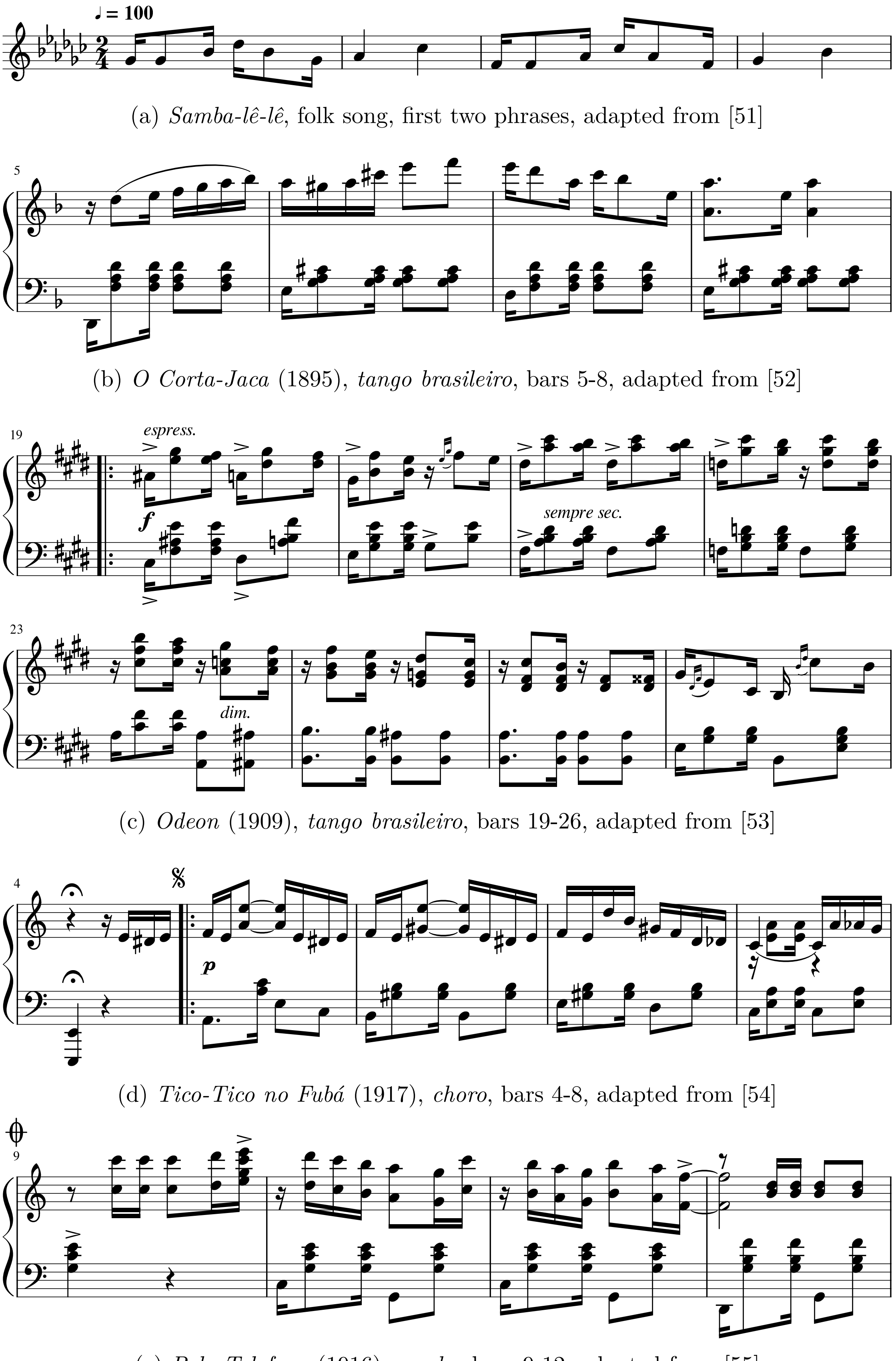


Figure 2.12: The "characteristic" syncope.

CANÇADO [50] analyzes typical syncopated rhythms of Brazilian music and their syntagmatic relations, by comparing *modinhas*, *lundus*, *tangos*, and *choros* to traditional rhythm cycles from Angola and West Africa. In her findings, she describes how both the "characteristic" syncope and the *habanera* rhythm (see Figure 2.14e), another rhythmic cell very familiar to mid-nineteenth- and early-twentieth-century Brazilian music, assume different functions in the studied corpus. The "characteristic" syncope followed by two eighth notes, and the *habanera* rhythmic cell serve as accompaniment; whereas two side-by-side "characteristic" syncopes and other variants including *habanera* syncope normally act as the melody line [50]. Figure 2.13 presents a few examples where both functions can be observed.

[20] Our nomenclature in this section diverges slightly from that of Sandroni, who conducted a highly recommended study in both his thesis and on "*Feitiço Decente*" [31]. Sandroni uses the term "characteristic syncope" for the whole phrase/bar shown in Figure 2.14a, whereas we call the same pattern "*maxixe* rhythm". We, however, understand the "characteristic syncope" as the short–long–short or 1+2+1 pattern (see Figure 2.12) that is featured on the first beat of the $\frac{2}{4}$ measure, following the description given by CANÇADO [50].



(a) *Samba-lê-lê*, folk song, first two phrases, adapted from [51]

(b) *O Corta-Jaca* (1895), *tango brasileiro*, bars 5-8, adapted from [52]

(c) *Odeon* (1909), *tango brasileiro*, bars 19-26, adapted from [53]

(d) *Tico-Tico no Fubá* (1917), *choro*, bars 4-8, adapted from [54]

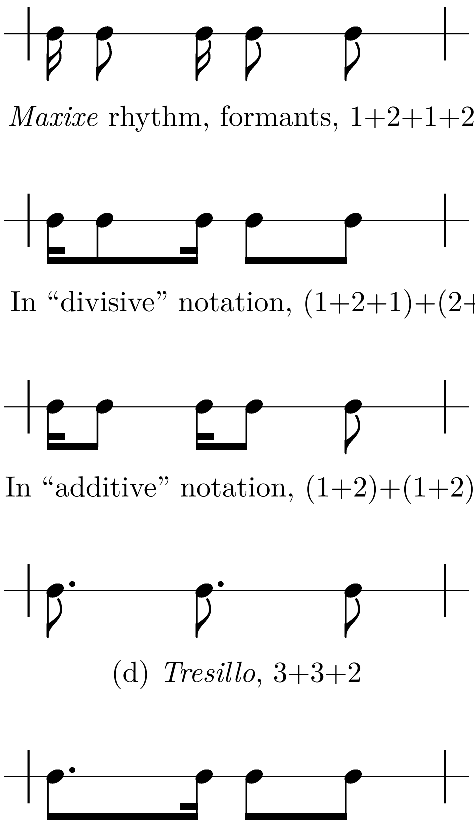
(e) *Pelo Telefone* (1916), *samba*, bars 9-12, adapted from [55]

Figure 2.13: Excerpts from Brazilian music showing the use of the characteristic syncope, the *maxixe* rhythm and the *habanera* rhythm both in the melody and in the accompaniment.

We shall take some time now to investigate the "characteristic" syncope, especially in the accompaniment scheme that was vastly featured in *maxixes* and first-generation *sambas*, for example, and how it was transformed at the end of the 1920s. We will henceforth refer to it as "*maxixe* rhythm", not because it was exclusive to this genre (as evidenced by Figure 2.13), but since it is in great part responsible for the "*amaxixado*" adjective given to first-generation *sambas*. As Sandroni argues [31], at first glance, it looks as though the *maxixe* rhythm is in a perfectly symmetric subdivision of the $\mathbf{\frac{2}{4}}$ measure, as shown in Figure 2.14(b). However, when the first-generation *samba* was in vogue, the "cultural reading" given to this characteristic syncope was one of rhythmic imparity, i.e., when binary and ternary low-level pulsation groups are interpolated [31]. This is better expressed by Figure 2.14(c), where the low-level events are agglutinated as (1+2)+(1+2)+2. The *maxixe* rhythm can then be regarded as a variant of the 3+3+2 *tresillo* rhythm[21] (Figure 2.14(d)) in which the ternary figures have been further subdivided into (1+2) groupings. The *habanera* rhythm can be understood in a similar fashion, when, starting with the *tresillo*, the second ternary figure is divided into (2+1) instead. With an abundant presence of these and yet more variants, the Brazilian music of the nineteenth century and the first decades of the twentieth century was conformed in what Sandroni calls a "*tresillo* paradigm" [31]. In particular, his analysis of *sambas* recorded from 1917 to 1921 indicates the presence of the paradigm variants in both the melody and the accompaniment [31].

A different paradigm was established by the *Estácio* group, as alluded by Ismael Silva in the interview transcribed before. The main property of this paradigm: its rhythm is more complex ("*bum bum paticumbumprugurundum*"), that is, more contrametric, than those belonging to the *tresillo* family. The new rhythm had a basic cycle of 16 sixteenth note pulses, spanning two measures in $\mathbf{\frac{2}{4}}$ time (against the 8 sixteenth note-cycle and single measure of the *tresillo*). The literature usually presents this cycle segmented in "unequal halves", either as 7+9 or as 9+7. We present two of these variants in Figure 2.15; yet, similarly to what we saw with the *tresillo*, the groupings can be divided differently or replaced with dotted figures. SANDRONI [31] also shows another variant, segmented as 5+11, which he associates with a transitional period between the two *samba* generations. All this variants were grouped by him in what is called the "Estácio paradigm". When listening to *samba* recordings dating from 1927 to 1933, Sandroni was able to find many examples of this paradigm, not only in the percussion, but also in the melody. This means that syl-

[21] Despite being habitually associated with Cuban music, the *tresillo* it is in fact present in many other Latin American countries marked by the Atlantic slave trade [31]. Its main characteristic is a strong and contrametric accent at the fourth sixteenth note level pulsation; this creates a sensation of imbalance (or imparity) of the syncope because this results in a segmentation of the 3+3+2 rhythm as 3+5.



(a) *Maxixe* rhythm, formants, 1+2+1+2+2

(b) In "divisive" notation, (1+2+1)+(2+2)

(c) In "additive" notation, (1+2)+(1+2)+2

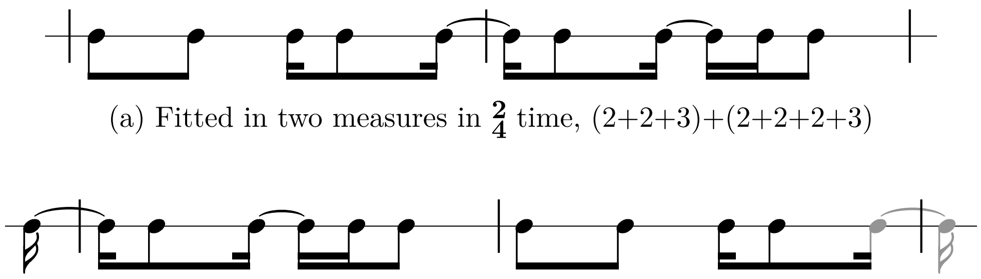
(d) *Tresillo*, 3+3+2

(e) *Habanera* rhythm, 3+(1+2)+2

Figure 2.14: *Tresillo* paradigm and the *maxixe* rhythm, the main rhythmic pattern in Brazilian music until the 1930s.

lables are articulated in a way that "suggests that [rhythm] of the *batucada*", which is highly contrametric. These rhythmic patterns from the Estácio paradigm (and also the ones from the *tresillo* paradigm) can be interpreted in light of what, in African and Afro-diasporic music, is called a "timeline" [56] — asymmetric rhythmic motifs, produced by hand clapping or by instruments (e.g., bells, high-pitched drums), that are repeated in *ostinato* and form a structural matrix around which the performance is organized.

Still with respect to this matter of contrametricity/commetricity in *samba*, there is a serious tendency in *sambas-enredo* and *escolas de samba* which is being detected and debated by scholars and musicians for at least the last 30 years. Several musicological studies [57, 58] have observed a considerable increase in the average tempo of live *samba-enredo* performances, an effect that is attributed to an increase in the number of paraders and to stricter parading time constraints. This has lead to the presence of more commetric structures specially in the choruses melodies, in a phenomenon called *marcialização do samba* [57] (i.e., the transformation of *samba-enredo* into *marcha*) or even derogatorily *marcha-enredo* [58]. This goes against the main postulates expressed at the beginning of *Carta do Samba* [49]: (1) to pre-



(a) Fitted in two measures in $\frac{2}{4}$ time, (2+2+3)+(2+2+2+3)

(b) Shifted with respect to the measure start point, (2+2+2+3)+(2+2+3)

Figure 2.15: Estácio paradigm variants. The pattern in (a) is also referred to as "*tamborim* pattern", since it corresponds to this instrument's *teleco-teco* cycle (see Figure 2.19c), despite also being found in lines produced by other instruments and the melody. Adapted from [24, 31].

serve the traditional characteristics of *samba* we must value the syncope; (2) mixed rhythmic forms (e.g., *samba-choro*), albeit legitimate, cannot replace *samba.*

As we have largely expressed in this section, the word "rhythm" relates to the way events in music are organized in time. In a more specific sense, it can describe a pattern of attacks "constrained overall by a meter or associated with a particular tempo" [3] or yet patterns of duration, which are based on the inter-onset intervals between successive events [13]. However, when we investigate the sound production and the language used by members of *escolas de samba*, the notion of "rhythm" (*ritmo*, in Portuguese) admits still another interpretation: it designates the main sonorous organization of the *bateria* ensemble, i.e., how different instruments interact and their sounds are superimposed [59].

The *ritmo* of a *bateria*, its sound organization "logic", is regarded by scholars as a practice

> akin to much of the percussion ensemble practices spread throughout sub-Saharan Africa. Generally speaking, cyclical individual parts are assigned to each instrument (or, in the *bateria* case, multi-piece instrument set), patterned upon a cluster of rhythmic, timbral, and, to a lesser extent, pitch qualities. [24, p. 130]

Each individual part in this superposition is, therefore, a recurrent rhythmic/timbral cycle that constitutes what performers call the instrument's *batida* [24] or *levada* — a term used not only for instruments in the percussion ensemble, but also for the main harmonic accompaniments provided by the *cavaquinho*[22] and the guitar, for example.

[22] *Cavaquinho* is a small four-stringed instrument of the guitar family. Originated in Portugal, it became very popular in Brazil, and was later introduced into Hawaii (where, in the nineteenth century, it led to the development of the ukulele) [3].

In the following section we describe the main instruments used in *samba* and, in particular, in an *escola de samba*. We then present a few of the *batidas* that are superimposed for the sonorous achievement of a *bateria*.

2.5 Instruments

Here we describe some of the main instruments used in *samba*. Even though we will specifically define how each one is used in an *escola de samba* setting, most of these instruments (or even better their functions) can be found in every *samba urbano* performance, whether it is a *roda* or a *bateria*.

Being related to African music practices, *samba* is mostly known for instruments of the membranophone class, such as *tamborim* and *surdo*. Nevertheless, one can also find idiophones (e.g., *agogô*, *reco-reco*) and chordophones (guitar, *cavaquinho*, etc.) in the usual practice, and even aerophones on rare occasions. Regarding their construction process, instruments used in *samba* today are almost all industrially made, but this has not always been the case [24]. For example, modern drumheads are made from plastic or acrylic, while drum shells are made of metal. Older instruments, however, were manufactured with membranes of calf or cat skin, and had wooden structures. Oftentimes, instruments were improvised from materials and even objects (e.g., tableware and cooking utensils, matchboxes) that were available to the impoverished population. The newer technologies employed in mass-produced *samba* instruments have also made the tuning process of these instruments more reliable, with the introduction of simple systems of bolts and counter hoops. In this section, we will limit our scope to the set of percussion (membranophones and idiophones) instruments used in *samba*, which are the focus of this thesis, and, in the following, we particularly describe the main instruments in their more common, newer, format (although some specifications also apply to older instruments) [24, 28, 60]:

- *Agogô* (Figure 2.16a): from the Yoruba word for “bell”, the *agogô* is an idiophone made by attaching together a few clapperless bells. The most common *agogô* contains two bells connected by a U shape piece of metal — these are also seen in *rodas de samba*. However, some *escolas de samba* employ *agogôs* with up to four bells (e.g., Império Serrano). Each bell produces a different pitch and, in the case of the two-bell *agogô*, the player can be press bells together creating a metallic clicking sound.

- *Chocalho* (Figure 2.16b): an indirectly struck idiophone. In *samba*, *chocalhos* are commonly found in two forms: *ganzá*, which is similar to a rainstick — a hollow (single-, double-, or sometimes, triple-) tube that is partially filled with grains, pebbles, beads or something of the like; and *rocar* (generally called simply "*chocalho*"), made with a wooden rod or an aluminium frame in which a series of jingles are attached. This latter form is more frequently found in *baterias*.

- *Cuíca* (Figure 2.16c): a friction drum, most commonly found with a 20 to 25 cm (8 to 10 in) single drumhead, although smaller and larger sizes can also be seen. A wooden stick is connected to the drumhead and lies inside the drum. Sound production is achieved by rubbing on this stick with a dampened cloth. The free hand can be used to change the pressure on the drumhead, varying the instrument's pitch.

- *Pandeiro* (Figure 2.16d): similar to a tambourine, this is a large hand-held frame drum with leather or synthetic head and jingles along its body. It is struck with the thumb, the fingertips, the heel, or the palm of the hand. In plays an important part in small settings; in *baterias* it is normally relegated to choreographic devices, not playing an important part in the sound formation.

- *Reco-reco* (Figure 2.16e): a scraper instrument. It comes in tubular form (made of wood) or as a metallic base over which one (or more) springs are fastened. It is scraped with a wooden or a metallic rod. Seen in both *roda* and *escola* settings, with the metallic instrument being more common in the latter.

- *Tamborim* (Figure 2.16f): a small frame drum of 15 cm (6 in) in diameter and a 4 cm deep frame. It is struck with a small wooden drumstick or two-/three-stemmed plastic stick. Sound is produced by striking the drumhead with the stick, while the fingers in the holding hand can be used to change the timbre. Not to be confused with the tambourine (*pandeiro*).

- *Caixa* (Figure 2.16g): a snare drum. *Caixas* come in different sizes, regarding not only the head diameter (10 to 14 in) but specially the distance between the two heads (for *tarol*, about 9 cm; *malacacheta* or *caixa-de-guerra*, 15 to 20 cm). It is played with two drumsticks and the "snare effect" is usually achieved by means of a set of strings placed across the top head. Fundamental in a *bateria*, this instrument is rarely seen in a *roda* setting.

- *Repique* (Figure 2.16h): also called *repinique* or *surdo repicador*, it is a medium-sized two-headed drum. Head size can vary from 8 to 12 in (more common), with the distance between heads usually set at 30 cm. It is alternately struck with a drumstick (on the center of the head, near the edge of the instrument or simultaneously on the rim) and the free hand.

- *Surdo* (Figure 2.16i): large bass drums, the *surdos* have the lowest pitch in a drum ensemble. In *baterias*, this instrument is normally divided into three sections — each tuned in a unique way and executing a different pattern: *surdo de primeira* or *de marcação* (lit. first/marker *surdo*), *surdo de segunda* or *de resposta* (lit. second/response *surdo*), and *surdo de terceira* or *de corte* (lit. third/cutter *surdo*). They are played with a mallet and muffled with the free hand.

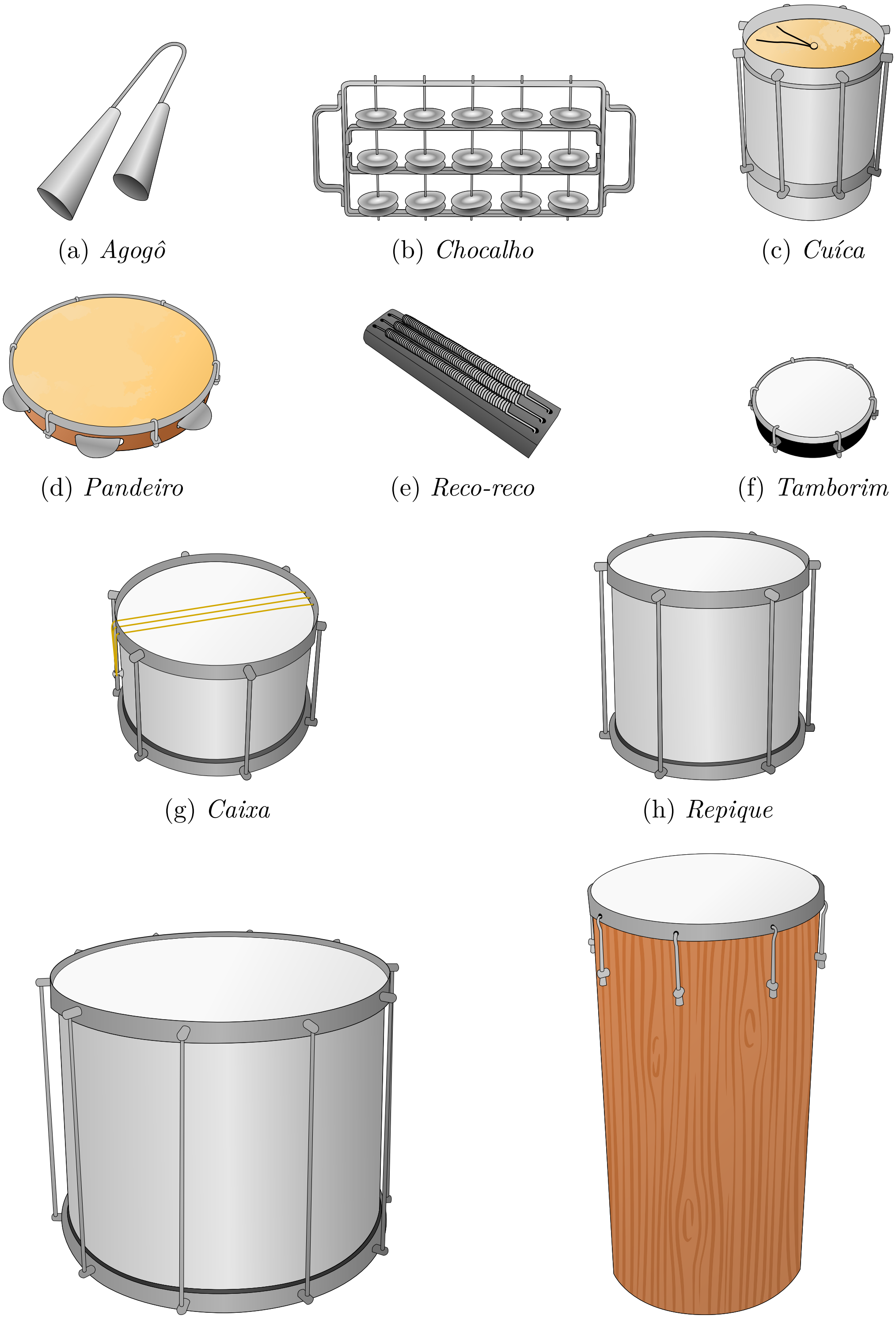
- *Tantã* (Figure 2.16j): a cylindrical hand drum more commonly used in the *roda*. It has only one head (12 to 14 in) and its use can be compared to that of *surdo de primeira* in a *bateria*.

Other percussion instruments that have been historically used in *baterias* or in *rodas* are: pratos (standard hand-held cymbal pairs, in *baterias* only), *frigideira* (frying pan) and *prato-e-faca* (kitchen knife and fork), which have recently displayed a reduction in usage. In *rodas de samba* and *pagodes*, we can also find *repiques de mão* and *repiques de anel* (respectively, hand- and ring-*repiques*).

The *bateria* is widely considered the “heart” of an *escola de samba*, not only for enlivening the other members in a *desfile*, but also because it is responsible for keeping the balance between singing, dancing and the “flow” of the *escola* in the *Avenida* [24]. The faster the *bateria* plays, the faster the members in other *alas* (wings) will parade. In a *desfile*, a *bateria* usually has around 300 instrumentalists (*batedores* or *batuqueiros*, “beaters”), several conductors — the *mestres* (masters of percussion), and the *diretor de ala* (wing director) also called *diretor de bateria* or *primeiro-mestre* (first master). Instruments in a *bateria* are referred to as *peças* (pieces) and, usually, the *mestres* in an *escola* are required to learn how to play all or most of the different *peças* [24], so that they can teach other instrumentalists and also conduct them in rehearsals and competitions.

In a *bateria*, *peças* can be further classified into two groups: (1) *miudezas* (minutiae) and (2) *couros-pesados* (lit. heavy leathers) [24]. *Miudezas* are all the hand-held instruments (e.g., *cuíca*, *tamborim*, *chocalho*) while *couros-pesados* correspond to the loudest segment of the drum section (*caixas*, *repique* and *surdos*), which act in a lower to middle register. We can also categorize the *peças* with respect to their function in a *bateria* [60]:

- *Surdos de marcação* are the foundation of the *bateria*, playing the second beat of *samba*’s duple meter. *Surdos de resposta* play the first beat and *surdos de corte* play along with the *surdos de marcação*, but in more complex patterns. The three *surdos* are not featured in all the *escolas* (e.g., *surdos de segunda* are not present in Mangueira’s *bateria*). The way these instruments are tuned



(a) *Agogô*

(b) *Chocalho*

(c) *Cuíca*

(d) *Pandeiro*

(e) *Reco-reco*

(f) *Tamborim*

(g) *Caixa*

(h) *Repique*

(i) *Surdo*

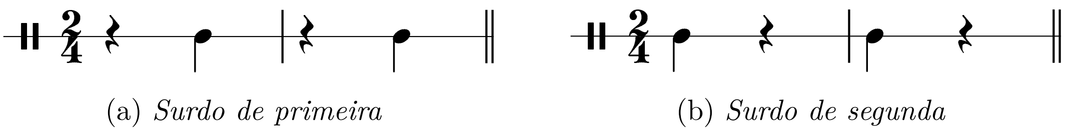
(j) *Tantã*

Figure 2.16: A few percussion instruments used in *samba*.

can also vary (e.g., *surdos de primeira* are usually tuned to a pitch lower than that of *surdos de segunda*, but in Mocidade, this relation is inverted);

- *Repiques* back up the *surdos* and are used in many moments to cue the *bateria*;
- *Caixas* play rhythmic figures that are characteristic of certain *escolas de samba* and (along with *surdos de corte*) play a great part in the *bateria* groove;
- *Pandeiros*, *cuícas*, *agogôs*, *reco-recos* help sustaining the rhythm and may also improvise phrases;
- *Tamborins* not only sustain the rhythm (as the other *miudezas*, but also play certain *convenções* (riffs) that highlight the melody of the *samba-enredo.*

Figures 2.17 to 2.21 illustrate a few patterns for the *surdo–tamborim–cuíca* trio, which is representative of the second-generation *samba*, and also for the *caixas*, which indicate an *escola's* identity. These patterns were transcribed to a score notation, but this surely brings many limitations since notation practices are unfamiliar to *samba*, which is traditionally transmitted in *escolas* without written aids [59]. Furthermore, traditional notation systems were developed for and describe well the intricate melodic-harmonic relations expressed in the common-practice music, which are without parallel in a *bateria* ensemble [59]. However, besides the rhythm, several other aspects (e.g., melodic, dynamic) are involved in the performance of each instrument's *batida*, which makes this notation issue even more problematical. Nevertheless, the aforementioned examples were notated in $\mathbf{\frac{2}{4}}$ time signature, widely applied when notating *samba*, although some scholars prefer using quaternary meters (see [59] for a discussion on this matter). In either structure, the sixteenth note is normally associated with the fastest pulse.



(a) *Surdo de primeira*

(b) *Surdo de segunda*

Figure 2.17: Cyclical parts for *surdo de primeira* and *de segunda*. Adapted from [60].

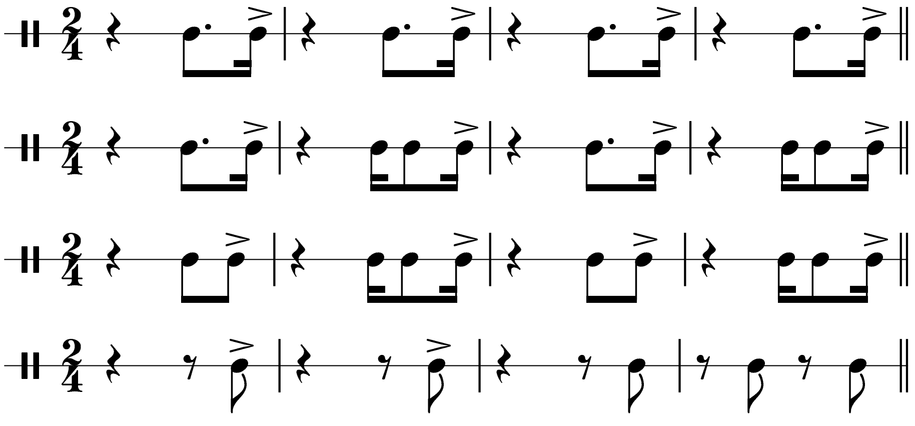
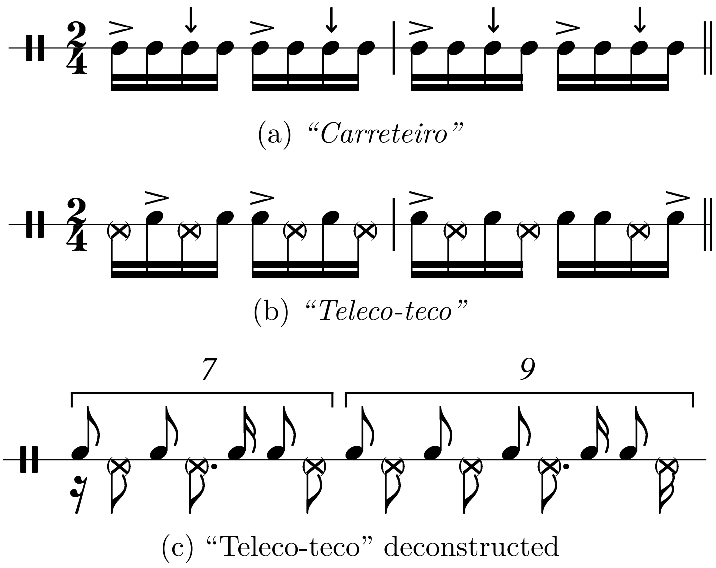


Figure 2.18: Cyclical parts for *surdo de terceira*. These instruments usually play the second beat of each measure like *surdos de primeira*, but are free to do fills and more complex rhythms, either improvised or in *convenções* unique to each *escola*. The symbol > indicates accented notes. Adapted from [60].



(a) *"Carreteiro"*

(b) *"Teleco-teco"*

(c) "Teleco-teco" deconstructed

Figure 2.19: Cyclical parts for *tamborim*. Two of the most common cycles are called: (a) *carreteiro* (where all the sixteenth notes are played) and (b) *teleco-teco*. The symbol ↓ indicates a *virada* (turn) of the instrument, which is then struck in an upwards movement. The (×) note heads (below the line) correspond to a note softly played with a finger of the hand holding the instrument, while notes on the line and above show when the instrument should be struck with the drumstick. Score (c) presents a deconstruction of the *teleco-teco* cycle, showing how to see the *tamborim* cycle of the *Estácio* paradigm. Adapted from [24, 60].

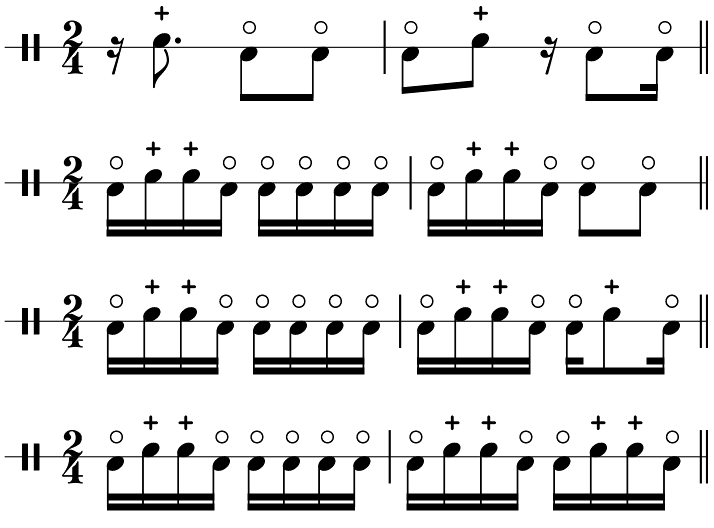
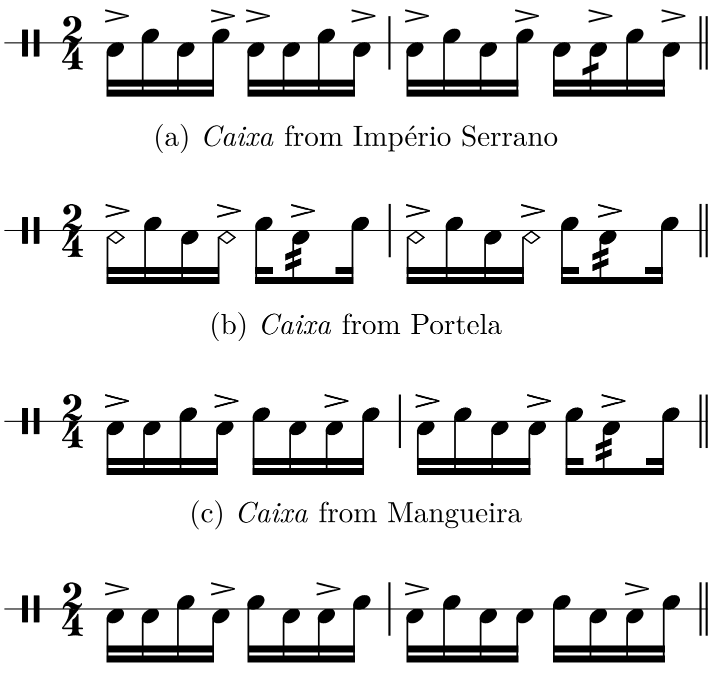


Figure 2.20: Cyclical parts for *cuíca*. The symbols + (note heads above the line) and ○ (note heads below the line) indicate, respectively, closed and open strokes, i.e., when fingers of the free hand press the skin from the outside while the other hand rub the stick (producing a higher tone pitch) or when the stick is rubbed without the pressing of the drumhead. Adapted from [60].



(a) *Caixa* from Império Serrano

(b) *Caixa* from Portela

(c) *Caixa* from Mangueira

(d) *Caixa* from Mangueira (older version)

Figure 2.21: Cyclical parts for *caixa* of different *escolas*. The ◇ note head is used for rim shots, while diagonal strokes across note stems indicate drum rolls. Notes below and above the line correspond to strong and weak hands, respectively. *Caixa* patterns show the identity of an *escola*. They originate from the religious drum playing associated to *orixás* (spirits in *candomblé*). Adapted from [60].

Chapter 3

Datasets

Machine-learning-based systems in music information retrieval (MIR) are becoming more complex to handle the increasing number of tasks and challenges they bring. As a result, accurately estimating the parameters of their typically large models requires a greater quantity and quality of data [61], especially because data must be often separated into training, test, and validation sets. Even though data augmentation techniques can be used to alleviate this bottleneck [61, 62], this kind of strategy is not able to solely solve the cultural bias still present in existing MIR data, methodologies, and conclusions [6].

Indeed, a great part of the research in this field focuses on musical traditions usually labeled as "Western". This is worrying, since by doing so we risk not being able to fully evaluate and reproduce specific musical properties found in many other cultures [6]. Some datasets attempt to be universal and to cover a large number of music styles, but end up sacrificing the very representation of what they are trying to portray. This is the case, for example, of the well-known Ballroom and Extended Ballroom datasets, in which the "Samba" class contains a mixture of songs of different origins.[1] Of those, only a few examples correspond to Brazilian rhythms, specifically identifiable as *bossa-nova*, *pagode*, and others [19]. In other datasets, music from non-"Western" traditions is given generic labels such as "Latin", or even "World" [19]. This underscores the importance of increasing the efforts towards the study of non-"Western" traditions found throughout this multicultural world.

Among the datasets devoted to non-"Western" music, one of the biggest projects today is CompMusic [6], which focuses on five particular music cultures: Arab-Andalusian, Beijing Opera, Turkish-makam, Hindustani, and Carnatic. Several

[1] There is a discussion to be had here regarding the origins and modern aspects of "ballroom samba". Without going into much detail, it is widely known that, despite its ties with early twentieth-century *maxixe*, modern ballroom samba (sometimes called international samba) is commonly associated with many other different "Latin" rhythms and is significantly different from the traditional *samba* styles still practiced in Brazil. It also differs, in both musical and dancing aspects, from the *samba de gafieira*, which is heir to *maxixe* in the national ballroom scene.

annotations are provided alongside the recordings, including melody (e.g., singer tonics, pitch contours), rhythm and structure (e.g., *tāḷa* cycles), scores (e.g., for percussion patterns), and lyrics. Some datasets of Latin-American music were also organized for MIR research. For instance, the dataset released in [63] comprises annotated audio recordings of Uruguayan *candombe* drumming, suited for beat/downbeat tracking. Aimed at music genre classification, the Latin Music Database [64] has Brazilian rhythms — *axé*, *forró*, *gaúcha*, *pagode*, and *sertaneja* — and music from other traditions: *bachata*, *bolero*, *merengue*, *salsa*, and *tango*. Focusing exclusively on Brazilian music, and built for music genre classification, the Brazilian Music Dataset [65] includes *forró*, rock, *repente*, MPB (*música popular brasileira*, lit. Brazilian popular music), *brega*, *sertanejo*, and disco.

For the tasks of drum transcription and, more specifically, drum playing technique recognition, the data that are available in the literature were mostly built around the standard drum kit (bass, snare, and tom-tom drums, hi-hats, and cymbals). We highlight the ENST-Drums dataset [66], which includes audio-visual recordings of three professional drummers playing individual strokes, phrases in different styles, soli, and accompaniment to real and to synthetic examples. Ten different labels were used in the annotations, encompassing brushing and stick techniques (e.g., rim shot, cross stick), among others. The MDB Drums dataset [67], which is comprised of drum kit annotations from a subset of the MedleyDB dataset [68], reports playing techniques for the snare drum, hi-hat, and cymbals. Synthetic data, which are present on a few datasets [62, 69], can also be used in this task; these ease the problem of generating reliable annotations and allow the construction of very large sets. Other datasets containing real-world playing techniques on the drum kit and also on Indian drums (*mridangam* and *tabla*) are mentioned in Chapter 6.

As we can see, even though several datasets of different rhythmic styles are available for use within the MIR community in its many tasks, *samba* is very poorly represented. In this chapter, we describe the two datasets exclusively of *samba* music that were organized and built for this research. We discuss the design criteria for compiling each dataset as well as the production of annotations and metadata. These datasets were originally reported in [19] and [20], the contents of which are here partially reproduced and also expanded to include other useful information. We also include in this chapter a brief description of two datasets that are also used in our experiments, but were assembled by other authors.

3.1 Brazilian Rhythmic Instruments Dataset

As mentioned in Chapter 1, the purpose of this work is the development of tools for the analysis and modeling of rhythm and rhythmic patterns using *samba* music as case study. The algorithms developed during the course of this thesis must thus incorporate prior musicological knowledge and provide musical insights, even though coming from a music technology perspective. To first test the effectiveness of such algorithms, a more controlled environment is preferred over the real occurrences of music phenomena, usually full of multilevel and interrelated information. The Brazilian Rhythmic Instruments Dataset (BRID), which we have selected for use in the initial steps of this project, complies with such requirement.

The BRID is a copyright-free dataset containing short solo- and multiple-instrument tracks in different Brazilian rhythmic styles, including *samba* and two of its subgenres. Currently, this dataset contains 367 short tracks of around 30 s on average, totaling 2 h 57 min (1.09 GB of data) at studio quality (44.1 kHz sample rate and 16-bit resolution). It was originally developed in the context of sound source separation [70], where the metronome-bound solo tracks were artificially mixed and served as ground truth for the separation process of the mixture track. Separation performance was also evaluated in the case of acoustic mixtures. However, the applicability of this dataset can most certainly be extended to other areas. In [19], we presented a few experiments showing how it can be used in rhythm computational analysis in particular.

Since this dataset does not contain singing voices, it is a good candidate for more controlled experiments, where the complex interactions between the interpreter and other melodic instruments with the percussive base may not be desirable. To the best of our knowledge, this is the first dataset of its kind, i.e., dedicated to Brazilian instruments and typical rhythms.

In the following sections, we briefly describe BRID and its contents. We also discuss how this dataset was organized and annotated.

3.1.1 Instruments and Rhythms

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*, *chocalho* (shaker), *cuíca*, *pandeiro*, *reco-reco*, *repique*, *surdo*, *tamborim* and *tantã*. To provide a variety of sounds, both membranophones and idiophones were featured. Also, whenever possible, instruments were varied in shape (e.g., oval or cylindrical shaker), size (e.g., 10- or 12-inch *pandeiro*), material (e.g., leather or synthetic drumhead), pitch/tuning (e.g., 1st, 2nd, and 3rd *surdos*, which are usually tuned in different pitch ranges for a *desfile de escola de samba*) and 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 (diameter of 12 in and either 4 or 6 snare wires), six *pandeiro* variations (either 10-, 11-, or 12-inch 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[2]). We refer the reader to Figure 2.16, where all instrument classes considered are portrayed. Tables 3.1 and 3.2 specify, respectively, each variation and the number of recordings each instrument is featured in.

Table 3.1: Instrument classes and variations (Var.).

Instrument	Label	Var.	Size (in)	Material	Drumstick	Additional info.
Agogô	AG	1	-	metal	wood	2 notes
Caixa	CX	1	12	nylon	wood	4 wires
		2	12	nylon	wood	6 wires
Chocalho	SK	1	-	wood	-	cylindrical
		2	-	metal	-	cylindrical
		3	-	plastic	-	double
Cuíca	CU	1	6	leather	-	-
		2	8	leather	-	-
		3	9.5	leather	-	-
Pandeiro	PD	1	10	nylon	-	-
		2	10	leather	-	-
		3	11	nylon	-	-
		4	11	leather	-	-
		5	12	nylon	-	-
		6	12	leather	-	-
Reco-reco	RR	1	10	metal	-	1 spring
		2	10	metal	-	2 springs
		3	11	metal	-	3 springs
		4	11	wood	-	-
Repique	RP	1	10	nylon	-	*repique de mão*
		2	12	nylon	wood	*repinique*
		3	12	leather	-	*repique de anel*
Surdo	SU	1	16	leather	mallet	-
		2	18	leather	mallet	-
		3	20	leather	mallet	-
Tamborim	TB	1	6	leather	wood	-
		2	6	nylon	wood	-
		3	6	nylon	plastic	-
Tantã	TT	1	10	leather	-	*tantã de corte*
		2	11	leather	-	*tantã de corte*
		3	12	leather	-	*tantã de corte*
		4	14	napa/nylon	-	*tantã de marcação*

[2] A leather-head *tamborim* is not played with a plastic single- or multiple-stemmed drumstick.

Table 3.2: Number of solo- and multi-instrument tracks per instrument class.

Instrument	Solo	Multi	Total
Agogô	12	14	26
Caixa	17	16	33
Chocalho	22	13	35
Cuíca	8	11	19
Pandeiro	85	44	129
Reco-reco	24	16	40
Repique	37	35	72
Surdo	26	42	68
Tamborim	20	24	44
Tantã	23	38	61

Table 3.3: Tempi/number of solo tracks per rhythm.

Rhythm	Label	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

The recordings present instruments being played in different Brazilian rhythmic styles. Although *samba* and two of its subgenres (*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 3.3 and 3.4, where tempo is given in beats per minute (bpm). All rhythms are in duple meter; *samba* and related genres are traditionally notated in this type of bar division, usually displaying a strong accent on the second beat [60]. During recording, only instruments and rhythms that are traditionally used in Brazilian music were considered to provide an authentic portrayal of each rhythm.

3.1.2 Dataset Recording

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

Table 3.4: Number of tracks per rhythm in multi-instrument recordings.

Rhythm	# Tracks
Samba	41
Partido-alto	28
Samba-enredo	21
Marcha	3

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

As mentioned before, 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 (each with years of experience in *samba* and other Brazilian genres) 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.[3] These are free improvisation patterns (still subject to the metronome track) which are very common in *rodas de samba*. It is worth mentioning that the musicians brought their own instruments for the recording sessions. Although the general characteristics of each instruments 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 ensemble performances, with the musicians playing together different rhythmic styles without a metronome reference, but with an indication of expected tempo. 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. One of the musicians featured in solo tracks was also recorded in these group settings. Table 3.5 summarizes musician participation in the recordings.

3.1.3 Track Labeling and Annotations

As previously informed, the recording of BRID was done for another work [70], and its intended use was in source separation tasks. Our specific contributions to the dataset were in its organization and cataloguing, together with the production of

[3] Except for *chocalho* (shaker) tracks.

Table 3.5: Musician participation in solo- and multi-instrument tracks.

Musician	Solo	Multi	Total
#1	91	-	91
#2	96	76	172
#3	87	-	87
#4	-	80	80
#5	-	78	78
#6	-	19	19

beats, downbeats, and onsets annotations, which allow it to be used in a wider range of problems. For the organization process, first the instrument classes and variations present in each track (solo recordings and acoustic mixtures in ensemble recordings) were recovered and a system of labels was developed for a naming convention. Missing tempo indications (in solo tracks) and musician identifications (specially in the case of multi-instrument tracks) were retrieved, as well as missing rhythmic style data. Track filenames were codified according to the procedure explained next.

Each audio track is given a unique filename, which starts with a four-digit number between brackets — a global identification number `[GID#]`, sequential for the entire dataset. In solo track (`S`) filenames, the `GID#` is followed by four groups of characters, whose format is either `SW-XXX-YY-ZZ` or `SW-XXX-YY-VZZ`, where `W` is the number for the musician playing in the track (see Table 3.5), `XXX` specifies the instrument class and variation being played, `YY` consists of a counter for tracks with the same pair musician–instrument, and `ZZ` (or `VZZ`) indicates the rhythmic style (or a *virada* for that style).

For acoustic mixture tracks (`M`), the `GID#` is followed by three groups of characters, whose format is `MW-YY-ZZ`. Here, `W` indicates the number of instruments recorded in the track (i.e., the number of musicians in the ensemble), `YY` is the counter for a given `MW` prefix, and `ZZ` means the same as in the case of solo tracks. The unique identifier (label) for each instrument class and for each rhythm can be found in Tables 3.1 and 3.3, respectively.

To exemplify, we can check two samples taken from the dataset: file `[0192] S2-PD3-01-SA`, which contains a solo *pandeiro* (variation 3: 11 in; leather head) recording, was executed by musician #2 in a *samba* style; and file `[0010] M4-10-SE`, which is a *samba-enredo* track performed by four musicians. A detailed list of instruments and musicians in each track is provided in Appendix A. In Table A.1, we can verify that in track `[0010]` the musicians were playing *caixa* (6 wires), *surdo* (18 in), *pandeiro* (11 in, nylon head), and *reco-reco* (metal, 2 springs).

Metronome tracks for solo recordings were made unavailable after the recording process and were not used in the case of ensemble recordings. Given that beat position is fundamental for some of the analyses conducted within the MIR domain, we carried on with the manual production of these targets. Downbeats were also annotated in this process. All time instants were procedurally aligned with the nearest onset — whenever the later was present —, defined as the local maximum of the spectral flux function [71] over a neighborhood of the annotated instant. Later, since another objective of this thesis is the study and characterization of rhythm patterns and microtiming in *samba* music, a precise and descriptive annotation of note articulations was required. Onsets were annotated only for solo tracks using the `madmom` Python package [72]. Then, note segments were clustered with respect to the stroke type in a related work [73], where the number of classes was perceptually determined for each file and instrument class. Then, using these clusters as starting points, we manually produced and refined stroke classifications.

All annotations were stored as plain text in .txt files. Each line in a file includes a timestamp and a label, which relates to an event in the corresponding recording. In the case of beat/downbeat annotations, the label is either "1" or "2", respectively if the beat is a downbeat or the second beat in a measure. In the case of articulation types, each label indicates a stroke class (e.g., labels "THUMB", "FINGERS", "SHELL" refer to three different *repique* strokes).

3.2 *Samba-Enredo* Dataset

A dataset of commercial *samba* recordings was also put together for this thesis. This dataset allows us to deal with the intricacies of music performances that more closely correlate to the actual musical phenomenon of *samba*, and provides a more representative portrayal of its particular rhythmic characteristics in a large collection of music data.

As stated in Chapter 2, *samba* has been developed in various forms throughout Brazil. Each one of the subgenres has its singularities in either rhythm, tempo, instrumentation, structure, improvisational aspects, etc. Even though they all share a common root, the amount of "swing" enforced by the musicians in each case can also differ greatly. Therefore, we deemed sensible to select several samples coming mostly of a single subgenre to form a cohesive dataset; this subgenre should preferably be one where the "swing" variance of its rhythmic properties was not so wide as to make the investigation impossible. Additionally, since this work was matured alongside the STAREL project, where a large dataset of *candombe* recordings was available, and *candombe* is performed in group processions (*comparsas*) during the Uruguayan Carnival, it would be interesting to also analyze a parading rhythm.

Samba-enredo satisfies all the above constraints, notably: it is tailored for and usually performed in *desfiles*; and it does not present too great a variation of playing speed in short time periods. Also, *sambas-enredo* are featured on a large number of commercial recordings in CD quality. In the next section we describe the *Samba-Enredo* Dataset (SAMBASET), currently comprised of recordings and metadata from 39 CDs of (mostly) *sambas-enredo*, which were acquired for the STAREL project. We also discuss in brief the process of annotation of beat/downbeat data.

3.2.1 Dataset Overview

As mentioned above, *sambas-enredo* are well documented in the phonographic industry. Apart from historical collections, since 1968 the yearly *sambas-enredo* that competing *escolas* will perform at the *Desfile* have been professionally recorded and marketed to the general public. Initially available as LP records, these official compilations began to appear regularly as CDs in 1990. Since then, the number of musicians (instrumentalists/choir) participating in each track has only increased.

SAMBASET covers different eras, from later renditions of old classics to the most recent *sambas-enredo* just out of the Sambadrome. Figure 3.1 indicates the distribution of *sambas* w.r.t. the year they were first performed (typically, the parading year). Three major collections make up the dataset; in chronological order:

- "*História das Escolas de Samba*" (HES): a collection of historical *sambas*, composed between 1928 and 1974, from four major *escolas de samba*,[4] arranged and interpreted by the instrumentalists of each *escola*. Recorded in 1974, published in four LPs by Discos Marcus Pereira (redistributed as CDs in 2011), the 48 tracks include a few *sambas-de-quadra/-de-terreiro* and *partidos-altos*.

- "*Escolas de Samba – Enredos*" (ESE): a collection of historical *sambas*, composed between 1949 and 1993, from ten traditional *escolas de samba*[5] in the voices of many idols from *samba*'s history, accompanied by a selected ensemble of instrumentalists and choir. There is a total of 100 tracks recorded and released as 10 CDs in 1993 by Sony Music, arranged by producer Rildo Hora. This collection includes a couple of tracks from different subgenres (*samba-de-terreiro* and *samba-exaltação*).

- "Sambas de Enredo" (SDE): official compilations of *sambas-enredo* recorded by members of the top *escolas*[6] from Rio de Janeiro, for each carnival parade between 1994 and 2018. The 25 CDs gather 338 tracks, published by RCA/BMG/Sony BMG (1994–2006) and by Universal Music (after 2007), with one

[4] Império Serrano, Mangueira, Portela, and Salgueiro.

[5] Beija-Flor, Estácio de Sá, Imperatriz, Império Serrano, Mangueira, Mocidade, Portela, Salgueiro, União da Ilha, and Vila Isabel.

[6] *Escolas* from the *Grupo Especial*. In 2018, these were 13 *escolas*, in order of their final score in the competition: Beija-Flor, Paraíso do Tuiuti, Salgueiro, Portela, Mangueira, Mocidade, Unidos da Tijuca, Imperatriz, Vila Isabel, União da Ilha, São Clemente, Grande Rio, and Império Serrano.

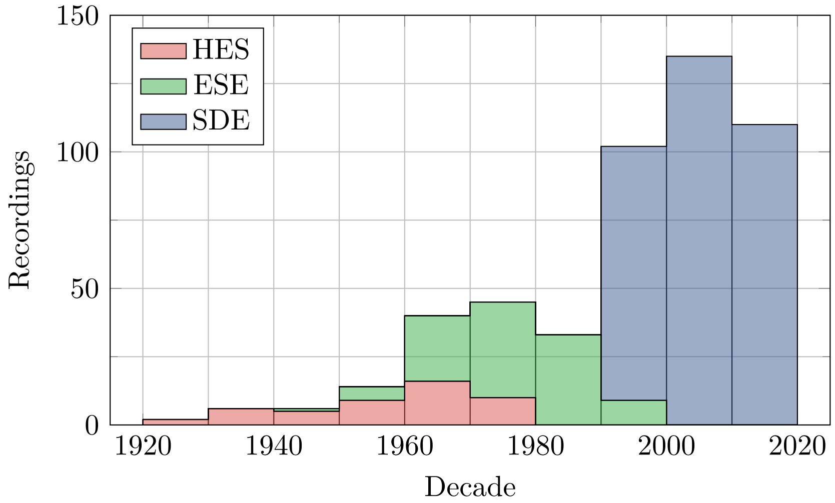


Figure 3.1: *Samba* recordings in SAMBASET per decade of first performance. Colors indicate the three different subcorpora in the dataset.

samba-enredo per track. In a few tracks, the first few seconds may include a short excerpt of a *samba-exaltação* or “battle cries” (calling the members of an *escola* to the *desfile*).

Table 3.6 shows the number of tracks for each *escola de samba* featured in the dataset, by genre. In total, there are 493 recorded *sambas* in 486 audio tracks,[7] resulting in over 40 h 30 min of content. All files are stereo with a sampling rate of 44.1 kHz and 16-bit resolution. Not only the three different collections allow for the coverage of different time periods, but they also have distinct sonorous characteristics. In HES, tracks feature only a few musicians playing very naturally and with great expression, as if they were in a *roda*. For several tracks in the official compilation (SDE), on the other hand, more than fifty instrumentalists play simultaneously while a choir of around the same size accompanies the main singer. For this reason, many tracks in SDE were recorded with the help of metronome tracks, although this information is not officially disclosed. Finally, ESE presents smaller ensembles and less expressiveness.

3.2.2 Metadata and Annotations

Metadata for albums and tracks were carefully curated and organized in an XML file. The information therein described was primarily obtained from CD booklets and later cross-checked with both the *União Brasileira de Composi-*

[7] Some tracks in the ESE collection contain more than one *samba*.

[8] An imbalance can be observed in both the distributions of genres and *escolas*. ST/SQ and OT tracks were only kept for the completeness of the dataset in regard to the CD collections. Moreover, the *escolas*’ playing styles are not so heterogeneous as to make this imbalance critical.

Table 3.6: Number of recordings in SAMBASET per *escola* and genre: *samba-enredo* (SE), *samba-de-terreiro/samba-de-quadra* (ST/SQ), and others (OT).[8]Other *escolas* were less featured in Grupo Especial from 1994 to 2018, presenting thus a smaller contribution to the SDE collection; these include: Império da Tijuca (2), Inocentes de Belford Roxo (1), Paraíso do Tuiuti (3), Renascer De Jacarepaguá (1), Rocinha (2), Santa Cruz (2), Unidos Da Ponte (3).

	Genres			
Escola	SE	ST/SQ	OT	Total
Mangueira	45	3	1	49
Portela	41	5	2	48
Salgueiro	42	5	-	47
Império Serrano	31	5	-	36
Mocidade	35	-	1	36
Beija-Flor	34	1	-	35
Imperatriz	35	-	-	35
Vila Isabel	33	-	-	33
União da Ilha	27	-	-	27
Grande Rio	25	-	-	25
Unidos da Tijuca	24	-	-	24
Viradouro	18	-	-	18
Estácio	16	-	-	16
Porto Da Pedra	15	-	-	15
Caprichosos	12	-	-	12
São Clemente	12	-	-	12
Tradição	11	-	-	11
Other *escolas* (7)	14	-	-	14
Total	470	19	4	**493**

tores[9](lit. Brazilian Union of Composers, UBC) and the *Instituto Memória Musical Brasileira*[10](Brazilian Musical Memory Institute, IMMuB). Whenever corresponding information was available, data were also checked against online database services such as FreeDB,[11] MusicBrainz[12] or Discogs.[13] Finally, we consulted *samba*-oriented forums and websites for additional, conflicting or missing information.

All XML tags can be seen in Figure 3.2, which shows an excerpt of the metadata file. While most of these labels are straightforward (e.g., title, composer, genre, samplerate), some require further clarification. First, the album_code refers to a unique code given to each album in the dataset. Albums from the HES and ESE collections were sequentially numbered, i.e., they are referred to by the codes `HES1` to `HES4` and `ESE1` to `ESE10`, respectively. For SDE, albums were specified via the

[9] `http://www.ubc.org.br/`

[10] `https://immub.org/`

[11] `https://gnudb.org`

[12] `https://musicbrainz.org`

[13] `https://www.discogs.com`

```
<metadata dataset="SAMBASET"
          curator="Lucas S. Maia"
          version="0.0.1">
...
  <album title="História das Escolas de Samba — Mangueira"
         arranger="Cartola"
         producer="J. C. Botezelli"
         instrumentalists="Various Artists"
         record_label="Discos Marcus Pereira"
         year_published="2011"
         length="00:29:48"
         total_tracks="12"
         album_code="HES1"
         barcode="7892141643634">
    ...
    <track track_number="6"
           title="Vale Do São Francisco"
           artist="Cartola"
           composer="Cartola and Carlos Cachaça"
           year_recorded="1974"
           year_first_performed="1948"
           genre="samba de enredo"
           length="02:49.226"
           samplerate="44100"
           bpm="78.4"
           start_time="00:07.895"
           end_time="02:49.226"
           checksum="d16974f135f0c374677c0e0db101cfea"/>
    ...
  </album>
...
</metadata>
```

Figure 3.2: Metadata file excerpt.

publishing year, which is also present in the album's title (i.e., SDE1994–SDE2018). The track_number is used with the album_code to name all audio files (e.g. the metadata in Figure 3.2 corresponds to file HES1.06). Track's start_time and end_time indicate the time each *samba* starts and ends, respectively. This is invaluable since many *samba-enredo* recordings are preceded by a short introductory speech or song motivating the performance, or succeeded by a "farewell" shout after the music has already stopped. The checksum attributes were filled with the MD5 hash of the track's WAV file, to allow the verification of audio data integrity. Finally, mean bpm values were estimated from the beat annotations described in the following.

SAMBASET has beat and downbeat annotations that were produced according to a semiautomatic procedure, after the results of a few experiments with state-of-the-art beat tracking systems (see Chapter 10). The DBNBeatTracker, available in the madmom package, was deemed a good candidate for providing reliable beat estimations [20]. Thus, first, automatically-generated beat annotations were obtained for all audio files using this system. In a second step, the estimates were checked and manually corrected, addressing eventual phase errors, and missing/extra beats. Since *samba-enredo* is always in duple meter, downbeats could be manually selected during this second phase. This two-step procedure greatly reduced the amount of manual work necessary to annotate beats and downbeats for this entire dataset.

A simple interface was built in Python 2 with Tkinter for a visualization of the dataset's albums and associated information. This interface allows the user to select and play any track and also search the dataset. A screenshot of the interface can be seen in Figure 3.3. A similar visualization, devoid of audio files for copyright purposes, is available at `http://www.smt.ufrj.br/~starel/sambaset` and shown in Figure 3.4.

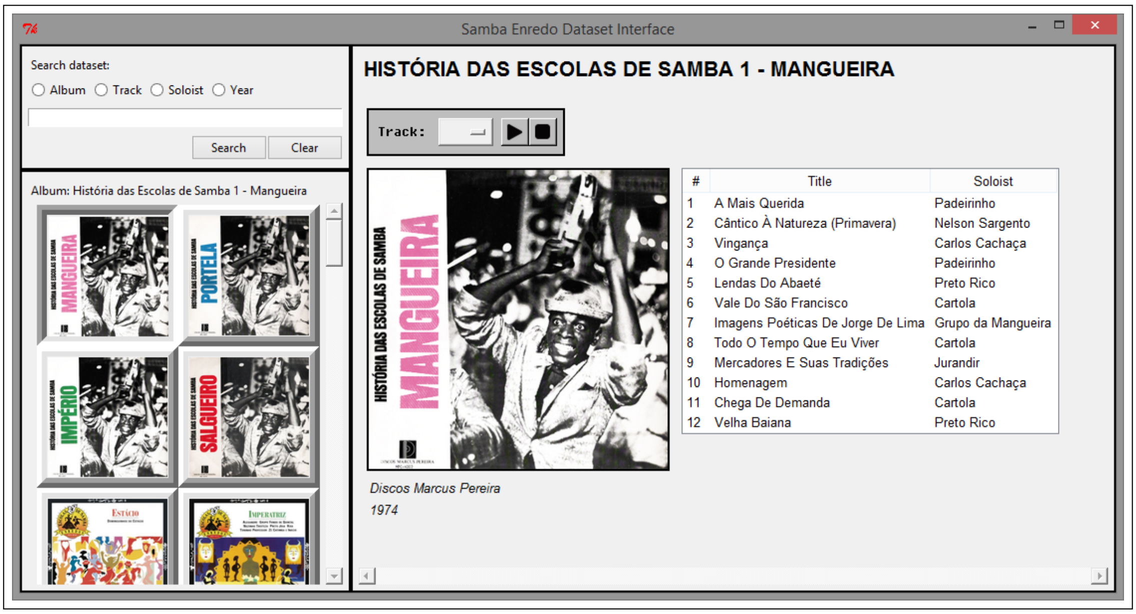
3.3 Other Datasets

Throughout our experiments in Part III, we have used other datasets that, although not prepared by us, provide interesting comparison points for the methodologies we have developed.

First, there is the Candombe dataset [63, 74]. *Candombe* refers to one of the most essential parts of Uruguayan popular culture. It is a style of dance and drumming music that can be traced back to the cultural practices brought to the Americas by enslaved African populations. Three types of drums are featured in *candombe*, each corresponding to a different frequency range and specific rhythmic patterns. *Chico* is the smallest (and highest-pitched) drum and functions as a timekeeper, describing the smallest metrical pulse. The *repique* is responsible for improvisational parts in the mid register. Finally, the *piano*, a large bass drum, plays the accompaniment. A timeline pattern, *clave* or *madera*, is shared by the three drums and is produced by hitting the drum shell with a stick. This pattern is commonly played by all drums at the start of a performance and helps establish the four-beat cycle, which is irregularly divided [38]. As with many musics of African tradition, *candombe* contains strong phenomenological accents that are displaced with respect to the metric structure [38]. The Candombe dataset contains 35 recordings of *candombe* drumming (2.5 h in duration) featuring three to five drummers performing different configurations of the three differently-sized drums. We note that tempo varies greatly and often increases along the performances in this dataset.

Furthermore, we have also explored the Ballroom dataset [75, 76], which is recognized as a standard of MIR literature. It consists of many distinct genres and was selected to serve as a counterpoint in our investigation in Chapter 11. It includes 698 tracks of eight ballroom dance genres (cha-cha-cha, jive, quickstep, rumba, samba, tango, Viennese waltz, waltz) of 31 s on average.

To allow a direct comparison of results across all datasets, *candombe* recordings


Samba Enredo Dataset Interface
Search dataset:
○ Album ○ Track ○ Soloist ○ Year
Search
Clear
Track:

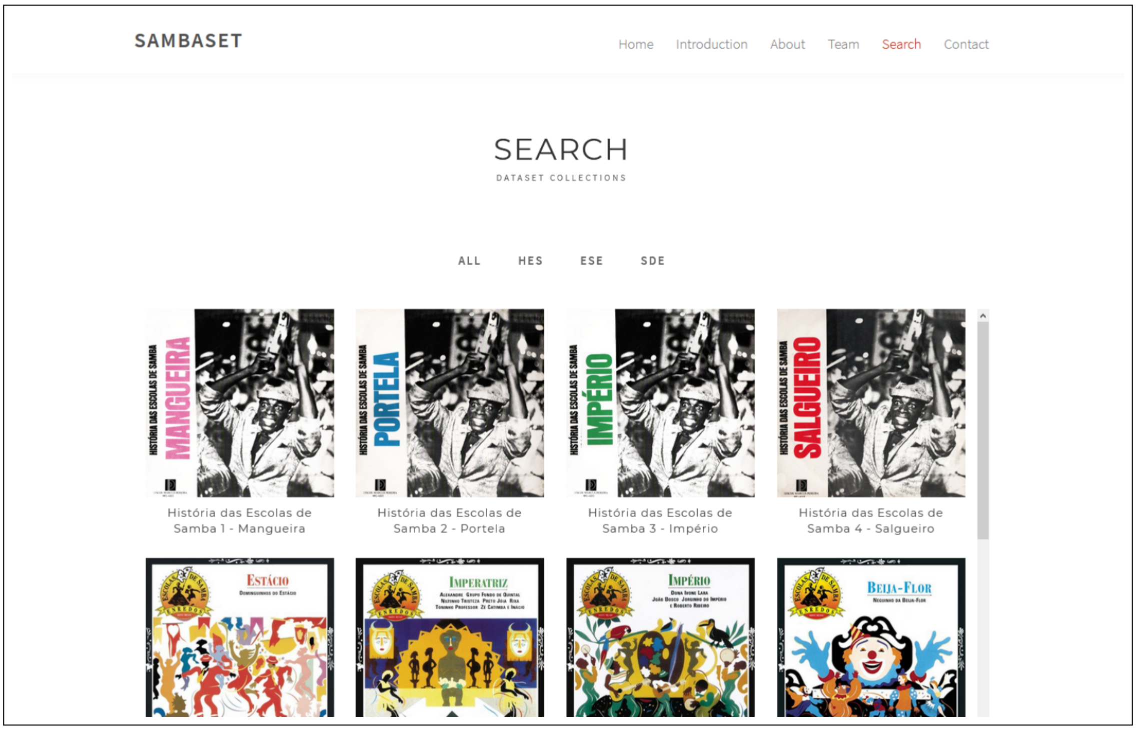
Album: História das Escolas de Samba 1 - Mangueira

HISTÓRIA DAS ESCOLAS DE SAMBA 1 - MANGUEIRA

#	Title	Soloist
1	A Mais Querida	Padeirinho
2	Cântico À Natureza (Primavera)	Nelson Sargento
3	Vingança	Carlos Cachaça
4	O Grande Presidente	Padeirinho
5	Lendas Do Abaeté	Preto Rico
6	Vale Do São Francisco	Cartola
7	Imagens Poéticas De Jorge De Lima	Grupo da Mangueira
8	Todo O Tempo Que Eu Viver	Cartola
9	Mercadores E Suas Tradições	Jurandir
10	Homenagem	Carlos Cachaça
11	Chega De Demanda	Cartola
12	Velha Baiana	Preto Rico

Discos Marcus Pereira
1974

Figure 3.3: Interface for querying the *Samba-Enredo* Dataset (SAMBASET).



História das Escolas de Samba 1 - Mangueira

História das Escolas de Samba 2 - Portela

História das Escolas de Samba 3 - Império

História das Escolas de Samba 4 - Salgueiro

Figure 3.4: Web version of SAMBASET interface.

have been segmented into 276 non-overlapping 30-second excerpts. The reasons for this are elucidated in Chapter 11.

Chapter 4

Signal Transforms

In this chapter, we briefly present the main signal transforms that are used throughout this thesis, while at the same time introducing our mathematical notation. Audio signals can be represented in many domains, each one being more useful to represent and analyze a certain kind of information. Signal transforms that translate among these different domains are of paramount importance in signal processing, particularly in audio processing.

4.1 Fourier Transform

Sounds are produced by the vibration of an object (e.g., a string, a membrane) and propagated by perturbations of a medium. In the case of a medium such as air, these perturbations are a series of local rarefactions and compressions that radiate from the sound source. At a given point in space, this sound wave is defined by changes in the local pressure, and can be represented by its waveform, $x(t)$, which is a function of time t (usually in seconds).

The Fourier transform (FT) is a mathematical tool that allows us to analyze the frequency content of $x(t)$, by comparing it to an infinite number of pure sinusoids of different frequencies. The Fourier transform of $x(t)$ is defined as [77]

$$X(\mathrm{j}\omega) = \int_{-\infty}^{\infty} x(t)\mathrm{e}^{-\mathrm{j}\omega t}\,\mathrm{d}t, \tag{4.1}$$

where j is the imaginary unit and ω, a continuous variable in radians per second (rad/s), is the angular frequency of each sinusoid

$$\mathrm{e}^{\mathrm{j}\omega t} = \cos\omega t + \mathrm{j}\sin\omega t. \tag{4.2}$$

We can interpret $X(\mathrm{j}\omega)$ in Equation (4.1) as a complex number that describes the magnitude and phase of its respective complex sinusoid in the superposition that represents $x(t)$, such that an inverse transform can be obtained by [77]

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\mathrm{j}\omega)\mathrm{e}^{\mathrm{j}\omega t}\,\mathrm{d}\omega. \tag{4.3}$$

One of the many properties of the Fourier transform is the convolution theorem, which states that the transform of the linear convolution of two signals in the time domain corresponds to multiplication in the frequency domain, and vice versa:

$$(x * y)(t) \quad \overset{\mathrm{FT}}{\longleftrightarrow} \quad X(\mathrm{j}\omega)\cdot Y(\mathrm{j}\omega) \tag{4.4}$$

$$x(t)\cdot y(t) \quad \overset{\mathrm{FT}}{\longleftrightarrow} \quad (X * Y)(\mathrm{j}\omega). \tag{4.5}$$

When an analog signal like $x(t)$ is brought to the digital world (e.g., digital recordings) it undergoes discretization and quantization, which allow it to be stored in binary form. The discrete-time version $x[n]$ is obtained from $x(t)$ through the sampling process, which consists of retaining one sample of the signal at every T_s seconds, where $T_\mathrm{s} > 0$ is called the sampling interval. In other words,

$$x[n] = x(nT_\mathrm{s}), \text{ for } n \in \mathbb{Z}. \tag{4.6}$$

Similarly to its continuous-time counterpart, the discrete-time signal $x[n]$ may also admit a Fourier representation. First, let us define a continuous-time signal $x_{T_\mathrm{s}}(t)$ that is equivalent to $x[n]$ as

$$x_{T_\mathrm{s}}(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT_\mathrm{s}), \tag{4.7}$$

with $n \in \mathbb{Z}$ and where $\delta(t)$ is the unit impulse at $t = 0$. We can observe that $x_{T_\mathrm{s}}(t)$ is defined by the product of $x(t)$ by an impulse train of period T_s, i.e., each unit impulse is shifted to $t = nT_\mathrm{s}$ and weighted by the value of $x(t)$ at the same instant. By applying the convolution theorem, we can readily see that the Fourier transform $X_{T_\mathrm{s}}(\mathrm{j}\omega)$ of our equivalent signal is the result of the convolution between $X(\mathrm{j}\omega)$ and the Fourier transform of the impulse train, which is itself an impulse train of period $\omega_\mathrm{s} = 2\pi/T_\mathrm{s}$. This means that $X_{T_\mathrm{s}}(\mathrm{j}\omega)$ is a periodically repeated version of $X(\mathrm{j}\omega)$. The Fourier representation of $x[n]$ is the counterpart to this $X_{T_\mathrm{s}}(\mathrm{j}\omega)$. We define this discrete-time Fourier transform (DTFT) as [77]

$$X(\mathrm{e}^{\mathrm{j}\Omega}) = \sum_{n=-\infty}^{\infty} x[n]\mathrm{e}^{-\mathrm{j}\Omega n}, \tag{4.8}$$

where $\Omega = \omega T_\mathrm{s}$ is the normalized angular frequency in radians per sample, so that $X(\mathrm{e}^{\mathrm{j}\Omega})$ is periodic in Ω with period 2π rad/sample.

The choice of T_s in the sampling process is not arbitrary. In fact, the Nyquist–Shannon theorem states that f_s, the reciprocal of the sampling period, must be greater than twice the maximum frequency present in the original signal, $x(t)$, if $x(t)$ is a real-valued baseband signal. If $x(t)$ is not bandlimited, it should first be made so by a low-pass analog filter; this guarantees that $x(t)$ can be recovered perfectly from $x[n]$. We can again use the equivalent signal $x_{T_\mathrm{s}}(t)$ to see why that is the case. First note that if f_s is much greater than the signal bandwidth, one can recover $X(\mathrm{j}\omega)$ by low-pass filtering $X_{T_\mathrm{s}}(\mathrm{j}\omega)$ with an adequate cutoff frequency. However, if T_s is too large and f_s is small when compared to the signal bandwidth, an overlap can occur between the shifted replicas of $X(\mathrm{j}\omega)$ in $X_{T_\mathrm{s}}(\mathrm{j}\omega)$. This phenomenon, called aliasing, potentially occurs when f_s is not greater than twice the bandwidth of $X(\mathrm{j}\omega)$.

We have not yet imposed a limitation on the length of signal $x[n]$, which cannot be infinite for digital applications. Not only that, but the continuous-frequency representation provided by the DTFT is also unfeasible in the same settings. If we limit the length of $x[n]$ to N samples, we can compute its discrete Fourier transform (DFT), which can be shown to be equivalent to a uniformly sampled version of the DTFT of the same signal at frequency values $\Omega_k = \Omega_0 k$, where $\Omega_0 = 2\pi/M$ [5]. The parameter M describes the frequency resolution of the DFT, and usually is set $M = N$, such that the DFT may be regarded as a change of basis. This way, the transform provides a representation $X[k]$ of same size as the input $x[n]$, i.e., N samples, and is defined by [5]

$$X[k] = \sum_{n=0}^{N-1} x[n]\mathrm{e}^{-\mathrm{j}\Omega_k n}, \tag{4.9}$$

with $k = 0, 1, \ldots, N-1$. Similarly to what happened in the sampling process of time signals, the discretization of the DTFT in the frequency domain at every $2\pi/N$ radians creates time replicas of $x[n]$ with a repetition period of N. The convolution theorem still holds for the DFT in the case of circular convolutions as signals are periodic in both time and frequency domains:

$$(x \circledast y)[n] \quad \overset{\mathrm{DFT}}{\longleftrightarrow} \quad X[k] \cdot Y[k] \tag{4.10}$$

$$x[n] \cdot y[n] \quad \overset{\mathrm{DFT}}{\longleftrightarrow} \quad (X \circledast Y)[k]. \tag{4.11}$$

It should be noted that for finite-length signals $x[n]$ and $y[n]$ the circular convolution and the linear convolution are identical if the signals are sufficiently zero-padded.

The family of Fourier representations translate a continuous- or discrete-time signal to the frequency domain, displaying its frequency components. As a result, we have no information on when these frequency components appear in the original signal, which could be desirable especially in the case of quasi-stationary or even nonstationary signals. A short-time analysis of the Fourier transform has been devised with this problem in mind. In this representation, known as short-time Fourier transform (STFT), a finite time window is used to select a portion of the signal; the Fourier analysis of the signal portion is computed and the window advances to the next time position. For a discrete-time signal $x[n]$, this procedure is mathematically described by [5]

$$X(\mathrm{e}^{\mathrm{j}\Omega}, m) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]\mathrm{e}^{-\mathrm{j}\Omega n}, \tag{4.12}$$

where $w[n]$ is a window function which is usually centered around $n = 0$. At each step, the window selects a portion of the signal around $n = m$ for analysis. We can, of course, compute a discretized version of the STFT by means of the DFT. It is also possible to advance the window by larger steps, with the inclusion of a hop parameter. The discrete-frequency version of the STFT of signal $x[n]$ is [5]

$$X[k, m] = \sum_{n=0}^{N-1} x[n]w[n-mh]\mathrm{e}^{-\mathrm{j}\Omega_k n}, \tag{4.13}$$

where $w[n]$ has length N and $h \in \mathbb{N}^*$ is the hop size. Thus, for a given m, $X[k, m]$ can be interpreted as the discrete Fourier transform of the modified signal defined by $x[n]w[n - mh]$. In general, the STFT is visually represented in two-dimensional form by its magnitude or power spectrograms,

$$Y_{\mathrm{mag}}[k, m] = |X[k, m]| \quad \text{or} \quad Y_{\mathrm{pow}}[k, m] = |X[k, m]|^2. \tag{4.14}$$

The STFT admits another interpretation [78]: for a fixed frequency Ω_k, $X[k, m]$ is the linear convolution between the window $w[n]$ with a modulated version of $x[n]$, i.e., $x[n]\mathrm{e}^{-\mathrm{j}\Omega_k n}$; the result of which is then decimated by h. The modulation shifts to the baseband all the frequency content of $x[n]$ around $\Omega = \Omega_k$, while the window $w[n]$, which normally has the form of a low-pass filter, selects only a small portion of the spectrum of $x[n]$ around $\Omega = 0$. From this filtered time signal, we then take one sample at every h samples. This filter-bank interpretation of the STFT will be quite useful in the following sections.

4.2 Constant-Q Transform

As seen before, the resolution provided by the DFT is uniform in frequency and the bins in this domain are linearly spaced with $\Delta\Omega_k = 2\pi/N$ rad/sample, where N is the length of the signal. Particularly, in the case of the STFT, where N is the length of the analysis window, a tradeoff arises between time and frequency resolutions. If we increase N, although the frequency resolution will be improved, we lose the ability to precisely detect events in time. Instead, if we decrease the number of analysis points, time localization is improved whereas frequency content loses resolution. This is known as the uncertainty principle, and it is well described and understood in the literature [5, 77].

The fixed time-frequency resolution the STFT has over the entire time-frequency plane is undesirable for certain applications. In the case of music signals and in auditory models of human hearing, the geometric relations among harmonics are such that the DFT has too fine a resolution in high frequencies, and low frequencies are underrepresented. Different time-frequency representations were developed to deal with this issue. For example, some representations circumvent this issue by averaging the energy in adjacent STFT frequency bins, grouped in a non-uniform fashion (e.g., mel spectrogram). Others, like the constant-Q transform (CQT) and its short-time computation, can be obtained from the time signal by processing it with a different family of filters than those used in the DFT.

The constant-Q transform [79] has geometric spacing and resolution in frequency which is controlled by the quality factor Q. Unlike in the uniform sampling of the DFT, we define a series of bin center frequencies

$$\Omega_k = \Omega_0 2^{\frac{k}{B}}, \tag{4.15}$$

for $k = 0, 1, \ldots, K-1$, where B is the number of frequency bins per octave (e.g., $B = 12$ for semitone resolution) and K is the total number of frequency bins used to compute the representation. The factor Q defines the constant selectivity of each filter as

$$Q = \frac{\Omega_k}{\Delta\Omega_k} = \frac{1}{2^{\frac{1}{B}} - 1}, \tag{4.16}$$

such that $\Delta\Omega_k$, the bandwidth for each bin, is proportional to the center frequency Ω_k. Remember that in the DFT the digital bandwidth of each bin is given by $2\pi/N$. In the CQT, the desired constant selectivity can be achieved by changing the length of the analysis window for each bin k, i.e.,

$$N_k = \frac{2\pi}{\Delta\Omega_k} = \frac{2\pi Q}{\Omega_k}. \tag{4.17}$$

Note that window lengths are real-valued. In practical implementations, N_k is usually rounded towards zero [79] or the nearest odd integer [80].

With that, the k-th spectral component of this representation is computed as

$$X_{\mathrm{CQ}}[k] = \frac{1}{N_k} \sum_{n=0}^{N_k-1} x[n] w_k[n] \mathrm{e}^{-\mathrm{j}\Omega_k n}, \tag{4.18}$$

where the window $w_k[n]$ is zero outside the range $[0, N_k - 1]$ and the center frequency is given by $\Omega_k = 2\pi Q/N_k$, which follows directly from Equation (4.17). The normalization factor N_k is included in this equation since each spectral component requires the summation of a different number of terms.

Finally, as in the case of the STFT, we can compute a short-time version of the CQT as

$$X_{\mathrm{CQ}}[k, m] = \frac{1}{N_k} \sum_{n=0}^{N_k - 1} x[n] w_k[n - m h_k] \mathrm{e}^{-\mathrm{j}\Omega_k n}, \tag{4.19}$$

where $X_{\mathrm{CQ}}[k, m]$ is the bin corresponding to the k-th frequency and m-th frame of $x[n]$ and h_k is the hop length (in samples) for a given channel k. In most applications, this hop is set $h_k \triangleq h$, i.e., it is made independent of the channel.

As we can observe, the CQT exploits the time-frequency resolution tradeoff by analyzing low-frequency components with longer observation windows $w_k[n]$ than in the high-frequency ones, which results in an improved frequency resolution in the low-frequency range and higher time resolution at the other end of the spectrum.

The CQT is more computationally expensive than the DFT, for which the fast Fourier transform (FFT) algorithm provides an efficient and elegant solution. However, efficient implementations of the CQT have been proposed in [80, 81].

4.3 Discrete Wavelet Transform

The discrete wavelet transform (DWT) is another interesting representation, which performs a multiresolution analysis via the projection of a time signal on a family of basis functions called wavelets [82]. These functions are defined by two parameters, k and m, that control, respectively, the time-scale dilation (scale) and the translation of a prototypical "mother" function, the analysis wavelet $\psi(t)$. The projection of $x(t)$ on the wavelets allows for an analysis of its contents with a variable degree of detail and is suitable for describing transient and nonstationary signals.

Let us first define $X_\Psi[k, m]$ that represents $x(t)$ at the k-th scale and displacement m as [82]

$$X_\Psi[k, m] = \int_{-\infty}^{\infty} x(t) \psi_{k,m}^*(t) \, \mathrm{d}t, \tag{4.20}$$

with $k, m \in \mathbb{Z}$ and where * denotes the complex conjugate. The "daughter" wavelets, $\psi_{k,m}(t)$, are usually obtained from the analysis wavelet through a dyadic derivation[1] such that

$$\psi_{k,m}(t) = 2^{\frac{k}{2}} \psi(2^k t - m), \tag{4.21}$$

[1] The dyadic derivation of the "daughter" functions in the wavelet transforms is sometimes expressed with negative powers, i.e., $2^{-k/2}$ and 2^{-k} (cf. [83]). In this case, the interpretation of the scale parameter and its effects are the opposite of what is discussed in this section, e.g., basis functions are stretched when the scale is increased, etc.

where $2^{k/2}$ is an amplitude scale factor that guarantees normalization along different scales. Many families of compactly supported bandpass functions have been proposed in the literature, e.g., the Haar, the Daubechies, and the Morlet wavelets [82].

From Equation (4.21), we can readily verify that, as the time-scale dilation parameter k grows, the basis function is compressed in time; correspondingly, it is stretched in the frequency domain and its center frequency shifts upwards. Conversely, the larger the scale, the more dilated in time the wavelet becomes, presenting a thinner support in the frequency domain. The geometric time scaling enforced by the dyadic derivation ensures that, as the scale increases (as do the central frequency and bandwidth of the basis function), fewer filters are arranged to cover the corresponding frequency region. Also note that the step for the translation of each "daughter" wavelet is dependent on the scale.

The original signal $x(t)$ can be resynthesized through [82]

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} X_{\Psi}[k,m]\overline{\psi}_{k,m}(t), \tag{4.22}$$

where $\overline{\psi}_{k,m}(t)$ are dilations and translations of a synthesis "mother" wavelet, $\overline{\psi}(t)$, that also obey the relation expressed in Equation (4.21). In practice, the signal $x(t)$ is limited in support, and we do not need to consider an infinite number of translations m when computing the inverse of Equation (4.22).

It is possible to reduce the number of coefficients that are used to represent $x(t)$. For this purpose, we introduce another set of functions, the analysis and the synthesis scaling functions $\varphi(t)$ and $\overline{\varphi}(t)$, and their corresponding families (subject to a derivation in the form of Equation (4.21)). With the analysis scaling function, we can compute the coefficients [82]

$$X_{\Phi}[k,m] = \int_{-\infty}^{\infty} x(t)\varphi^{*}_{k,m}(t)\,\mathrm{d}t. \tag{4.23}$$

The multiresolution principle states that a space that contains signals with high degree of detail will also contain those of lower resolution [82]. In other words, if $\mathbf{\Phi}_k$ is the span (i.e., the set of signals that can be represented by a basis set) at scale k, then $\mathbf{\Phi}_k \subset \mathbf{\Phi}_{k+1}$. The scaling functions and the wavelets are related in such way that, for a given scale k, the subspace $\mathbf{\Psi}_k$ in which the wavelets reside is orthogonal to $\mathbf{\Phi}_k$, and complements the latter to form $\mathbf{\Phi}_{k+1}$. We can then choose an initial scale k_0, that indicates the coarsest span generated by the $\varphi_{k_0,m}(t)$ functions. Lastly, by taking into consideration that $x(t)$ is limited in resolution, we may also impose an upper bound on k, for example $k \leq K$, up to which the wavelet family will provide the necessary detail to the analysis at hand. Figure 4.1 depicts the nesting of the subspaces generated by the scaling and wavelet functions.

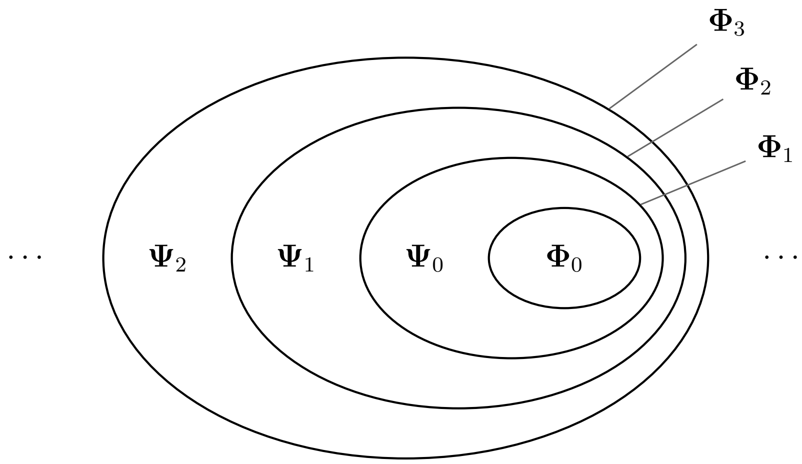


Figure 4.1: Vector spaces for scaling and wavelet functions starting at $k_0 = 0$.

Thus, without loss of generality, we set the initial scale $k_0 = 0$ and K as the maximum scale. Then, defining $\varphi_m(t) = \varphi_{k_0,m}(t)$ and $X_\Phi[m] = X_\Phi[k_0, m]$, we can alternatively reconstruct the time signal $x(t)$ as (cf. Equation (4.22))

$$x(t) = \sum_{m=-\infty}^{\infty} X_\Phi[m]\overline{\varphi}_m(t) + \sum_{m=-\infty}^{\infty} \sum_{k=0}^{K} X_\Psi[k, m]\overline{\psi}_{k,m}(t). \tag{4.24}$$

Since lower resolution scaling functions must reside in the space defined by higher resolution ones, we can express any scaling function in $\mathbf{\Phi}_k$ with the functions that define $\mathbf{\Phi}_{k+1}$ through the following relation [82]:

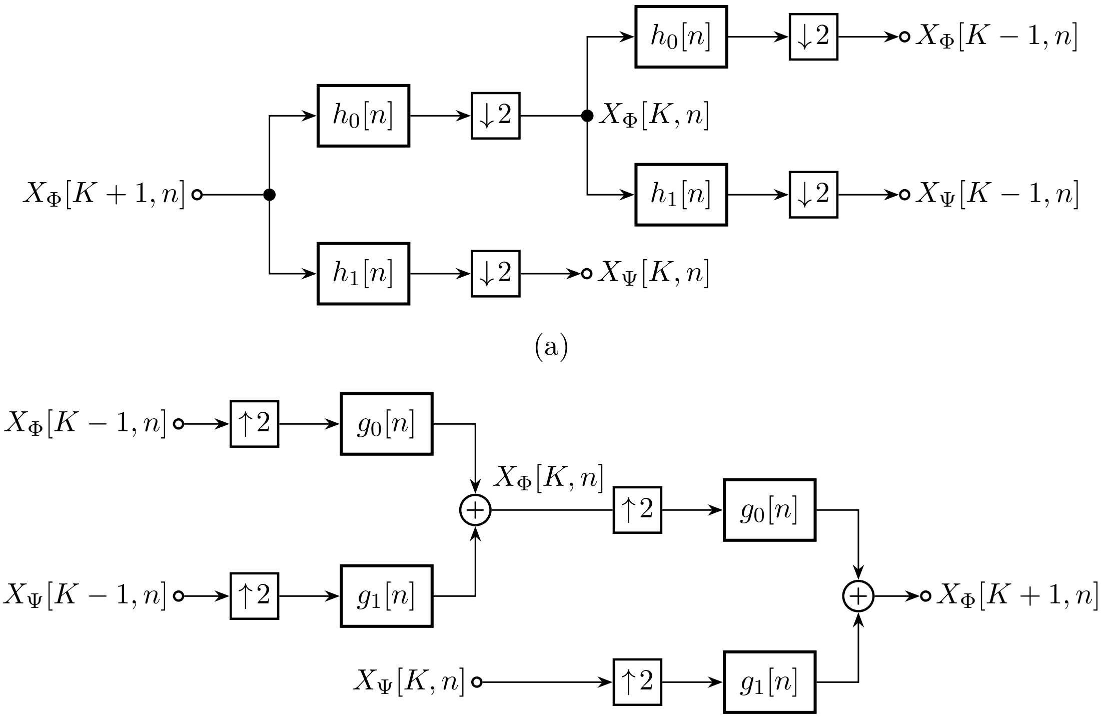
$$\varphi(2^k t - m) = \sum_n h_0[-n]\sqrt{2}\varphi(2^{k+1}t - 2m - n), \tag{4.25}$$

with $n \in \mathbb{Z}$. Similarly, because the wavelets at the k-th scale reside in $\mathbf{\Psi}_k$, which is the orthogonal complement of $\mathbf{\Phi}_k$ in $\mathbf{\Phi}_{k+1}$, they can be expressed by functions in this space as the weighted sum

$$\psi(2^k t - m) = \sum_n h_1[-n]\sqrt{2}\varphi(2^{k+1}t - 2m - n), \tag{4.26}$$

with $n \in \mathbb{Z}$. The synthesis scaling and wavelet functions share analogous relations, with weights correspondingly given by $g_0[-n]$ and $g_1[-n]$ [82].

From the multiresolution formulations given by Equations (4.25) and (4.26), it can be shown [82] that the coefficients $X_\Psi[k, m]$ and $X_\Phi[k, m]$ are expressible in terms of $X_\Phi[k + 1, m]$, the scaling coefficients at the $(k + 1)$-th scale. This allows a quite efficient computation of the wavelet and scaling coefficients at coarser scales via a critically decimated dyadic filter bank of analysis, the input of which are the values $X_\Phi[k + 1, n]$ which are filtered by two finite impulse response (FIR) filters,



(a)

(b)

Figure 4.2: Filter banks of (a) analysis and (b) synthesis for the efficient computation of wavelet transform coefficients.

with impulse responses $h_0[n]$ and $h_1[n]$. Similarly, we can travel from coarser scales up to scales of more detail with a synthesis filter bank determined by $g_0[n]$ and $g_1[n]$.

Strictly speaking, the wavelet transform is only defined for continuous time signals, $x(t)$. In practice, however, we refer to the wavelet transform of a discrete time signal $x[n]$ from scale 0 up to K as the coefficients obtained when filtering this signal through the analysis filter bank described before [83]. Note that this is equivalent to computing the wavelet decomposition of a continuous time signal $\hat{x}(t)$ that, at the $(K+1)$-th scale, is represented by coefficients $X_\Phi[K+1,m] = x[m]$.

Wavelet transform of a time signal can be visually represented by a scalogram, a time-scale plot of its coefficients. For complex wavelets, moduli of the coefficients or squared moduli are used, analogously to usual spectrogram representations.

4.4 Modulation Spectral Transform

The time-frequency representations we reviewed earlier manage the tradeoff between two different domains. They are quite useful in the representation of nonstationary signals, where they resort to an appropriate set of window lengths and hop sizes such that signal properties remain approximately stationary within every analysis frame. A different way to model and study this class of signals is by means of modulation analyses. In this framework, a real broadband signal $x[n]$ is interpreted as a superposition of different narrowband components, and each of these subbands $s_k[n]$ (e.g., analytic subbands of a filter bank) is itself expressed in product form by a modulating signal and a carrier, $m_k[n]$ and $c_k[n]$ respectively [84]:

$$x[n] = \sum_{k=0}^{K-1} s_k[n] = \sum_{k=0}^{K-1} m_k[n] c_k[n]. \tag{4.27}$$

Here, the modulating term represents the envelope of the subband signal (amplitude modulation) while the carrier contains its temporal fine structure (frequency modulation). There is no unique way to demodulate a single subband $s_k[n]$, i.e., determine $m_k[n]$ and $c_k[n]$ for all n; in fact, for many nontrivial cases, $s[n] = m[n]c[n]$ defines an underdetermined system of equations. The mathematical possibilities for demodulation, albeit legitimate, are not all useful and can be reduced in number by the use of certain conventions or constrained by the physical meaning of the signals under analysis [85, 86], e.g., boundedness of the amplitude modulator or the bandwidth of the carrier's instantaneous frequency. We describe in the following two different approaches to demodulate the subband signals.

4.4.1 Incoherent Demodulation

In the incoherent demodulation, we assume modulators are real-valued and non-negative, and separate the complex analytic subbands into magnitude and phase. We then have

$$m_k[n] = |s_k[n]| \quad \text{and} \quad c_k[n] = \mathrm{e}^{\mathrm{j}\angle s_k[n]}, \tag{4.28}$$

respectively the Hilbert envelope and the complex carrier term of the k-th subband.

The incoherent demodulation gets its name from the fact that the modulator is estimated directly from the subband signal, without previous estimation of the carrier. This particularly common demodulation method does not guarantee a bounded modulator nor a band-limited carrier. Indeed, since all the phase information is represented by the carrier, its instantaneous frequency can exceed the frequency range of the original signal or even contain infinite discontinuities [85].

4.4.2 Coherent Demodulation

In the coherent demodulation, we must first recover the carrier signal $c_k[n] = \mathrm{e}^{\mathrm{j}\phi_k[n]}$, which can then be used for estimating the modulator from the subband signal by

$$m_k[n] = s_k[n] c_k^*[n]. \tag{4.29}$$

There are different ways to accomplish this carrier estimation. One example is the spectral center-of-gravity (COG) method [85], in which the instantaneous frequency (IF) of each analytic subband is framewise estimated via the spectral centroid (see Section 5.2.2). In practice, the squared magnitude of the STFT of $s_k[n]$ is used in the computation of the spectral density at each frame. By virtue of this frame-by-frame process, this IF estimate is smooth, which results in a bandlimited carrier.

The carrier phase $\phi_k[n]$ is computed through discrete integration of the instantaneous frequency $\Omega_{\mathrm{IF}}^k[n]$, such that

$$\phi_k[n] = \sum_{p=0}^{n} \Omega_{\mathrm{IF}}^k[p]. \tag{4.30}$$

The carrier signal is now determined and the corresponding (now possibly complex) modulator can be retrieved by Equation (4.29). It follows from this demodulation scheme that, if the original signal is bounded (in frequency), the modulator will be well-behaved [85].

In either the incoherent or coherent demodulation cases, the modulation spectrum can be obtained through a suitable time-frequency transformation (e.g., FT, STFT) of the modulators $m_k[n]$, for all k subbands, resulting in a joint acoustic-frequency–modulation-frequency representation [87].

4.5 Scale Transform

The scale transform (ST) is a particular case of the Mellin transform, and defined as [88]:

$$D_x(c) = \int_0^{\infty} x(t)\mathrm{e}^{(-\mathrm{j}c-1/2)\ln t}\,\mathrm{d}t, \tag{4.31}$$

where $c \in \mathbb{R}$ is the scale variable. Its inverse is given by [89]

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} D_x(c)\mathrm{e}^{(\mathrm{j}c-1/2)\ln t}\,\mathrm{d}t. \tag{4.32}$$

The main property of the scale transform we will exploit in this work is the scale invariance [89], by means of which the transforms of signals $x(t)$ and its scaled version $\sqrt{a}x(at)$, with $a \in \mathbb{R}^+$, have the same magnitudes, differing only in phase. The values $a < 1$ and $a > 1$ correspond to scale expansion and compression, respectively.

The scale transform of a signal $x(t)$ can be computed from the Fourier transform of an exponentially warped version of the same signal, weighted by an exponential window. The Fourier transform of a signal defined as

$$x_{\text{scale}}(t) = x(\mathrm{e}^t)\mathrm{e}^{t/2} \tag{4.33}$$

is

$$\begin{aligned} X_{\text{scale}}(\mathrm{j}c) &= \int_{-\infty}^{\infty} x_{\text{scale}}(t)\mathrm{e}^{-\mathrm{j}ct}\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty} x(\mathrm{e}^t)\mathrm{e}^{t/2}\mathrm{e}^{-\mathrm{j}ct}\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty} x(\mathrm{e}^t)\mathrm{e}^{(1/2-\mathrm{j}c)t}\,\mathrm{d}t. \end{aligned} \tag{4.34}$$

After a change of variable $t^\star = \mathrm{e}^t$ ($t = \ln t^\star$),

$$\begin{aligned} X_{\text{scale}}(\mathrm{j}c) &= \int_{0}^{\infty} x(t^\star)\mathrm{e}^{(1/2-\mathrm{j}c)\ln t^\star}\,\mathrm{d}(\ln t^\star) \\ &= \int_{0}^{\infty} x(t^\star)\mathrm{e}^{(1/2-\mathrm{j}c)\ln t^\star}\frac{1}{t^\star}\,\mathrm{d}t^\star \\ &= \int_{0}^{\infty} x(t^\star)\mathrm{e}^{(1/2-\mathrm{j}c)\ln t^\star}\mathrm{e}^{-\ln t^\star}\,\mathrm{d}t^\star \\ &= \int_{0}^{\infty} x(t^\star)\mathrm{e}^{(-\mathrm{j}c-1/2)\ln t^\star}\,\mathrm{d}t^\star \\ &= D_x(c). \end{aligned} \tag{4.35}$$

Fast computation of the scale transform exploits this relation with the Fourier transform [89].

Other approaches provide a discrete approximation to that integral of Equation (4.31), assuming a constant value over logarithmic intervals. In the direct scale transform (DST) [90], the integral is approximated as

$$D_x(c) = \frac{1}{1/2-\mathrm{j}c}\sum_{n=1}^{\infty}\left[x(nT_\mathrm{s}-T_\mathrm{s}) - x(nT_\mathrm{s})\right](nT_\mathrm{s})^{1/2-\mathrm{j}c}. \tag{4.36}$$

This transform can be made efficient by precomputing the basis function matrix with elements $(nT_\mathrm{s})^{1/2-\mathrm{j}c}$ and has the added benefit of avoiding the non-linear interpolation of signal $x(t)$ that is required to obtain $x_{\text{scale}}(t)$ in conventional implementations.

Part II

Drum Sound Classification

Chapter 5

Features for Drum Sound Classification

In this chapter, we present an overview of the preprocessing steps and feature descriptors that are used for drum sound classification. These features are investigated and selected to be used in conjunction with supervised classification techniques reported in Chapter 7 for the task of drum stroke recognition.

Here, we bundled the features into two major groups. First, the literature descriptors, which correspond to features commonly used for the purpose of instrument recognition. These can be further divided into temporal-, spectral-, and cepstral-related features, according to the domains from which they are extracted. A different taxonomy of descriptors for music, speech and environmental sounds is available in [91]. The second group is composed of proposed features and includes the scattering transform, which, to the best of our knowledge, has not been applied in this task, and a modulation descriptor using a cascade of CQTs first presented here.

Most of the following features are thoroughly described in [92]; for the remaining features and whenever necessary we provide complementary references.

5.1 Signal Envelope

A signal envelope $v[n]$ is a "smooth" function[1] that traces the outline of a discrete-time signal $x[n]$, thus providing an approximation of its instantaneous amplitude. For a narrowband signal such as the ones described in Section 4.4, the modulator signal $m_k[n]$ acts as an envelope of the subband signal. However, envelopes can also be extracted from broadband signals.

There are several ways of constructing $v[n]$. Strictly speaking, two such outlines can be computed for any given signal — an upper and a lower boundaries. Con-

[1] Strictly speaking, an envelope of a signal can only be said "smooth" in the continuous domain. In the discrete domain, "smooth" refers to a signal without abrupt changes.

sidering as the real envelope only an upper boundary of the rectified signal $|x[n]|$, we could, for instance, compute either the instantaneous peak amplitude or the instantaneous root mean square (RMS) value of the input over a series of overlapping short windows, which respectively correspond to

$$v_{\text{peak}}[n] = \max_{m_0 \leq m \leq n} |x[m]| \tag{5.1}$$

$$v_{\text{rms}}[n] = \sqrt{\frac{1}{M} \sum_{m=m_0}^{n} x^2[m]}, \tag{5.2}$$

where $m_0 = n - M + 1$, and M, the window size, must be chosen to at least match the largest period expected in the signal [93]. Alternatively, an envelope can be derived from a non-symmetric low-pass filtering [94] that models the characteristics of short attack and long decay times commonly observed in musical signals.

Yet another possibility is to compute an analytic envelope of $x[n]$. First, we define the analytic signal

$$y[n] = x[n] + \mathrm{j}\mathcal{H}\{x\}[n], \tag{5.3}$$

where $\mathcal{H}\{\cdot\}$ is the Hilbert transform; and then construct an envelope by filtering its magnitude by an adequate low-pass filter $h[n]$ [95]:

$$v[n] = (h * |y|)[n]. \tag{5.4}$$

Finally, the true amplitude envelope (TAE) exploits a technique commonly employed for the computation of spectral envelopes in its dual domain [96]. Therefore, a time signal is first processed (rectified, zero-padded to the nearest power of 2, and concatenated with a time-reversed version of itself) to simulate the magnitude spectrum of a real signal. Then the true envelope technique is applied, which consists in liftering in the cepstral domain (see Section 5.2.3) and iteratively updating the estimated envelope. The optimal order for the liftering operation can be estimated from the fundamental period of the signal. In Figure 5.1, we present as an example the envelopes of a single drum stroke obtained from each of these five methods.

Observing the shapes of the envelopes in Figure 5.1, we can model an envelope of a generic percussive sound as a curve composed of two phases: an attack (of rapid energy increase) and an exponential decay phase [97]. This is called the attack–decay model, which is depicted in Figure 5.2 for continuous time. From the envelope signal, we can compute a small number of time-related descriptors that are discussed in the following (e.g., the log attack time, decay time constant, etc.).

The envelope can also be used in the detection and segmentation of events. However, unless signals are very percussive in nature, this feature does not usually

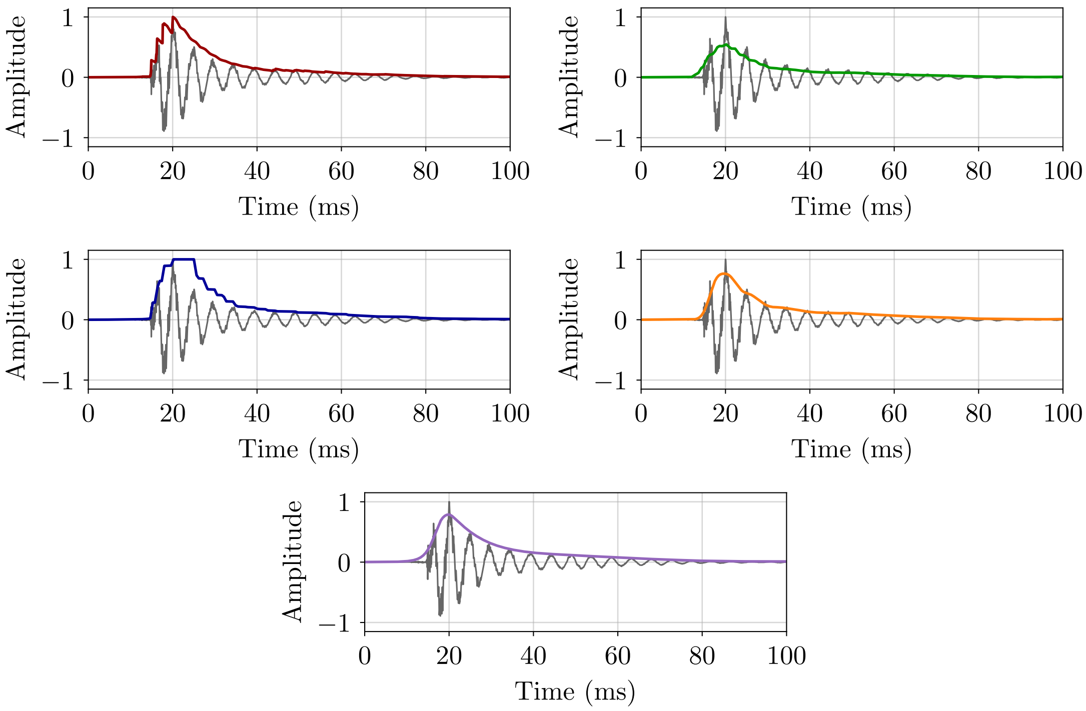


Figure 5.1: Examples of the computation of envelopes of a single drum stroke using: non-symmetric low-pass filter (top left), instantaneous RMS (top right), instantaneous window maximum (middle left), filtered analytical signal amplitude (middle right), and true amplitude envelope (bottom). Since the lowest frequency in this excerpt is about 200 Hz, we used a window length of 5 ms (with 0.5-ms hop length) where required. For the non-symmetric low-pass filter, we selected attack and release times of 0 and 5 ms, respectively. The true envelope was obtained with Hamming liftering of order 110 in the cepstral domain.

produce reliable results with peak picking post-processing methods, with its time derivative being better suited for this task [98]. For this reason, the literature on onset detection focuses less on this kind of time-domain descriptor, instead exploring onset detection algorithms that are based on other families of functions (e.g., from spectral, phase, and complex domains). Since they are not the subject of this thesis, these functions are briefly reviewed before our experiments in Chapter 7.

5.2 Descriptors from the Literature

In this section, the most commonly used descriptors in the classification of percussive sounds are reviewed. We assume that the signal $x[n]$, of length N, is the result of a segmentation process, i.e., it describes the waveform of a single drum stroke. Temporal descriptors are computed directly from the signal or from some envelope, $v[n]$. Spectral descriptors, as expected, are functions of the spectrum whereas cepstral descriptors are functions of the cepstrum. In both cases, descriptors are computed

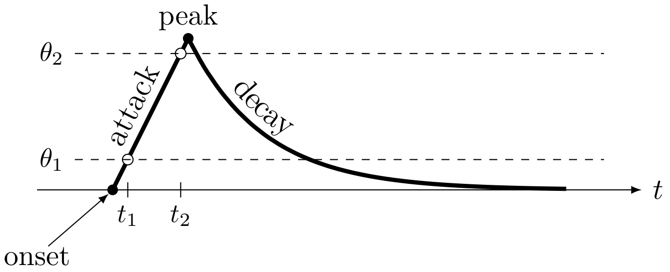


Figure 5.2: Attack–decay envelope model for percussive signals. We indicate in the figure two thresholds, θ_1 and θ_2, that help define the attack duration. These parameters are empirically chosen from the type of signal.

for each time frame and summarized by their mean and standard deviation. For simplicity, in this section we will denote the spectrum of a frame as $X[k]$ in place of $X[k, m]$, and we will only consider the K positive frequencies (possibly including DC), unless otherwise specified.

5.2.1 Temporal Descriptors

Log Attack Time

The log attack time is the logarithm of the total duration of the attack portion of a given signal. There are many ways to estimate the start and end of an attack, and in this work we use the fixed threshold method [92] considering the nature of the signals we study. To compute this estimation, we first define the attack start and stop thresholds at $\theta_1 = 20\%$ and $\theta_2 = 90\%$ of the envelope peak, respectively. We then compute the log attack time,

$$\Delta_{\log} n = \log(n_2 - n_1), \tag{5.5}$$

where n_1 and n_2 are, respectively, the indices of the samples where the envelope signal first exceeds the start and stop thresholds, respectively, during the attack phase. We show the relation between time instants and thresholds in Figure 5.2 for a continuous envelope.

Crest Factor

The crest factor is the ratio between the maximum absolute value and the RMS of the signal [99], i.e.,

$$\beta = \frac{\max\{|x[n]|\}}{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]}}. \tag{5.6}$$

This measure indicates how pronounced is the height of the signal peak relative to the RMS value. The minimum possible crest factor, $\beta = 1$, occurs when the signal has no peaks.

Temporal Decrease

The temporal decrease is a measure of the decay rate of the envelope signal. To compute it, we first take the decay portion of the signal,

$$v_{\text{decay}}[n] = v[n + n_{\max}]u[n] \tag{5.7}$$

where $n_{\max} = \arg\max_n\{v[n]\}$ is the sample at which the peak occurs and

$$u[n] = \begin{cases} 0, & n < 0, \\ 1, & n \geq 0 \end{cases} \tag{5.8}$$

is the unit step function. This decay can be modelled as an exponential function

$$v_{\text{decay}}[n] = A\mathrm{e}^{-\frac{n}{r}}u[n], \tag{5.9}$$

with a given decay rate r (in samples), which can be directly estimated through linear fit of the natural logarithm of $v_{\text{decay}}[n]$ [92].

In this work, for the computation of r, we follow the procedure described in [93] by observing that, for $n > 0$,

$$\begin{aligned} v_{\text{decay}}[n+1] &= A\mathrm{e}^{-\frac{n+1}{r}} \\ &= \mathrm{e}^{-\frac{1}{r}}\left(A\mathrm{e}^{-\frac{n}{r}}\right) \\ &= \alpha v_{\text{decay}}[n], \end{aligned} \tag{5.10}$$

if we define $\alpha = \mathrm{e}^{-1/r}$. This means that $v_{\text{decay}}[n+1]$ is directly proportional to its predecessor. The proportionality constant α can then be computed via the first-order autocorrelation as

$$\sum_{n=0}^{N_{\mathrm{d}}-2} v_{\text{decay}}[n]v_{\text{decay}}[n+1] = \sum_{n=0}^{N_{\mathrm{d}}-2} v_{\text{decay}}[n](\alpha v_{\text{decay}}[n]) = \alpha \sum_{n=0}^{N_{\mathrm{d}}-2} v_{\text{decay}}^2[n], \tag{5.11}$$

where the envelope decay has length N_{d}. Thus,

$$\alpha = \frac{\sum_{n=0}^{N_{\mathrm{d}}-2} v_{\text{decay}}[n]v_{\text{decay}}[n+1]}{\sum_{n=0}^{N_{\mathrm{d}}-2} v_{\text{decay}}^2[n]}, \tag{5.12}$$

from which the decay rate can be immediately retrieved as $r = -1/\ln\alpha$ in samples, or as rT_{s}, in seconds.

Zero-Crossing Rate

The zero-crossing rate (ZCR) measures the number of sign changes in the signal, i.e., from positive- to negative-valued samples and vice versa, with respect to the length of the signal or region upon which it is being computed:

$$\zeta = \frac{1}{2(N-1)} \sum_{n=0}^{N-2} |\operatorname{sgn}(x[n]) - \operatorname{sgn}(x[n+1])|, \tag{5.13}$$

where the sign function is defined here by counting zero as a positive value, i.e.,

$$\operatorname{sgn}(a) = \begin{cases} 1, & a \geq 0, \\ -1, & a < 0. \end{cases} \tag{5.14}$$

Noisy signals usually present a high ZCR, whereas lower ZCR values are a characteristic of better behaved (e.g., pitched) signals. This measure has been used in the literature as a discriminator of voiced/unvoiced sounds, as an estimate for the local fundamental frequency of monophonic signals, and even as a feature in the classification of percussive sounds produced by different instruments [97].

As we mentioned, the ZCR can be calculated for an entire signal or signal region, yielding a scalar. However, it can also be computed for each time frame, in which case it is generally presented accompanied by its mean and variance over the set of frames [95].

Moments

The moments of a function are scalars that are related to the shape of its curve. Since they provide information about the shape of a curve with a finite set of numbers [100], moments have been used in many applications including the classification of percussive sounds, where they can be computed from both envelope signals and spectra [92, 99, 101–103]. In this work, we can benefit from the interpretation given to moments in the field of statistics by observing that, similarly to a probability mass function (PMF), the functions (envelopes and magnitude spectra) over which we will be computing moments are non-negative by definition. In this section, we will present the definition for the moments of the envelope signal; the derivation for a magnitude spectrum is formally identical and should follow naturally if we replace the envelope with the magnitude spectrum and substitute frequency for time.

Before we can extract these values, we need to normalize the target function $v[n]$, of length N, by defining

$$\tilde{v}[n] = \frac{v[n]}{\sum_{n=0}^{N-1} v[n]}, \tag{5.15}$$

that adds up to one, like a proper PMF. Then, the p-th moment is given by

$$m_p = \sum_{n=0}^{N-1} n^p \tilde{v}[n]. \tag{5.16}$$

Note that $m_0 = 1$, and that the first moment about the vertical axis, m_1, is the mean or centroid of the distribution-like signal $\tilde{v}[n]$.

We can also compute a set of moments about the mean, called central moments, which are defined as

$$\mu_p = \sum_{n=0}^{N-1} (n - m_1)^p \tilde{v}[n]. \tag{5.17}$$

We can readily observe that low-order moments are not very informative, since $\mu_0 = 1$ and $\mu_1 = 0$. Higher-order central moments, however, provide more detailed information about the shape of a distribution.

First, the variance, which is a measure of the dispersion of the distribution around the centroid, and is expressed by the second central moment

$$\sigma^2 = \mu_2. \tag{5.18}$$

The square root of the variance, σ, is called the standard deviation.

The skewness measures the asymmetry of the distribution about the mean, and is calculated via the ratio

$$\gamma = \frac{\mu_3}{\sigma^3}. \tag{5.19}$$

Symmetric distributions have a skewness of zero, and negative/positive skewness indicates whether the curve is skewed to the left/right of the centroid.

Lastly, the kurtosis is obtained from the fourth central moment by doing

$$\kappa = \frac{\mu_4}{\sigma^4}, \tag{5.20}$$

and allows the assessment of the tail of the distribution, i.e., if the function is light- or heavy-tailed. It is common to present the kurtosis by its "excess",

$$\bar{\kappa} = \frac{\mu_4}{\sigma^4} - 3; \tag{5.21}$$

this is in fact a comparison between the function and a normal distribution, for which the fourth central moment is $\mu_4 = 3\sigma^4$. Distributions with short tails have negative excess kurtosis, whereas long-tailed ones (often sharper at the peak) show positive values in this measure.

To allow the comparison of moments derived from signals of different lengths, we can define length-normalized versions of the moments about the vertical axis and about the mean as

$$\begin{cases} \tilde{m}_p = \dfrac{m_p}{(N-1)^p}, \\ \tilde{\mu}_p = \dfrac{\mu_p}{(N-1)^p}, \end{cases} \tag{5.22}$$

respectively.

Strong Decay

This feature is computed from the non-linear combination between the signal energy and its temporal centroid such that [104]

$$\varsigma = \sqrt{\frac{E}{m_1}}, \tag{5.23}$$

where the signal energy is given by

$$E = \sum_{n=0}^{N-1} x^2[n]. \tag{5.24}$$

A signal has a high strong decay if it has high energy and centroid close to its start, and a low strong decay otherwise.

5.2.2 Spectral Descriptors

Spectral Energy

We can compute the root mean square spectral energy of a frame, which, due to Parseval's identity, is equivalent to the time-domain RMS (up to a factor of $1/\sqrt{N}$), as

$$E_{\text{rms}} = \sqrt{\frac{1}{K}\sum_{k=0}^{K-1} |X[k]|^2}. \tag{5.25}$$

Even more interesting is the characterization of the spectral energy in the different signal subbands, i.e., how energy is distributed along different portions of the spectrum. These energies are commonly computed over non-overlapping rectangular frequency bands. The energy in a subband b is

$$E_b = \sum_{k=k_{1,b}}^{k_{2,b}} |X[k]|^2, \tag{5.26}$$

for which the limits $k_{1,b}$ and $k_{2,b}$ are empirically determined.

Moments

Moments can also be used to characterize the spectral shape. Similarly to what we did for time signals, we define the p-th spectral moment of a frame as

$$m'_p = \sum_{k=0}^{K-1} k^p \tilde{X}[k], \tag{5.27}$$

where

$$\tilde{X}[k] = \frac{|X[k]|}{\sum_{k=0}^{K-1} |X[k]|} \tag{5.28}$$

is the normalized magnitude spectrum.

Central moments for spectra follow the same definition of Equation (5.17), i.e.,

$$\mu'_p = \sum_{k=0}^{K-1} (k - m'_1)^p \tilde{X}[k]. \tag{5.29}$$

We point out that the spectral centroid,

$$m'_1 = \frac{\sum_{k=0}^{K-1} k|X[k]|}{\sum_{k=0}^{K-1} |X[k]|}, \tag{5.30}$$

was shown to correlate well to the perception of brightness of music tones [105], which is an aspect of timbre related to the energy proportion of high- to low-frequency regions of a sound spectrum.

Spectral Flatness

The spectral flatness, sometimes also called the tonality coefficient,[2] is a measure of how noise-like the magnitude spectrum is, reaching the maximum value of one for a flat spectrum (e.g., white noise). If the energy is concentrated in few spectral regions, the flatness measure is low, approaching zero for pure-tone-like sounds.[3] It is computed as the ratio between the geometric and arithmetic means of the magnitude spectrum of a time frame,[4]

$$\mathcal{F} = \frac{\sqrt[K]{\prod_{k=0}^{K-1} |X[k]|}}{\frac{1}{K}\sum_{k=0}^{K-1} |X[k]|}. \tag{5.31}$$

[2] Not to be confused with the concept of "tonality" in music, namely, the organization of tones around a reference (tonic). Note also that, in the case of this descriptor, "flatness" and "tonality" indicate opposite concepts: the "flatter" the spectrum, the less "tonal" it is, and vice versa.

[3] A full-range sinusoidal sweep within a frame would be regarded as "flat" by this descriptor, despite containing many different tones.

[4] For numerical accuracy, the geometric mean is usually computed as the arithmetic mean of the logarithm of the magnitude values, which is then converted back via exponentiation.

Its value is often converted to a decibel scale with a range of $(-\infty, 0)$.

Spectral Crest

The spectral crest is a measure of the "peakiness" of a spectrum. It is defined as the ratio between the peak and mean values of the magnitude spectrum. Mathematically, for a given frame, we can write it as

$$C = \frac{\max |X[k]|}{\frac{1}{K}\sum_{k=0}^{K-1} |X[k]|}. \tag{5.32}$$

Observe that the value of the spectral crest is bounded between 1 and K, provided that the spectrum is not zero for all k. The spectral crest is lower the flatter the magnitude spectrum, and higher if most of the energy is contained in fewer bins.

Strong Peak

The strong peak feature measures how pronounced the spectral peak is with respect to its bandwidth, B, defined at the -6 dB (half amplitude) points around the maximum [104]

$$\xi = \frac{\max |X[k]|}{B}. \tag{5.33}$$

The sharper and higher the spectrum peak at a frame, the "stronger" it is qualified; whereas for frames with flat spectra, the value of the strong peak is 0 by default.

The bandwidth B can be computed in a linear frequency scale, as the difference between the extreme frequencies to the right and to the left of the peak that lie above the half-amplitude threshold, or in log-scale, as the logarithm of their ratio. This latter method allows a comparison between peaks in different ranges that is better correlated with human perception [106].

Spectral Roll-off

The spectral roll-off is a measure of the signal bandwidth defined as the frequency below which a certain percentage θ (e.g., 95%) of the spectral energy lies, i.e., the index ρ such that,

$$\sum_{k=0}^{\rho} |X[k]|^2 = \theta \sum_{k=0}^{K-1} |X[k]|^2. \tag{5.34}$$

Spectral Contrast

The spectral contrast is a measure of the relative distribution of magnitudes in selected subbands of the spectrum. We can interpret it as a measure of the relation between harmonic and non-harmonic frequency components in each of the subbands [107], where strong peaks usually correspond with the former and valleys reflect the stochastic noise-like characteristics of the latter.[5] It is computed as follows. First, the spectrum is divided into octave-scaled subbands, typically six. For each band k, the algorithm sorts the bins according to their magnitudes and computes the peak and valley averages, respectively P_k and V_k, of a percentage of the topmost and bottommost values. The spectral contrast values are then computed as

$$C_k = \log \frac{P_k}{V_k}. \tag{5.35}$$

Spectral contrast and valley values for all subbands are commonly concatenated in a vector and used as a feature after decorrelation with a Karhunen-Loève transform (KLT). It has shown good results in genre classification [107, 108] and in the detection and identification of drum sources extracted from polyphonic mixtures [109].

AKKERMANS et al. [107] propose a modification to the computation of spectral contrast coefficients by using as the logarithm's base the average subband magnitude of the frame, A_k. Authors posit that this updated contrast measure better describes the shape of the subband, since it is able to distinguish between spectra with similar P_k and V_k, but of different profiles. They suggest a different perceptually inspired subband division that guarantees that all subbands contain enough bins. The new measure is called shape-based spectral contrast and is shown to improve discriminability in classification while being more robust to compression [107].

5.2.3 Cepstral Descriptors

The cepstrum of a signal or frame is computed as the inverse Fourier transform of the logarithm of a signal spectral magnitude. Its domain is that of the quefrency, which is in a sense a measure of time and was coined by inverting the order of syllables in the word "frequency". Similarly the term "cepstrum" was derived from "spectrum", by reversing the first syllable. We can write the real cepstrum as

$$c[\nu] = \frac{1}{M} \sum_{k=0}^{M-1} (\log |X[k]|) \mathrm{e}^{\mathrm{j}\Omega_k \nu}, \tag{5.36}$$

where $\nu = 0, 1, \ldots, L$ is the quefrency variable, L is the desired number of cepstral coefficients, and $\Omega_k = 2\pi k/M$, and M is the number of DFT bins.

Due to the convolution theorem and to the application of the logarithm, the cepstrum has the interesting property of turning into summands components that have been convolved in the original time signal. This deconvolution is quite useful in the analysis of signals that can be represented by a source–filter model, for example,

[5] Note that, unlike the spectral flatness, which assesses the general spectral shape, the spectral contrast describes the dynamic range within the spectrum.

such as human speech. If the cepstral components lie in different quefrency regions, they can be filtered in this domain (a process called liftering, an anagram of “filtering”) and the components can be retrieved in the time domain [110]. Music signals, in most cases, do not admit this kind of modelling. However, cepstral analysis is useful for the analysis of musical instruments, since it is able to characterize the spectrum overall envelope (low quefrency) and its periodicities (higher quefrencies).

Different perceptually motivated filter banks have been used in the literature as a replacement for the linearly distributed Fourier transform filter bank when computing the cepstrum. Three of these variants are presented in the following; we do not include here the constant-Q cepstral coefficients proposed by Brown in [111], which was used with modest results in instrument recognition and which, more recently, has shown good results in speaker verification [112]. When used in classification tasks, cepstra are frequently computed framewise and presented along with their first and second discrete-time derivatives, δc and $\delta^2 c$, usually summarized over time with the mean and standard deviation.

Mel-Frequency Cepstral Coefficients

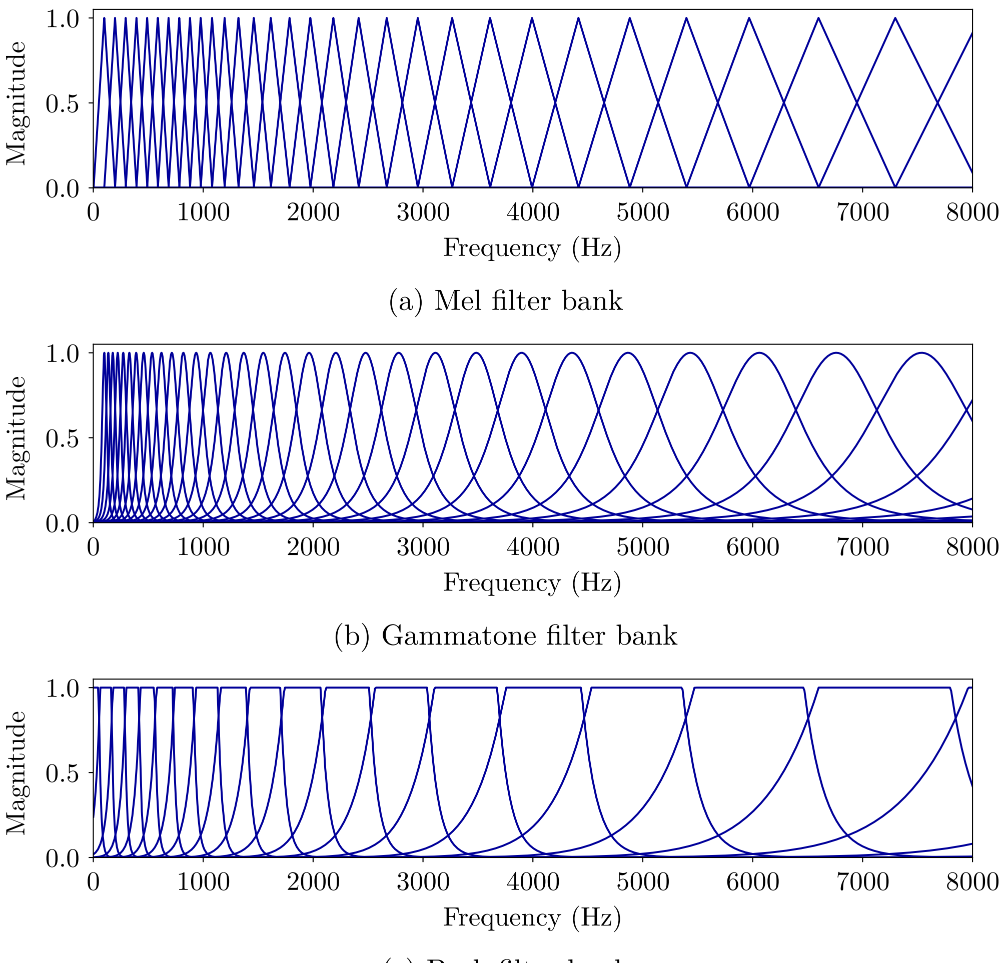
The mel scale was derived from psychophysical studies after observing that human perception of the pitch of simple tones does not follow a linear scale. One possible way to convert linear frequencies (in Hz) to the mel scale is [113]

$$m(f) = 2595 \log\left(1 + \frac{f}{700}\right), \tag{5.37}$$

in order that a 1 kHz tone, 40 dB above the hearing threshold, is equivalent to 1000 mels [114]. For frequencies above 1 kHz, the perceived pitch increases in logarithmic fashion, whereas well below this reference point it behaves almost linearly. Other transformations from hertz to mel can be seen in [115].

The spectrum of a frame can be represented in the mel scale by weighting the magnitude of the DFT bins with a filter bank of $\hat{K}$ uniformly-spaced filters in the mel scale. Each k-th filter is a bandpass triangular-shaped curve with constant bandwidth so that neighboring channels overlap by 50%. Filters are also commonly normalized to unit area or energy. Figure 5.3a exemplifies this filter bank. Then, for each $\hat{k} = 0, 1, \ldots, \hat{K} - 1$, we integrate over k the product of the magnitude spectrum and the corresponding triangular filter to obtain the modified spectrum, $Y[\hat{k}]$, in the mel scale. Finally, the mel-frequency cepstral coefficients (MFCC) for that frame are computed with the aid of a type-II discrete cosine transform (DCT) as

$$c_{\text{mel}}[\nu] = \sum_{\hat{k}=0}^{\hat{K}-1} \log(Y[\hat{k}]) \cos\left[\frac{\pi}{\hat{K}}\left(\hat{k} + \frac{1}{2}\right)\nu\right]. \tag{5.38}$$



(a) Mel filter bank

(b) Gammatone filter bank

(c) Bark filter bank

Figure 5.3: Filter banks for the computation of mel-, gammatone-, and Bark-frequency cepstral coefficients. Plots (a) and (b) present example filter banks that were designed with 40 channels up to the Nyquist rate ($f_s/2 = 22.05$ kHz). Filters are spaced in frequency according to the mel and gammatone scales, respectively, with the first center frequency at 100 Hz. In (c) we show a 24-band filter bank for obtaining the Bark spectrum. For the purpose of illustration, only a portion of the spectrum is shown and filters are unnormalized.

Gammatone-Frequency Cepstral Coefficients

Another way to include a perceptual flavor in the cepstral representation is to first process the signal through a set of gammatone filters. These filters model the tonotopic organization of the basilar membrane by simulating the way different regions of the cochlea are excited by specific frequency ranges, the critical bands, which get increasingly broader with increasing frequency [113]. The impulse response of each gammatone filter is defined in the time domain by a sinusoidal carrier whose amplitude is modulated by a curve shaped like a gamma probability density function [116].

In continuous time, the gammatone filters can be expressed by

$$g(t) = At^{p-1}\mathrm{e}^{-\lambda t}\cos(2\pi f_{\mathrm{c}}t + \phi)u(t), \tag{5.39}$$

where A is the amplitude, p is the order of the filter, λ is the decaying factor, f_{c} is the center frequency, and ϕ is the carrier initial phase. Filters are distributed along frequency in proportion to their respective bandwidths, which are given by the equivalent rectangular bandwidth (ERB) scale derived from notch-noise experiments [117]

$$\mathrm{ERB}(f) = 24.7\left(4.37\frac{f}{1000} + 1\right). \tag{5.40}$$

The signal is analyzed by the gammatone filter bank (see Figure 5.3b) and the output of each channel is commonly downsampled to an appropriate frame rate (e.g., 100 Hz), fully rectified and compressed with a cubic root operation [118]. Then, the gammatone-frequency cepstral coefficients (GFCC) can be obtained from each frame, $Y[\hat{k}]$, of the resulting "cochleagram" as

$$c_{\mathrm{gamma}}[\nu] = \sqrt{\frac{2}{\hat{K}}}\sum_{\hat{k}=0}^{\hat{K}-1} Y[\hat{k}]\cos\left[\frac{\pi}{\hat{K}}\left(\hat{k} + \frac{1}{2}\right)\nu\right]. \tag{5.41}$$

Observe that since the coefficients have been derived with a cubic root non-linearity, they do not characterize a proper cepstrum. Logarithmic compression can be used instead, which provides scale invariance but less robustness to noise in speaker identification [119].

Bark-Frequency Cepstral Coefficients

An estimation for the critical bandwidths along the basilar membrane is given by the Bark scale, which was devised by Zwicker after loudness summation experiments [120]. It can be regarded as defining a linear basilar distance measure whose relation to frequency can be expressed by

$$z(f) = 13\arctan\left(0.76\frac{f}{1000}\right) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right], \tag{5.42}$$

for z in Bark and f in Hz. This equation is not used much in practice since it does not admit an inverse. A simple (and invertible) approximation is given by

$$z(f) = \left[\frac{26.81f}{1960 + f}\right] - 0.53, \tag{5.43}$$

for which minor corrections are needed if the Bark values smaller are than 2.0 or greater than 20.1. We refer the reader to [121] for more expressions of the critical-band rate.

The procedure for computing Bark-frequency cepstral coefficients (BFCC) is very similar to the one described for MFCCs if we replace the warping function of Equation (5.37) by that of Equation (5.42). The triangular filters can also be substituted by trapezoidal ones that have constant gain within each critical band and linear decay slopes (in dB) at band transitions [122]. It is common to summarize the spectral magnitude/energy in 24 to 26 Bark bands, since this is about the number of adjacent critical bands found in the basilar membrane [123]. A Bark filter bank is exemplified in Figure 5.3c.

The use of Bark bands has proven to be successful in the classification of percussive sounds [124], and corresponding cepstra were shown to produce slightly more acurate results than MFCCs in the same task [125, 126], which could be attributed to the fact that the Bark weighting better preserves mid tones in detriment of higher frequencies when compared to mel scaling [127].

5.3 Proposed Descriptors

In this section, we define a different class of features that has not been explored much for the analysis of instrument timbre, and even less for the description of drum signals. We present a novel formulation, based on the CQT, for deriving the modulation spectrum of a signal. We also describe the wavelet scattering transform [128], which allows the representation of a signal in different levels of detail while also capturing the interferences between frequencies within this signal.

5.3.1 CQT-based Modulation Spectrum Coefficients

The modulation spectral transform, whose fundamentals were previously discussed in Section 4.4, has been used in both speech and music processing for many different tasks, e.g., speech and speaker recognition [129–133], speech detection in noisy environments [134], vocal emotion recognition and conversion [135, 136], audio and music genre classification [137–139], multiple-pitch and multiple-instrument recognition [140], music emotion classification [141], and source separation [142, 143].

As we alluded before, the modulation spectrum of an audio signal can be derived by the cascaded application of two filter bank analyses or time-frequency transforms along the time dimension. This leads to a representation in a joint acoustic- and modulation-frequency domain. There is no consensus regarding a best practice in the extraction of this representation and different researchers suggest different methods, which might meet the principles of an underlying subjective model, for example. A modulation spectrum can be obtained through the computation of the DFT of each channel of a signal's magnitude spectrogram [130, 131, 139, 142]. Spectrogram bins can also be grouped prior to the second transformation to provide an octave-scale acoustic frequency resolution [138]. A constant-Q resolution can also be achieved for the modulation frequency by appropriately grouping STFT bins or by using a continuous wavelet transform [144] — this takes into consideration how modulation structures are actually perceived by the human auditory system. Other efficient implementations use the time-domain aliasing cancellation filter bank for both stages [145], or for the first stage only, which is followed by a hierarchical lapped transform with octave-band structure [146]. Some works exploit structures inspired in auditory models for both stages. A few authors achieve this by computing the first analysis with a gammatone filter bank and then empirically selecting frequency bands from each channel in the second stage (usually following a log or quasi-log distribution) [132, 135, 137, 143]. Mel-scale aggregation and other auditory-inspired (critical band) filter banks are also featured in some works at the base decomposition [129, 131, 134, 136, 147, 148]. In approaches that use auditory filter banks at the first stage, additional operations must be done to demodulate the signal at each channel, i.e., derive a temporal envelope, before computing the second analysis. Common choices are half-wave rectification followed by low-pass filtering [129, 143, 147], or Hilbert envelope extraction [132, 136]. In either case, it is customary to reduce the amount of data by decimating the resulting envelopes. We note that coherent demodulation schemes have received far less attention [84, 149].

When dealing with classification tasks, authors might directly use the modulation spectrum as feature representation. However, the regular computation using a cascade of STFTs usually generates too large a dimension for classification to be feasible. Therefore, some techniques have been presented to reduce the dimensionality of the modulation spectrum by aggregating bins on either axes or by smoothing the overall representation with a DCT [131]. Of course, logarithmically-scaled filter banks at both stages also achieve a reduction in dimension at the expense of the final resolution. LEE et al. [138] propose a derived feature, the octave-based modulation spectral contrast, and perform a genre classification over feature vectors containing mean and standard deviation of the spectral contrast and valleys (see Section 5.2.2) extracted from all modulation subbands. In other related works, alternative modulation features are obtained from short-time cesptral representations [150, 151] or from spectro-temporal modulations jointly computed over both axes of a time-frequency representation [152]. These features reflect different aspects of timbre.

Here, we compute the modulation spectrum using a different approach that provides logarithmic resolution on both stages of the transform by using constant-Q filters. First, we perform a short-time CQT analysis of the time signal $x[n]$. This base transform has geometric spacing and resolution in frequency that are determined by Q_1, the quality factor, and we select an initial center frequency Ω_0 according to the minimum frequency present in the signal. The number of filters in this analysis is limited by the Nyquist frequency, $f_\mathrm{s}/2$. We can express this time-frequency representation as (cf. Equation (4.19))

$$X_\mathrm{CQ}[k_1, m_1] = \frac{1}{N_{k_1}} \sum_{n=0}^{N_{k_1}-1} x[n] w_{k_1}[n - m_1 h_1] \mathrm{e}^{-\mathrm{j}\Omega_{k_1} n}, \tag{5.44}$$

where k_1 and m_1 are respectively the acoustic frequency bin and the frame indices, Ω_{k_1} is the center frequency of the k_1-th bin, and $w_{k_1}[n]$ is a window function of length N_{k_1} (which depends on the bin number, as seen in Equation (4.17)). The hop parameter, h_1, is made equal for all channels and it plays a very important part in determining the maximum modulation frequency that is to be represented on the second stage. Indeed, it defines the frame rate of $X_\mathrm{CQ}[k_1, m_1]$, $f_\mathrm{m} = f_\mathrm{s}/h_1$, thus limiting the modulation analysis range to $f_\mathrm{m}/2$.

We then transform the magnitude of $X_\mathrm{CQ}[k_1, m_1]$, which represents the amplitude fluctuations of the k_1-th frequency component as a function of time (frame counter m_1), with a second short-time CQT. This yields a two-dimensional representation of modulation frequency versus time (frames) for this channel. The second CQT possibly has different quality factor, Q_2, and initial frequency, Ω'_0. This operation is repeated for all acoustic frequency components, and we obtain

$$Y[k_1, k_2, m_2] = \frac{1}{N_{k_2}} \sum_{m_1=0}^{N_{k_2}-1} |X_\mathrm{CQ}[k_1, m_1]| w'_{k_2}[m_1 - m_2 h_2] \mathrm{e}^{-\mathrm{j}\Omega_{k_2} m_1}, \tag{5.45}$$

for which k_2 and m_2 correspond to the modulation frequency and the frame indices, respectively, $w'_{k_2}[n]$ is a window function of length N_{k_2}, and Ω_{k_2} is the modulation center frequency of the k_2-th bin. Finally, reminiscing of the periodogram, we can average the magnitude (or the squared magnitude) $|Y[k_1, k_2, m_2]|$ over the time variable, m_2:

$$\bar{Y}[k_1, k_2] = \frac{1}{M_2} \sum_{m_2=0}^{M_2-1} |Y[k_1, k_2, m_2]|, \tag{5.46}$$

where M_2 is the total number of frames after the second stage. This average results into an estimate of the acoustic versus modulation frequency characteristics of the original signal, which we call the modulation spectrogram. Figure 5.4 illustrates the process for the computation of $\bar{Y}[k_1, k_2]$.

Observe that not all combinations of k_1 and k_2 lead to valid coefficients in the modulation spectrogram. We can discard the values of $\bar{Y}[k_1, k_2]$ for pairs (k_1, k_2)

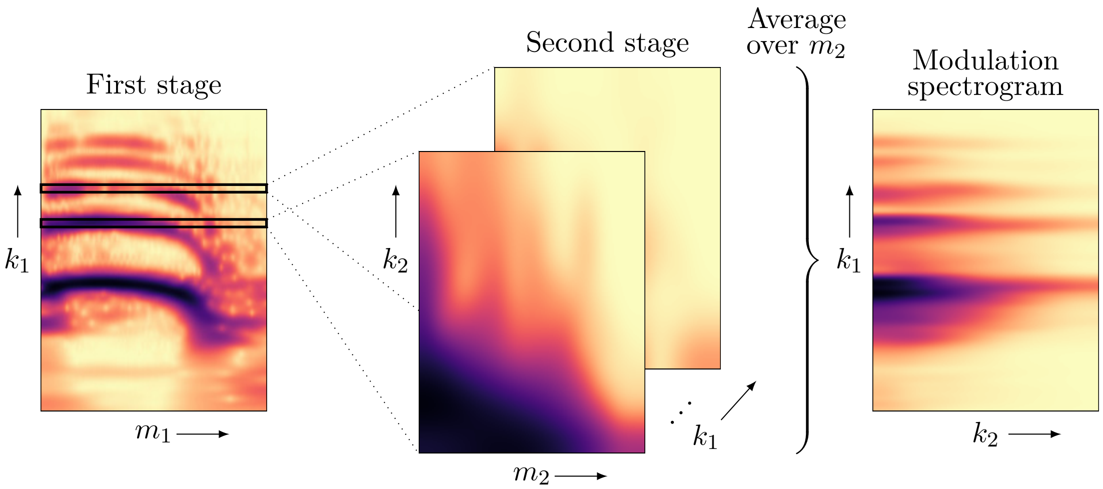


Figure 5.4: Stages in the proposed computation of the modulation spectrogram.

that correspond to a modulation filter with a center frequency higher than that of the acoustic filter, which have no physical meaning. We could alternatively take into account the bandwidths of each filter in the first stage since they are responsible for limiting the frequencies of the corresponding estimated modulators. A similar approach is used in the computation of the scattering coefficients (see Section 5.3.2).

The efficient implementations of the short-time CQT algorithm we mentioned before can be used to make the computation of the modulation spectrogram less expensive. We observe that the recursive subsampling algorithm of SCHÖRKHUBER and KLAPURI [80] includes some limitations in the hop length that could result in an insufficient number of frames at the input of the second stage. This could possibly be remedied by lowering the resolution (i.e., the number of bins per octave) of the first transform, all other conditions remaining constant. Alternatively, the variable-Q transform (VQT) could be used as input representation instead of the CQT, since it allows the user to reduce the length of the analysis windows (at the expense of the frequency resolution), and would therefore soften the hop constraints [153].

5.3.2 Scattering Coefficients

The wavelet scattering transform [128] (WST) has gained growing attention in the last few years for its ability to provide robust features for audio and image classification tasks. It yields a sparse signal representation (i.e., concentrates energy at a few coefficients), which is locally invariant to time shifts and stable to small-scale deformations caused by warping (e.g., time warping).

This representation is computed by a process not unlike that of a convolutional neural network (CNN), which consists of three main steps: filtering, non-linearity transformation, and pooling. In the wavelet scattering transform, however, the signal is convolved with functions from a fixed (i.e., non-trainable) K-band complex wavelet filter bank. This filter bank is constructed by dilating a mother wavelet $\psi(t)$, whose Fourier transform $\Psi(\omega)$ has bandpass magnitude response and center frequency normalized to unit. For audio applications, the daughter wavelets are obtained by the scaling equation

$$\psi_k(t) = 2^{\frac{k}{Q}} \psi(2^{\frac{k}{Q}} t), \tag{5.47}$$

with $k, Q \in \mathbb{Z}$. Observe that the Fourier transform of the dilated wavelet $\psi_k(t)$ is expressed by

$$\Psi_k(\omega) = \Psi(2^{-\frac{k}{Q}} \omega), \tag{5.48}$$

and, therefore, has center frequency $2^{\frac{k}{Q}}$. The Q parameter indicates the number of filters per octave, in such a manner that the bandwidth of each wavelet in the frequency domain is of the order of Q^{-1}. Following the filter bank analysis, we discard the complex phase of each channel's output and retain only the absolute value with a modulus non-linearity. This is equivalent to the Hilbert envelope demodulation seen in Sections 4.4.1 and 5.1. Finally, at the last step, the scattering coefficients are obtained by time-averaging the moduli with a low-pass filter with impulse response $\varphi(t)$ of duration T. This parameter determines the local invariance scale. This time-averaging removes the high-frequencies present in the scalogram coefficients. However, this information is not lost by the scattering transform, but recovered in a subsequent layer of wavelet convolutions and modulus operators. The output of this layer is also low-pass filtered by $\varphi(t)$, which again ensures local time-shift invariance. The process can be repeated again to recover smaller details in as many layers as desired. The wavelet scattering transform is therefore determined by a cascading of operations, as pictured in Figure 5.5.

We now go into more detail of what the WST computes at each layer. By default, for the zeroth layer, we have as output

$$\mathcal{S}_0\{x\}(t) = (x * \varphi)(t). \tag{5.49}$$

At the first layer, $x(t)$ is analyzed by a filter bank of wavelets $\psi_{k_1}(t)$ and, after the non-linearity, we obtain

$$\mathcal{U}_1\{x\}(t, k_1) = |x * \psi_{k_1}|(t). \tag{5.50}$$

The Fourier transform of the low-pass filter, which we indicate by $\Phi(\omega)$, and that of the wavelets $\Psi_k(\omega)$ are designed to cover the entire signal spectrum, with the center frequencies and bandwidths of the wavelets approximately following the mel-frequency scale. The frequency resolution in this layer is determined by Q_1, which is usually set to eight wavelets per octave in the high-frequency range, whereas $Q_1 - 1$ linearly spaced filters are placed in the lower frequency region [128]. The output

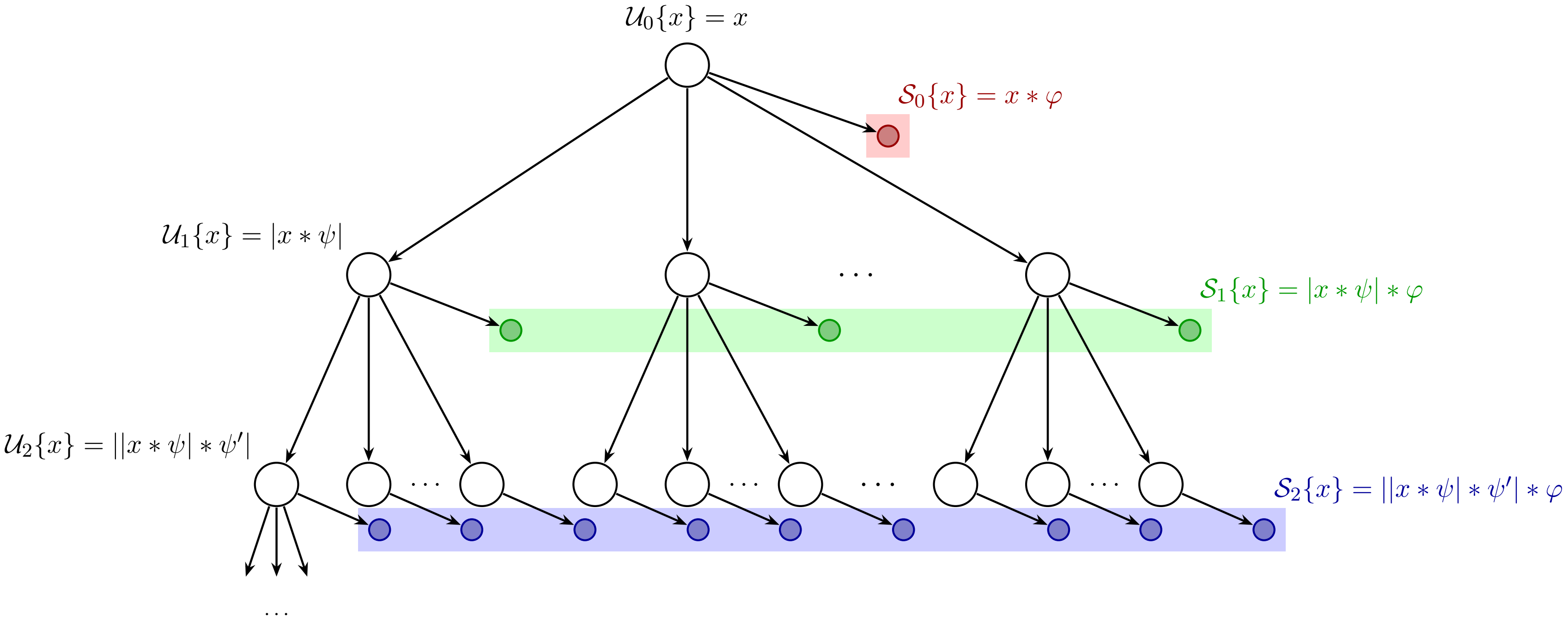

$\mathcal{U}_0\{x\} = x$
$\mathcal{S}_0\{x\} = x * \varphi$
$\mathcal{U}_1\{x\} = |x * \psi|$
$\mathcal{S}_1\{x\} = |x * \psi| * \varphi$
$\mathcal{U}_2\{x\} = ||x * \psi| * \psi'|$
$\mathcal{S}_2\{x\} = ||x * \psi| * \psi'| * \varphi$

Figure 5.5: Wavelet scattering network. The graph is presented in a simplified notation and explicitly shown up to order $l = 2$. We removed the time and frequency variables, and notated the first- and second-order wavelet filter banks as ψ and ψ', respectively. The main (clear) nodes at each level represent the wavelet modulus computation for wavelets of different center frequency and bandwidth, except for the zeroth level (top). Colored nodes, in their turn, indicate the network outputs, i.e., the scattering coefficients. Each output node correspond to a unique path defined by the frequency variables $(k_1, k_2, \ldots, k_l)$ up to the desired level l.

from this layer is the first-order scattering coefficients,

$$\mathcal{S}_1\{x\}(t,k_1) = (|x * \psi_{k_1}| * \varphi)(t), \tag{5.51}$$

which can be shown to approximate a mel spectrogram [128]. We then represent the detail lost at this step using another filter bank, $\{\psi_{k_2}(t)\}$, in which the wavelets are distributed with a possibly different resolution to allow a sparse representation (e.g., $Q_2 = 1$, one wavelet per octave). This detail,

$$\mathcal{U}_2\{x\}(t,k_1,k_2) = ||x * \psi_{k_1}| * \psi_{k_2}|(t), \tag{5.52}$$

is again low-pass filtered to make up the second-order coefficients

$$\mathcal{S}_2\{x\}(t,k_1,k_2) = (||x * \psi_{k_1}| * \psi_{k_2}| * \varphi)(t), \tag{5.53}$$

which carry significant information about interferences and amplitude modulations [128]. For the l-th layer, we thus have

$$\begin{cases} \mathcal{U}_l\{x\}(t,k_1,k_2,\ldots,k_l) = ||\cdots||x * \psi_{k_1}| * \psi_{k_2}| * \cdots| * \psi_{k_l}|(t) \\ \mathcal{S}_l\{x\}(t,k_1,k_2,\ldots,k_l) = (||\cdots||x * \psi_{k_1}| * \psi_{k_2}| * \cdots| * \psi_{k_l}| * \varphi)(t) \end{cases} \tag{5.54}$$

as the detail recovered from the $(l-1)$-th layer and the l-th-order scattering coefficients, respectively. Typically, the scattering transform is computed with maximal order $l = 2$, since most of the energy lies in the first- and second-order coefficients, and is quickly dissipated in lower levels [128]. In this respect, the final representation can be written as

$$\mathcal{S}\{x\} = (S_0\{x\}, S_1\{x\}, S_2\{x\}), \tag{5.55}$$

where we have removed the time and frequency variables for simplicity.

Note that, similarly to what we mentioned before in the case of the modulation spectrum, not all paths $(k_1, k_2, \ldots, k_l)$ yield significant coefficients. As ANDÉN and MALLAT [128] point out, since $|x * \psi_{k_1}|(t)$ has a limited bandwidth equivalent to that of the magnitude response of $\psi_{k_1}(t)$, in the analysis of the second filter bank one should only consider values of k_2 such that the supports of the filter and of the demodulated signal intercept in the frequency domain. In efficient implementations of the wavelet scattering transform, signals are subsampled by a factor that depends on their bandwidths or on the scale of the low-pass filter, and convolutions are calculated with the help of the FFT algorithm.

The scattering representation can be normalized so that coefficients at different orders are decorrelated, which can improve accuracy in classification tasks [128].

This requires that, at any order l, coefficients be renormalized by the components of the previous order [128], i.e.,

$$\tilde{\mathcal{S}}_l\{x\}(t, k_1, \ldots, k_{l-1}, k_l) = \frac{\mathcal{S}_l\{x\}(t, k_1, \ldots, k_{l-1}, k_l)}{\mathcal{S}_{l-1}\{x\}(t, k_1, \ldots, k_{l-1}) + \epsilon}, \tag{5.56}$$

where ϵ is a silence detection threshold. It is also habitual to aggregate the scattering coefficients by averaging them along the time variable.

Finally, we point out that scattering can also be carried through in the frequency dimension. Two time-frequency scattering techniques are possible to achieve representations that are locally invariant to time and frequency transpositions as well as robust to small deformations in both domains. First, in the separable scattering, a frequency scattering can be computed by analyzing the first- or second-order time-scattering scalogram with a wavelet filter bank defined along the log-frequency axis [154]. In the joint scattering, however, the final representation is obtained by decomposing a scalogram with a two-dimensional wavelet transform, followed by modulus and time averaging. In this case, the two-dimensional kernels are usually defined as the product of two independent time and frequency wavelets. ANDÉN et al. [155] show that a joint time-frequency scattering transform achieves state-of-the-art results in the classification of musical instruments and of acoustic scenes. The Scattering transform has also been used in other works for the recognition of instruments [156], instrument shapes [157], and instrumental playing techniques [158, 159].

Chapter 6

Physics of Sound Production and Literature Review

In the “Western” music tradition, the primary role of percussion is to support and demarcate the rhythm. Given the impulse-like characteristics of percussive sounds, signals from these kinds of sources are commonly modeled by automatic drum transcription (ADT) systems as composed of a short attack region which is followed by an exponential decay profile (see Section 5.1). The typical ADT pipeline starts from a sound recording of one or multiple percussion instruments (with or without the presence of other non-percussive parts and singing voices) and includes either the detection and classification of note-segments (i.e., portions of the recording corresponding to each single note) or the retrieval of onset times from separated single-instrument streams. In both cases, most works reported in the literature assume this rhythm-keeping view of percussion, and it becomes sufficient in the process of transcription to determine which percussion instruments have been played and when. This approximates drum transcription to the task of instrument recognition [160].

During the Romantic era, Western compositions began to undergo a series of changes that have had a lasting impact on musical performance. In particular, many composers started to exploit not only the rhythmic properties of percussion instruments, but also their “tone coloring” capabilities [2]. A great exponent of this movement was Hector Berlioz, who is known for expanding the use of percussion in the orchestra, in particular, of the kettledrums (timpani), which were featured in some pieces with up to 16 simultaneous instruments executed by eight instrumentalists. Moreover, Berlioz notated with considerable clarity how he expected percussion (kettledrums, cymbals, tenor and bass drums, etc.) to be played, providing instructions for the use of hard- or soft-ended drumsticks and for the muffling of the instruments [2], for example. In the late nineteenth and early twentieth century, composers like Rimsky-Korsakov, Debussy, Stravinsky, and Bartók furthered the developments in this matter [2]. We highlight here the indications given in later pieces by Bartók (e.g., *Cantata Profana*, Sz. 94), which specified the striking locations at the center or the edge of the snare drum, as well as the coupling/decoupling of the snares, to provide greater timbral variety [2]. The interested reader can further delve into this subject by referring to the book by BLADES [2], "Percussion Instruments and their History", where the author (also a percussionist) thoroughly describes the innovative use of percussion in orchestral music and how it evolved during the Romantic and post-Romantic periods.

The importance of percussive timbre and its subtle variations becomes apparent earlier in the music of other (non-Western) sources. For example, the control of pitch and quality in the sound produced by African drummers was narrated with awe by European travellers of the eighteenth and nineteenth centuries. They note, for example, how different methods — hand, hand and stick, or double-stick drumming — are employed depending on the type of instrument and function [2]. Timbral variation was also obtained by changing the striking positions, the weight and release of each attack, or by muting the drumhead, not to mention the use of a diversity of hand shapes and parts (in "hand" or "hand and stick" techniques) [2]. As SCHLOSS [93, p. 49] points out:

> One important difference between African and Western percussion is the notion of a "stroke-space," which can be defined as the universe of possibilities of ways of striking the drum; it is the vocabulary of strokes. Although in Western percussion the emphasis is usually on the uniformity of tone, in African percussion, how the drum is struck is almost as important as when.

In "Studies in African Music", JONES [4] describes the difficulties found in manually transcribing the music of the Ewe tribe in West Africa. The British musicologist alludes to the notion of a drumming grammar, which uses this vocabulary of strokes, and in which any alteration of note quality in a particular drum pattern reflects the intention of the drummer and transforms the pattern into a different one. He then suggests that a good transcription of Ewe music should record three facts about every drumbeat [4]: striking hand, position on the drumhead, and whether the remaining hand was also employed for sound production. Similar techniques and interpretations were reported by researchers on the playing of Asian drums. The Indian drums, *mridangam* and *tabla*, for instance, can produce a variety of notes depending on hand position and attack point, as meticulously investigated by RAMAN [161]. The variation in the number of striking fingers and the adjustment of membrane tension during the performance are a part of a musician's technique for some of the Japanese *taiko* drums, which allow the production of many different tones [162].

Evidently, the different "stroke types" must be taken into consideration when producing an automatic transcription for African or Asian drumming or, in fact, for every music tradition or piece in which this "stroke-space" is used to convey meaning. For example, a precise transcription of a faithful rendition of the snare drum in Bartòk's *Cantata Profana* should display the same indications, by means of whichever accorded symbology, as those provided by the composer. This rationale for the transcription of timbral variations is also applicable to other playing techniques, e.g., rudiments (roll, flam, etc.), and to the notation of dynamics [163].

In this chapter we investigate the classification of percussive timbre, with particular interest in the timbral variations produced on the same instrument. First we describe the physics behind sound production in drums and provide more detail on what contributes to differences in timbre. We then briefly review the ADT literature pointing to a few works that study this subject.

6.1 Sound Production in Percussion Instruments

As we mentioned before, the objective of ADT methods is the identification and transcription of percussion instruments in recorded audio. More specifically, most of the literature has as object of analysis a subset of the classes of membranophones and idiophones where sound production is achieved by striking (e.g., with the hands or a drumstick/mallet) the tightly stretched membrane or the instrument body itself. To help us classify and interpret the sounds produced by various kinds of drums as well as the diversity of strokes of a single instrument, we briefly describe in this section how these complex vibrating systems operate and what aspects of their construction and playing may alter the yielded sound quality.

Two properties are required for the generation of mechanical vibrations in a body [162]: stiffness and inertia. These properties guarantee that, when subject to mechanical deformation, the body will: (1) resist it by trying to stay in equilibrium and (2) overshoot its equilibrium position when fighting against this displacement. Thus, the vibration motion is set in place through the work of a restoring force that alternates the system's stored energy between its potential (elastic) and kinetic forms. This exchange is not perpetual and, in real systems, oscillation amplitudes are damped and decay to zero as energy is lost in the form of heat and sound.

The shape of the oscillations can be calculated by solving the system's wave equation and satisfying a certain number of boundary conditions. These solutions can be composed of a superposition of several normal vibration modes, each of which is also a solution to the original wave equation. In distributed mass systems (e.g., strings, membranes) it is customary to graphically represent these oscillations as standing waves, as exemplified in Figure 6.2.

To illustrate this phenomenon, let us first consider as the vibrating body a uniform circular membrane — a model of what is commonly found in many membranophones. An ideal membrane has no stiffness, but can be made elastic by clamping its boundary and maintaining throughout a constant surface tension T, measured in N/m. The wave equation for this membrane is better expressed in polar coordinates in terms of three independent variables — the radius (distance from the center of the membrane), the azimuthal angle, and time; its general solution can then be separated as a product of three functions, one on each of these variables. On the radial direction, the solution has the form of a Bessel function, and on the azimuthal direction, it is cosine-shaped. A full derivation of the solution for the wave equation in a circular membrane can be seen in [162].

If the membrane has a radius of R meters and an area density σ, in kg/m^2, then the fundamental vibration mode $(0, 1)$, in which the entire membrane moves in phase (with maximum displacement at the center and a node at the perimeter), has frequency given by [162]

$$f_{0,1} = \frac{2.405}{2\pi R}\sqrt{\frac{T}{\sigma}}. \tag{6.1}$$

General vibration modes are indicated by tuples (m, n), where m and n are respectively related to the number of nodal diameters and of nodal circles (including one at the boundary). In fact, m governs the order of the Bessel function, $J_m(\cdot)$, and the cosine frequency of the azimuthal component, while n indicates the n-th strictly positive root of the Bessel function that lies on the boundary condition, such that there is no displacement at $r = R$. The oscillation frequencies for all the (m, n) modes can thus be obtained through different iterations of the Bessel function and its roots. Figure 6.1 presents a few of these modes by highlighting nodal lines and showing the alternating displacement patterns that are produced on the membrane's surface. Figure 6.2 shows the $(1, 2)$ mode in a three-dimensional representation, to provide greater detail.

While in one-dimensional systems (e.g., strings, bars) overtone frequencies are ideally harmonically related to the fundamental, i.e., the ratios $f_{0,1}/f_{m,n}$ follow the harmonic series, this is not the case of the ideal circular membrane, as displayed in Figure 6.1. However, multiple factors can affect the oscillation frequency and ratios of normal modes in real drum membranes, including, but not limited to [162]: air loading,[1] bending stiffness, stiffness to shear, and coupling effects with other vibrators (e.g., secondary membrane, drum shell, snares). In some instruments, these effects can cooperate to convey a strong sense of pitch in the produced sound. For example, in kettledrums, where air is trapped between the stretched membrane and

[1] Air loading refers to the coupling between air and the membrane. Its effect can be different whether air is trapped inside the instrument or if the membrane is open to air on both sides.

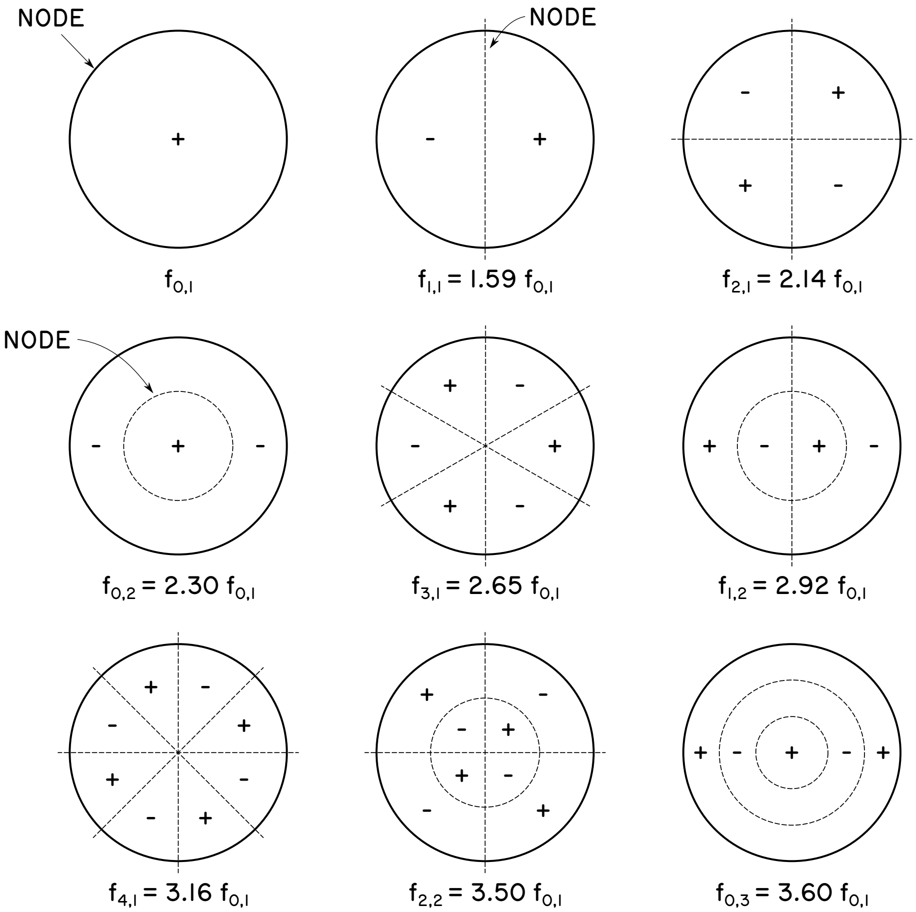

NODE
NODE
$f_{0,1}$
$f_{1,1} = 1.59\ f_{0,1}$
$f_{2,1} = 2.14\ f_{0,1}$
NODE
$f_{0,2} = 2.30\ f_{0,1}$
$f_{3,1} = 2.65\ f_{0,1}$
$f_{1,2} = 2.92\ f_{0,1}$
$f_{4,1} = 3.16\ f_{0,1}$
$f_{2,2} = 3.50\ f_{0,1}$
$f_{0,3} = 3.60\ f_{0,1}$

Figure 6.1: Vibration modes for an ideal and uniform circular membrane. Each mode (m, n) is associated to a frequency $f_{m,n}$, where m is the number of nodal diameters and n is the number of nodal circles, including the boundary [162]. Here we give the first nine modes (in ascending order, from left to right and top to bottom) and their relative frequencies as a (non-integer) multiple of the fundamental $f_{0,1}$, which depends on membrane parameters (applied tension, radius, area density). Positive and negative signals indicate the alternating displacement characteristic of regions between nodal delimiters.

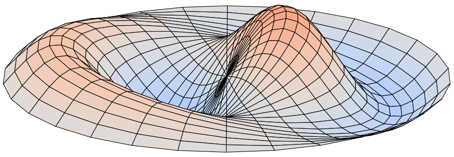


Figure 6.2: Three-dimensional model of vibration mode $(1, 2)$.

a large hemispherical bowl, the effect of air loading is responsible for lowering the frequencies of the first (low-frequency) modes [162]. When kettledrums are properly tuned, their most prominent overtones are then disposed in a quasi-harmonic series starting at the $(1,1)$ mode, i.e., the frequencies of strong sounding overtones present an almost integer relation to this principal mode, an organization which produces a defined pitch associated to $f_{1,1}$. The harmonicity between partials is also a distinguishing feature of the sound of Indian drums (*mridangam*, *tabla*) — here, however, this effect is obtained through the application of a tuning paste on the drumhead. This mix of iron oxide, starch, and gum is loaded on the drumhead in many thin layers and smoothed down in such way that the thickness of the membrane is greater at its center. The net effect of this composition, once dried, is that of changing the shape of the standing waves and allowing the production of five overtones in harmonic sequence, starting at the fundamental [161].

Overtone frequencies play a huge part on the perceived sound, but the duration and relative energy of each partial are just as important in determining the instrument's timbre. In bass drums, for instance, although the frequencies of the $(0,1)$ mode and of the first few nonsymmetric modes may approximate a harmonic relation, the presence of loudly sounding inharmonic partials of higher frequency (above 200 Hz) provides the instrument with a characteristic indefinite pitch [162]. In most instruments, these properties of duration and amplitude can be altered by choices made in the construction phase or during the tuning process, as well as due to the playing action (e.g., location and intensity of the striking force). Damping times depend on factors like the materials used in the drumhead and in the shell, and the tension of the drum membranes. If the drumhead is not muffled by the musician's hand or mallet (and all vibrations on the membrane are suddenly interrupted), we can verify that, in a free-sounding drum, different overtones decay at different rates. In general, low-order modes are more pronounced and decay more rapidly. Higher modes, on the other hand, present a smaller decay rate, which is mostly attributed to the alternating phase patterns that form on the membrane (see Figure 6.1) and reduce the overall energy radiation efficiency [164]. Moreover, a strike at the center of the membrane can initially transfer a lot of energy to inharmonic axisymmetric modes (e.g., $(0,1)$, $(0,2)$, etc.), whereas off-centered attacks excite other modes. For example, an excitation of the kettledrum at the normal off-center position[2] favors the harmonic modes, $(1,1)$, $(2,1)$, $(3,1)$, etc. Studies have shown that nonlinear coupling processes that transfer energy between different sounding modes can occur [162]. For all these reasons, the amplitude and energy relations of the fundamental and the overtones is particularly complex to model.

Similar considerations can be applied to certain idiophones, particularly those

[2] About one-fourth of the distance from the rim to the center of the membrane [162].

made out of circular plates (e.g., cymbals, gongs). Due to the variety of shapes of the main vibrator (i.e., the plate), we will not further describe sound production in these instruments. The interested reader is referred again to [162] and references cited therein for a more detailed description of idiophones.

In summary, we observe that the diversity of percussive excitations in both membranophones and idiophones can be generally characterized by three components, according to Bell's comprehensive drum timbre lexicon [165]:

1. Mode, which refers to "simple" (e.g., single strike, shaking, scraping, rubbing, plucking) or "complex" (e.g., flam and roll rudiments, *sforzando* dynamics) articulations;

2. Implements, i.e., the type of beater that is used to excite the instrument (e.g., drumstick, brush, mallets, hands);

3. Region, i.e., the placement of the excitation (e.g., edge, center).

These components and their complex interactions with the form and materials of the different instruments produce the entire gamut of sounds in the percussion family.

6.2 Approaches to Drum Sound Classification

In this section, we review a few approaches to drum sound classification that are found in the ADT literature. While the majority of works in this subject deal with drum instrument classification, i.e., *which* drum is being played, we here provide a more in-depth look at theses and papers that focus on the analysis of timbral varieties and playing techniques, i.e., *how* a certain drum is being played. Nevertheless, we also consider works of the first type, which can contribute with great insight into the choice of features, for example. Similar reviews can be seen in [166–170].

Since this research subject has seen little exploration, studies differ in many respects such as the instruments and timbres under analysis (e.g., drum kit, instruments pertaining to certain music traditions), the feature extraction process (i.e., how to represent the signal), the classification technique (unsupervised or supervised) and algorithm (e.g., k-means, KNN, SVM, neural networks), among others. Some works evaluate their systems on datasets of individual drum sound events, whereas in others the audio from a real drum sequence recording has to first be segmented (e.g., by means of an onset detection scheme) for then the segments can be classified. In this latter case, musicological knowledge (i.e., language models) can be embedded in the system to improve the classification accuracy by exploiting context clues such as neighboring notes or periodic measure-length patterns. Other difficulties arise from the ensemble characteristics of recorded samples, that is, if they contain "monophonic" or "polyphonic" percussion, and if other non-percussive instruments are present.

Schloss, Bilmes, and *Conga* Strokes

In a seminal work presented by SCHLOSS [93], a system for the transcription of percussive music was showcased with recordings of *conga* drum performances. Schloss discusses several of the major transcription tasks, from track segmentation and stroke recognition to tempo tracking and high-level rhythm analysis. The recordings in his dataset contained a sequence of non-simultaneously occurring strokes on a pair of (high/low) *conga* drums that he manually classified into four basic categories [93]:

- OPEN — "Hand snaps away from drumhead, allowing maximum ringing of drumhead in normal mode";
- MUFF — "Hand 'sticks' to drumhead, damping and also raising pitch of tone by about a minor third";
- BASS — "Palm of hand hits center of drumhead, causing lowest perceived pitch";
- SLAP — "Hand hits center of drumhead while damping edge, causing sharp attack and higher perceived pitch";

These gave rise to eight total classes considering both drums. His "low-level analysis" included the steps of detection/segmentation and classification, and can be summarized as follows. First, an envelope was obtained from the waveform through a min-max sliding window of length $T_0 = 1/f_0$ seconds, where f_0 represents the lowest frequency expected in the data. The system computed a series of slopes, $\{S_n\}$, from this envelope via linear regression of every set of n successive points. Then, note attacks were inferred from large abrupt variations in the slope series, at which point the system was also informed by the local energy maxima of a high-pass filtered version of the signal. To help reducing the detection of false positives, Schloss assumed a "forbidden attack region" after each detected attack, meaning that no other attacks are to be found at the immediately following instants. The segmentation step ensued where each segment corresponded to the time interval between two successive attacks. Rests were also inferred during this process as the regions where local amplitudes did not surpass the power of the signal filtered with a moving average filter. For each segment containing a note, the system determined whether the stroke was damped or undamped by modeling the exponential decay with a one-pole fit and applying a heuristic decision threshold to the estimated decay constant. Schloss also used this information to select from which portion of the signal to extract spectral features for classification. If the note was undamped (i.e., decay rate is large), he considered a “steady state” portion of the waveform around the middle part of the segment (100 to 200 ms). Instead, if the note was damped (i.e., exhibited a small decay rate), he analyzed the entire segment. Spectra from the selected portions of each segment were non-uniformly partitioned into three empirically determined bins (0 to 1 kHz; 1 to 7 kHz; 7 kHz to the Nyquist rate) and the energy distributions over these bins were used as the main classification feature. The peak spectral frequency, assumed to be the fundamental in undamped strokes, was also recorded. Next, each stroke was matched to the nearest centroid of a set of pre-recorded templates in the feature space, and distances were normalized by the standard deviation of each template class. The system output a “notelist”, which contained onset time, duration, amplitude, and stroke type of every detected event, along with a confidence measure for each stroke, which was used as input for subsequent higher-level analyses. Schloss reports good accuracy rates and few undetected notes (false negatives).

BILMES [171] also examined *conga* recordings in his investigation of expressive timing in percussive rhythm. Instead of a single musician playing a pair of drums, he recorded an ensemble (Los Muñequitos de Matanzas) composed of three drums of different sizes (*tumba*, *conga*, and *quinto*), *guagua* (a bamboo slit drum), *claves* (a pair of wooden cylinders played by concussion), and singers. Microphones and musicians were carefully positioned during the recording process to minimize bleeding between stems. To investigate the conga strokes, Bilmes segmented individual tracks, then classified the set of segments using an unsupervised approach. At the segmentation step, a pair of linear phase FIR low- and high-pass filters were used to analyze the audio input, and short-time energies of both filtered signals were computed with a small window. Linear regression was used to calculate slopes from each energy envelope, providing candidate attack times, similar to Schloss. Attack points were defined by combining information from both slope series. Then, each stroke segment was determined from its attack time to the point where local energy got below a fraction of the corresponding maximum. If, for any segment, this fraction was never reached, its endpoint was set to a few seconds before the attack time of the next stroke. For clusterization, the system computed the following set of features from each segment: an approximation of the CQT of the segment obtained by averaging DFT bins into one-third octave bands; the length-normalized energy of the stroke segment; and an exponential fit of the envelope decay (as in [93]). These formed a feature vector of length 28 representing the stroke space. Features were normalized to zero mean and unit variance across all vectors, and the Karhunen-Loève transform was used to reduce dimensionality through the selection of the most significant eigenvectors. The processed feature vectors were finally clustered with the k-means algorithm, where the number of clusters k was subjectively adjusted. Despite being able to recognize by ear from six to ten different strokes (depending on the drum), Bilmes found that $k = 4$ or $k = 5$ provided the best accuracy due to large class imbalance between different stroke types. He also points out that oftentimes it is difficult for a listener to identify a stroke being played in isolation; the task becomes easier and rather immediate once proper context is given (e.g., tatum number, position in phrase).

Studies on Features and Classifiers for Drum Instrument Classification

HERRERA et al. [101] conducted a large-scale evaluation of drum sound classification techniques involving several instruments. The investigated dataset was composed of 634 individual sound samples of drum kit instruments: bass drums, snare drums, tom-tom drums, hi-hats, and cymbals. During the process of assembling the dataset, the researchers made a conscious effort to ensure that the samples varied in dynamics and that, for each class, recordings of different instruments were available. However, they intentionally excluded playing techniques that resulted in significant timbral deviations from the expected "standard sounds". Examples of these include strokes using brushes or rim shots. For each sample, the attack–decay boundary was determined from the signal amplitude. Then, to describe each drum sample, they selected 48 different features, grouped into four categories:

- attack-related features — attack energy, temporal centroid (of the entire signal), logarithm of attack duration, attack zero-crossing rate, and temporal centroid to attack duration ratio;
- decay-related features — spectral flatness, spectral centroid (and variance), strong peak,[3] spectral kurtosis, spectral skewness, zero-crossing rate (and variance), and strong decay[4] — all of which were computed at the decay portion of the signal;
- relative energy features — distribution of energy over eight empirically chosen bands (40 to 70 Hz, 70 to 110 Hz, 130 to 145 Hz, 160 to 190 Hz, 300 to 400 Hz, 5 to 7 kHz, 7 to 10 kHz, and 10 to 15 kHz);
- time average of thirteen mel-frequency cepstral coefficients (MFCC) and respective variances.

[3] The "strong peak" feature describes how pronounced is the peak of the spectrum, achieving high values the greater and thinner the main peak is.

[4] The "strong decay" feature describes the position of the temporal centroid with respect to the decay portion of the signal, and is high if the temporal centroid is close to its beginning.

The classification was divided into three levels, with increasing degrees of class detailing. In “super-category”, instrument sounds were coarsely classified into “membranes” or “plates”. At the “basic-level”, we find the proper instruments classes (as described before). Finally, in “sub-category”, certain instrument classes were finely subdivided: low/medium/high toms; open/closed hi-hats; ride/crash cymbals. Techniques for classification were chosen from tree families: instance-based algorithms (k-nearest neighbors — KNN, K*), statistical modelling (canonical discriminant analysis), and decision trees (C4.5, PART); which were combined with two feature selection schemes (correlation-based feature selection and ReliefF). References for all classification and feature selection techniques can be found in the original paper [101]. The authors performed preliminary tests to select sets with fewer but more relevant features, such as: zero-crossing rates, spectral skewness and kurtosis, temporal centroid, low-order MFCCs, relative energy in few bands. They report high accuracies in the three classification levels with performance degrading a bit the more specific the class (respectively 99%, 97%, 90%). It was also shown that the subset of relevant features provided best classification overall.

In a follow-up paper, HERRERA et al. [124] expanded the number of instrument classes (*bongo*, clap, *clave*, *conga*, cowbell, side-stick,[5] *tabla*, tambourine, timbale, triangle) adding up to 1976 samples of acoustic and synthetic individual drum sounds. They included new Bark-based descriptors and used a couple of other classification algorithms (e.g., kernel density estimation) to evaluate the “basic-level” classification task over 200 different feature subsets, obtaining at most 85.7% accuracy with a set of 40 features. In another interesting experiment, authors presented a classifier that was able to correctly identify manufacturers using a distinct subset of features for each of four instrument classes (bass drums, snare drums, tom-tom drums, and hi-hats); results ranged from 86.3% to 99% accuracy. In both studies [101, 124], experiments were conducted with ten-fold cross-validation.

Another large-scale study by VAN STEELANT et al. [99] identified bass drum and snare drum sounds in loops using Support Vector Machines (SVM) with linear and Gaussian kernels. Tracks were programmatically generated by combining 2028 samples of acoustic and electronic percussion instruments from five different classes (bass drums, snare drums, hi-hats, cymbals, and tom-tom drums). Each track consisted of a loop of drum sounds in quantized positions, and the superposition of different instruments was allowed. Bass drum and snare drum samples were never simultaneously present in any mix. Since the timing information of each drum loop was known a priori, authors bypassed the problem of onset detection and instead

[5] Side-stick (or cross-stick) is a drumming technique usually executed on the snare drum where the drummer rests the end of the drum stick over the membrane and hits the rim with the tip of the stick.

extracted features from a 70-ms region at the beginning of each attack. Each stroke was then represented by an 89-dimensional feature vector, which included:

- energy-related features — root mean square (RMS) of the entire segment and in three frequency bands, in absolute and relative values;
- temporal features — ZCR, crest factor (peak amplitude to RMS ratio), and temporal centroid;
- spectral features — spectral centroid, skewness, kurtosis, and roll-off;[6]
- MFCC — mean and variance of twelve coefficients, and of their first and second derivatives;

and features were normalized to the range $[-1, 1]$. They conducted different experiments with a reduced set of 14 features and with the complete set; data were split between train and test sets (87.5%/12.5%) and seven-fold cross-validation was employed to select models. Good accuracy, precision, and recall values show that the SVM classifiers were able to generalize and correctly indicate the presence of bass/snare drum even when "noise" (non-percussive MIDI accompaniment) was added to the mixtures.

Playing Techniques on Drums of Indefinite Pitch

TINDALE et al. [172] extracted a set of temporal and spectral features to investigate subtle timbral variations on snare drum recordings. A dataset was prepared with recordings of three expert players on different drums. Players were instructed to excite the drum using different implements and placements, in order to generate seven different timbres: rim shot, brush stroke, and regular strikes at center, near-center, halfway, near-edge, and edge positions. In total, 1260 strokes were recorded at CD quality (16-bit, 44.1 kHz). The set of features was computed on four different windows lengths: attack (from onset to peak), and 512, 1024, and 2048 samples from the onset. Playing technique classification was conducted with feed-forward neural networks, KNN, and SVM, using ten-fold cross-validation and evaluated on four different scopes: all strokes; five placements on the membrane; rim shot, brush, and edge; center, halfway, and edge. The KNN classifier was reported as the most consistent, with accuracies ranging from 89.6% to 99.3% on all tasks and windows.

An expansion of Tindale's work was presented by PROCKUP et al. [18]. Authors investigated timbres not only in snare drum samples, but also from strokes on two different tom-tom drums. They considered different stick heights (from 8 to

[6] The spectral roll-off is usually computed as the frequency point bellow which a certain predefined percentage (e.g., 85%) of the energy of the spectrum is contained. Steelant et al. [99] defined it with the magnitude of the spectrum instead.

32 cm) and dynamics (light, medium, heavy), as well as three strike positions (center, halfway, edge) and types (regular strike, rim shot, buzz roll, cross-stick). With at least four examples of each parameter combination, the entire dataset consisted of 1804 individual strokes, which were used for training and testing the articulation recognition model. They used a set of "compact" features to represent signals in the timbre space. For that, first they extracted the evolution of single dimensional features over time; these included RMS energy, roughness,[7] brightness, and energy ratios between four spectral bands (0 to 1000 Hz, 0 to 534 Hz, 534 to 1805 Hz, and 1805 Hz to the Nyquist rate) and the entire spectrum. Then, they fitted third- and sixth-degree polynomials to these temporal sequences, and used the fitting coefficients as their final representation. They conducted two experiments on the features extracted from the dataset using SVMs with RBF kernels and five-fold cross-validation. In the first experiment, the authors evaluated the classification accuracy of each feature and of combinations of (normalized) features on the subsets of each individual drum. Authors also classified samples using MFCCs and their first and second derivatives. MFCCs topped the accuracy rates for individual features and drums (except the floor tom), but were outperformed by all the investigated combination of features, especially that of MFCCs and brightness evolution (sixth-degree fit coefficients). Next, in order to test the generalization potential of the selected features, they examined the classification task on the entire set of audio samples. Results on the second experiment were analogous to those of the first one, with the highest accuracy rate of 97.2% for the combination of brightness and MFCCs, showing that the system was able to generalize the stroke type for the distinct drums.

SOUZA et al. [173] presented a study of sounds produced by five different cymbals (china, crash, hi-hat, ride, and splash) of various materials, builds, and sizes. In their dataset of just over a thousand samples, the main cymbal types were divided into subclasses based on the playing techniques used to produce them. These included "open", "closed", and "chick" (for hi-hats); "edge", "body", and "bell" (strike placement); "roll"; and "choke" (for when the cymbal was muffled after being struck). Five sets of features were extracted from the data — temporal, spectral, line spectral frequencies (LSF), linear-frequency cepstrum, MFCCs — and classification was performed using naive Bayes, random forest, C4.5, KNN, and SVM models estimated with ten-fold cross-validation. They conducted one experiment for the classification of cymbal types and another one for playing techniques. In both cases, the combination of SVM and LSF features yielded the best results, with accuracies of 96.6% at the cymbal type recognition task and 86.5% for technique classification.

[7] Roughness is a measure of the sensory dissonance elicited by a sound, and is usually computed considering fast modulations that occur between peaks that lie close together in the spectrum [120].

Performance on the latter was shown to improve by 5% when the LSF descriptor was combined with features from the temporal domain.

WU and LERCH [102] reported a different approach for investigating snare drum playing techniques (strike, buzz roll, flam, and drag[8]) in polyphonic recordings. Their method performed classification on features computed over the activations at the output of a non-negative matrix factorization algorithm (NMF). The objective of the NMF algorithm is to decompose a non-negative matrix $\mathbf{V}_{N\times M}$ (e.g., a spectrogram) into a product of two matrices $\mathbf{W}_{N\times R}$ and $\mathbf{H}_{R\times M}$, which are also constrained to be non-negative, such that an approximation $\hat{\mathbf{V}} = \mathbf{WH}$ can be obtained numerically. Matrices $\mathbf{W}$ and $\mathbf{H}$ are respectively called basis (or template) and activation matrices, and R determines the number of distinct template–activation pairs into which $\mathbf{V}$ is to be decomposed. Wu and Lerch used a partially-fixed NMF (PFNMF), in which the basis matrix $\mathbf{W}$ was split into two parts. One part contained a fixed dictionary of previously learned drum templates (for hi-hats, snare and bass drums) and the other was a randomly initialized template that was used to learn harmonic components from the mixture. In their system, authors first computed a set of template activations from examples of the different articulations using a regular NMF and solo recordings, and then analyzed the polyphonic recordings with the PFNMF. Once activations for the snare drum were extracted from the mixed samples, normalized and smoothed with median filtering, they segmented a 400-ms window around each annotated onset. Segments were shifted to guarantee that peak values were always centered and, for each segment, the following features were computed:

- distribution features — spread, skewness, crest, centroid, and flatness;
- inter-onset interval (IOI) features;
- peak features — indices and ratios between side peaks (in descending order) and the main peak;
- dynamic time warping (DTW) features — the cumulative cost of the DTW distance between the segment and the four template activations.

These features were put together into a 19-dimensional feature vector. A feature vector of traditional timbre features (spectral centroid, ZCR, MFCC, etc.) was also computed for comparison. A multi-class SVM classifier was trained on a solo training dataset assembled by the authors, which contained strike and buzz roll excerpts from the dataset presented in [18], and synthetically-generated flam and drag excerpts. This classifier was then used in two experiments with real-world recordings from the ENST-Drums dataset [66]: one with informed onsets and the other without

[8] Flam and drag are snare drum playing techniques commonly used in military bands where the primary stroke is preceded by one (flam) or two (drag) grace notes.

annotations. The classifier with the common timbre features achieved the highest accuracy on the cross-validation for the training set, but performed poorly on the remaining data. The classifier with features derived from activations performed well on the training set (above 90% accuracy) and moderately on the real-world recordings with and without annotations (76.0% and 64.6% on average, respectively) when no background music was present. When music was added, performance dropped to 40.4% accuracy, which was likely due to the noisy activation function that was extracted from the polyphonic mixture in this case.

CHESHIRE et al. [174] carried out not a classification task, but an A/B listening test to verify if participants could distinguish between different dynamics on the snare drum when the loudness of all sound samples was normalized. With that objective, they recorded samples from a single snare drum in a professional setting using four (two condenser and two dynamic) microphones in close placement to minimise room reverberation effects. Sound pressure level (SPL) measurements were performed during the recordings and helped set a target of high and low velocity strikes for the drummer. Then, for each microphone, one example track of each velocity was created containing a two-bar drum pattern. During the test, participants were subjected to these phrases ten times for each microphone in a random order, and were asked, half of the time, in which phrase strikes had a perceived lower velocity, and, in the other half, in which phrase strikes had higher velocity. The 15 listeners had between 21 and 50 years, and 3 to 30 years of experience in the audio field. Tests showed that the participants were able to successfully distinguish between the two dynamics. Authors also conducted an objective test with a second set of recordings and several of the already discussed temporal and spectral features. Statistical testing revealed a significant difference between the two types of strokes.

Playing Techniques on Brazilian Instruments

Although most of the works in the ADT literature focus on timbral variations produced in drums commonly found in a drum kit, instruments from other traditions, such as some Brazilian percussion instruments, have received moderate attention as well. For instance, ROY et al. [175] explored the subtleties in the strokes produced by a *pandeiro virtuoso* using the "Extractor Discovery System". Using genetic programming, this system automatically selected and combined elementary operators into "analytical features", which were evaluated according to a fitness metric — a supervised classification with SVM, in this case. Examples of operators included, but were not limited to: FFT, log-compression, Hann window, temporal and spectral centroids, MFCCs, etc. The combination of operators was controlled by the type of inputs and outputs they require, and genetic transformations were applied at each generation of the genetic algorithm to guarantee variability and a non-decreasing fitness. The authors constructed a dataset of 2448 *pandeiro* sounds to test the system. This dataset contained a balanced number of examples from six categories [175]:

- Tung — a bass sound;
- Ting — a higher pitched bass sound;
- PA — a slap sound (similar to the *conga* slap);
- pa — a softer slap, hitting the drumhead on the center;
- Tchi — the sound of the jingles;
- Tr — a *tremolo* of the jingles;

Each signal was windowed into non-overlapping frames of 1.4 ms and tests were conducted on features extracted from three regions: pre-attack (the frame corresponding to the detected onset and its predecessor), post-attack (the onset frame and its successor) and full sound. The results were shown to outperform those of a reference feature set.

In his thesis, DA COSTA [73] proposed method for an unsupervised classification of strokes on solo tracks from the BRID dataset described in Section 3.1.3. He first listened to examples of files from the ten instrument classes, determining the expected number of stroke types for each instrument. He downsampled all signals to a sampling rate of 11 025 Hz and computed spectrograms with $N = 256$ samples (50% overlap). This rather coarse frequency resolution was deliberately chosen to reduce the variability of examples from the same stroke type. Then, informed by the onset annotations, he segmented the spectrograms starting 5 ms before each onset and ending up to 200 ms after it, depending on the instrument. Segments were weighted with a half Hamming window, giving more weight to earlier samples closer to the attack. Finally, the weighted spectrogram segments were used as features and clustered with k-means. Classes in different files from the same instrument were not directly comparable, but Da Costa estimated an overall accuracy of 75% to 80%.

Playing Techniques on Indian Drums

Playing techniques were also investigated in the case of the Indian drums. As we mentioned in Section 6.1, these drums, namely the *mridangam* and the *tabla*, present strong harmonicity between partials and elicit a clear sense of pitch. GILLET and RICHARD [176] describe a transcription system of isolated[9] *tabla* strokes using

[9] *Tabla* strokes are usually treated as monophonic even though, in many situations, both drums (*dayan* and *bayan*) are played simultaneously and articulate different *bols* or, in other cases, their

language models for improved performance. Authors noticed that, in *tabla* performances, the same stroke can sometimes be identified by different *bols* (i.e., symbol), depending on the context,[10] hence the importance of informing the classification with local clues. In their system, audio signals from three solo recordings were first segmented using the time derivative of a normalized version of the amplitude envelope, to compensate for local energy fluctuations. These recordings differed in many aspects, including the quality of the instruments and their tuning (in C♯3 and D3). In total, they obtained 64 drum phrases which were segmented into 5715 strokes. Authors represented the power spectrum of each stroke as a mixture of $N = 4$ Gaussian distributions and used a set of twelve elements (mean, variance, and relative weight of each distribution) as the feature vector. Classification was performed with a hidden Markov model (HMM) that, by modelling transition and emission probabilities, took advantage of the local dependencies in *tabla* drum sequences, and yielded high accuracy (6.5% error rate) over the entire test set. Another experiment was conducted in which training and test sets corresponded to different instruments and recording conditions. In this situation, performance degraded, but remained around 90%. Results were compared to simpler classifiers — KNN, naive Bayes estimator, kernel Density estimator —, which were outperformed.

This work on the recognition of *tabla* strokes was later greatly expanded by CHORDIA [103]. In this paper, the author experimented with the dataset and annotations provided by GILLET and RICHARD [176], but also added manually segmented and annotated 11 119 strokes from original studio recordings. Instead of modelling each sample with its power spectrum density, the author extracted 31 features including temporal centroid, attack time, ZCR, spectral centroid, spectral skewness, spectral kurtosis, and 13 MFCCs. The dimensionality of this feature set was reduced with the principal component analysis (PCA). Four different classifiers were used in this work — multivariate Gaussian classifier, feed-forward neural network, probabilistic neural network, and tree classifier — and language modelling with HMMs was also attempted to achieve a fair comparison with the previous work. As in [176], Chordia conducted two experiments by training and testing on the same and on different subsets, which he names "cross-validation" and "novel generalization" tests respectively. The author reports that, on Gillet and Richard's dataset, there was a significant performance gain (about 10%) in the cross-validation test when, in both works, no language modelling is employed. In fact, his results for classifiers without context clues match those of [176] with HMMs, which the author attributes to the use of more sophisticated features. A similar trend was reported

sounds overlap due to ringing. Nevertheless, the sequence of strokes is always regarded as a sequence of individual — simple or compound — *bols*. A similar thing happens in the case of the *mridangam*, where strokes at both drumheads are represented by a single syllable.

[10] Also, sometimes the same *bol* can refer to different timbres [103].

for this dataset in the novel generalization test. Chordia also shows that, across all datasets, no performance gain was achieved in his system by using language models, and, instead, performance was degraded in some cases. The author explains that this is likely due to poor data quality. Overall, neural networks displayed the highest accuracies in both tests and, for all classifiers, highest rates were obtained when the recording conditions of train/test sets were the same.

More recently, ROHIT et al. [177] separated *tabla* strokes into four major categories according to which drum was hit and whether the stroke was damped or resonant. They named these archetypes "damped", "resonant treble", "resonant bass", and "resonant both", noting that "damped" refer to strokes on either drum that do have a fast decay and that, for resonant strokes of a single drum (bass or treble), the remaining drum was silent or articulated a damped sound. Authors recognized the similarities between these classes and three of the mostly studied instruments in ADT: the bass and snare drums, and the hi-hats. They traced a correspondence between "resonant bass" and the bass drum, "resonant treble" and the snare drum, and "damped" and the hi-hat. "Resonant both" was expressed as a simultaneous activation of "resonant bass" and "resonant treble". This mapping allowed authors to adapt multi- and single-class CNN-based models derived for the transcription of the Western drumkit to the task of *tabla* category recognition. They experimented with training the models from zero and applying some transfer learning strategies, as well as implementing some data augmentation (pitch-shifting, time-scaling, attack-remixing, and *tabla*-specific) methods. As a baseline, they used a random forest classifier with a large set of acoustic features, including spectral moments, MFCCs, spectral energy and flux in bass (50 to 200 Hz) and treble (200 to 2000 Hz) regions, log attack time, temporal centroid, ZCR. Results show that the classifier based on neural networks outperformed that based on random forest with regular features in a large dataset of 26 000 strokes, with the single-class model achieving on average 70.4% F-measure and outperforming the other configurations in the recognition of each class, except in the case of the scarcest target class, which was best handled by the fine-tuned multiclass model.

Finally, we highlight works devoted to the similar task of classifying *mridangam* strokes (also called *aksharas*). In [178] and [179], isolated stroke samples were analyzed with the NMF and classified with respect to the resulting activations. ANANTAPADMANABHAN et al. [178] recorded two solo performances on the *mridangam*, one tuned to D♯ and the other to E. Each set of recordings was separated into common phrases (134 in D♯ and 114 in E) containing a total of 1170 strokes. They also recorded the first five individual modes of each instrument,[11] following

[11] In the aforementioned study on the Indian musical drums [161], Raman describes in detail how the modes of a harmonic drum like the *mridangam* can be individually excited through the careful

a procedure described by RAMAN [161]. With the spectrogram representation of each mode, they ran the NMF algorithm for $R = 1$ until convergence, which yielded a dictionary of basis functions representing each of the modes. Then, they projected the recorded phrases onto the basis dictionary and computed activations for template modes in these sequences. Since each stroke might be composed of more than one mode, authors aggregated the activations to determine onsets for each stroke. Next, they segmented the strokes at the located onsets, keeping only 80% of the frames in-between successive onsets to avoid interference due to ringing. Ten classes of strokes were individually modelled (cf. [180, VI-B]) with distinct HMMs and four-fold cross-validation. They experimented with the classification of only simple strokes and of all (simple and compound) strokes, achieving higher accuracy in the former (82.3% and 84.7% for D♯ and E, respectively, with four fundamental modes; and 78.3% and 88.4% with five modes). In the latter, i.e., for the entire set of strokes, accuracy rates were 72.6% and 75.0% for D♯ and E, respectively, with four fundamental modes, and 73.2% and 87.4% with five modes. Authors reported larger class confusion in compound *aksharas* formed of similar simple strokes.

In the previous work, experiments were performed separately for each tuning; in a subsequent paper, ANANTAPADMANABHAN et al. [179] attempted to transcribe *mridangam* strokes independently of the performance's tonic. They evaluated their algorithm in a dataset containing a total of 7170 *mridangam* strokes in six different tunings (from B to E in steps of a semitone); for the D♯ and E tonics authors used the same recordings as the previous work, whereas new recordings were made for the other tonics using a tunable instrument. Overall, ten different strokes were featured in the dataset. The CQT was selected to represent the audio in the time-frequency domain since transpositions of the same stroke in different tunings were expected to display the same shape in the CQT magnitude, differing only by a linear shift on the frequency axis. Authors then applied the DFT over the CQT frequency bins, and kept only the magnitude of the obtained representation. This procedure resembles the method for the extraction of MFCCs, in which the DCT is applied over the log-power spectrum of a mel-frequency representation of the signal. As in [178], each recording was projected onto a dictionary of NMF bases. Here, however, this dictionary was learned directly from five different solo recordings (in the C♯, D, G, and G♯ tunings, respectively) which were not included in the dataset,

placement of fingers over the membrane followed by simple percussion. Similarly to what is done in the case of string instruments, fingers that rest on the membrane are used for the creation of nodes (i.e., nodal lines) in the mode that one desires to excite. For example, to excite the second harmonic, the drummer can gently touch the membrane with one finger in the direction of a nodal diameter and then strike the membrane with a finger of the opposing hand in the perpendicular direction. Raman also discusses how the third, fourth, and fifth harmonics can be produced by either one of multiple corresponding modes or by a superposition of two or more individual modes in any ratio.

instead of from samples of the instrument modes. Another difference consisted on the extraction of onset times, which was carried through by picking peaks of a spectral flux function. After the factorization, NMF activations were summarized in-between detected attacks by computing the relative energy, standard deviation, maximum and minimum values over frames, producing a $4R$-dimensional feature vector for each stroke. Vectors in the feature space were then classified with SVM (radial basis function kernels) in ten-fold cross-validation. Two main experiments were reported. In the first one, they compared the results on the D♯ and E subsets with the previous work, for simple and compound strokes. They show recognition improvements of up to 10.3% (from 72.6% to 82.9%) in the best case when train and test sets come from the same tuning, and of 4.2% (66.2% to 70.4%) otherwise. In the second experiment, each of the six recordings was considered as a fold and used for testing the system trained with the remaining five. Accuracies were as high as 75.6% and 72.6% for C and C♯ tunings, and as low as 56.7% and 57.1% for D♯ and E. Authors pointed out that best results were obtained for those tracks that had similar recordings conditions (B to D), particularly those with more neighboring tonics one semitone apart (C and C♯). We should note that in both tests authors used a CQT with twelve bins per octave over six octaves (starting at $f_0 = 70$ Hz) and $R = 20$ bases in the factorization step.

Finally, KURIAKOSE et al. [181] modeled up to 41 different *aksharas* in seven different recordings (and five distinct tonics) using two different approaches. For the first approach, they used a group delay novelty curve to find onsets and segment performances as a series of stroke events. Thirteen MFCCs (including the energy) were computed over 20-ms frames (with 90% overlap) for each segment, along with their first and second derivatives, forming a 39-dimensional feature vector. Then, following a similar approach to that of [178], they built different (three-state, single mixture) HMMs to model each individual stroke. Lastly, an HMM language model corrected the stroke transcription. In the second method, onsets were not detected beforehand and MFCCs were extracted over a larger 100-ms window. Generally, they report that the method for isolated stroke recognition performed better, especially when language modelling was used, for train/test sets on the same tracks. Accuracies in the best case were of 75.7% for long concert recordings and 95.8% for studio recordings. In a test with only the studio-recorded phrases, in four different tonics, authors evaluated the transcription with different observation representations. Results showed that replacing MFCCs with the CQT-based features of [179] or with cent filter-bank cepstral coefficients (CFCC) yielded greater accuracy when training and testing on all tonics (from 60% to 74% and 77%, respectively), and when training on one tonic and testing on the remaining three (from 50% to 62% and 66%).

Chapter 7

Investigation of Drum Sound Classification

In this chapter we apply some of the knowledge reviewed on Chapters 5 and 6 to the classification of drum strokes from BRID solo tracks. As we mentioned in the dataset definition (Section 3.1), this dataset encapsulates some of the main rhythmic patterns in *samba* and related genres. This poses challenges that were not faced by many of the works discussed on the previous chapter that dealt with individually recorded drum strokes both in training and testing [18, 124, 172, 173, 175] or, at least, only during training [102]. First, when investigating performance recordings, besides all acoustic properties (e.g., reverberation) that might affect the classification accuracy and which most systems that operate on isolated samples also have to face, we must take into consideration that information pertaining to one note might "leak" into the succeeding event due to mechanical characteristics of sound production. These effects are unavoidable in segment–and–classify approaches. Second, and most important, we are not looking into classifying sound events produced by different drums (e.g., in a drum kit) as most of the earlier works; instead, we look into distinguishing the different playing techniques used for sound production in a single drum, such as exemplified by [18, 103, 172, 173, 175–179, 181]. Finally, due to the richness of instrument variations in the dataset, the different recorded articulations types identified with a single instrument class may present significant differences in timbre.

In this chapter, we use an SVM-based supervised classification scheme, following previous works that had similar objectives [18, 99, 102, 172, 173, 175, 179]. The SVM leverages a diverse ensemble of features to recognize drum playing techniques from samples that were first detected and segmented from audio streams. The system has to be robust enough to detect drum events of very different characteristics (e.g., energy profile), but also to correctly classify more subtle changes in playing technique that arise in the sound production of the same instrument (see Figure 7.1).

We describe in the next sections the samples that were annotated for our experiment, as well as our approach to onset detection and segmentation. Finally, we describe the features and procedures for classifying drum sounds.

7.1 Subset Definition

We conduct our investigation with a set of samples from the BRID solos representing two instruments: *tantã* and *repique de mão.* These instruments are commonly performed together at *rodas de samba* and *pagodes*, usually with interwoven rhythmic patterns. Both instruments are played while placed over one or two legs, with one hand supporting the instrument by its shell and the other free hand striking the membrane. Moreover, they present an interesting and manageable number of different articulations. For *tantã*, we highlight the following categories [182]:

- FINGERS — the lowest-pitched sound, produced by slapping the membrane near the rim with the fingers.
- HAND — produced with an open hand on the skin near the center.
- SHELL — produced when the hand that supports the instrument strikes the drum shell.

Repique de mão, a single-membrane adaptation of *repinique*, has these main articulations:

- THUMB — the lowest-pitched sound, produced by the thumb close to the edge of the membrane.
- FINGERS — produced by the fingers at the center of the drumhead.
- SHELL — supporting hand striking the drum shell.

BOLÃO [182] also reports another articulation for *repique*: “RIM”, which is produced by the fingers striking the rim of the instrument. However, these are not featured in our dataset or, at least, could not be recognized afterward during the annotation procedure. We note that the articulations described above can be grouped into certain “archetypes” with respect to their musical meaning and the characteristics of the sound they produce, notably: an open tone/low-pitched sound (FINGERS in *tantã*, THUMB in *repique de mão*), a closed/muffled sound (HAND/FINGERS), and a sound produced on the drum shell (SHELL/SHELL) that is used to fill in the rhythm.

In total, there are 23 recordings of *tantã* and 12 of *repique de mão* in different styles (*samba*, *partido-alto*, *samba-enredo*, *marcha*, and corresponding *viradas*). In

Table 7.1: *Tantã* articulations in the subset.

	Variation				
Category	TT1	TT2	TT3	TT4	Total
FINGERS	282	275	238	397	1192
HAND	309	172	206	384	1071
SHELL	419	320	350	599	1688
Total	1010	767	794	1380	**3951**

Table 7.2: *Repique de mão* articulations in the subset.

Category	RP1
THUMB	876
FINGERS	620
SHELL	917
Total	**2413**

particular, for *tantã*, four different variations (variations 1 to 4), which are characterized by different drumhead sizes and materials, are included in the subset. Recordings vary also according to the performer: musician #1 plays variation 2; musician #2 plays variations 3 and 4; and musician #3 plays variations 1 and 4. For *repique de mão*, only a single instrument was available — the other *repiques* in BRID are *repiniques* and *repiques de anel* (see Table 3.1). Tables 7.1 and 7.2 show the number and types of articulations for *tantã* and *repique de mão*, respectively.

Figure 7.1 displays an excerpt of a *tantã* recording in which the three articulations are present. We can easily verify that the articulations are very different: not only in their energy profiles (amplitudes, durations, etc.), but also with respect to their frequency contents. For example, the first and last strokes in this excerpt have a strong bass-like component. The remaining strokes present more energy in the mid- and high-frequency ranges, which is more akin to broadband noise. Moreover, we can observe that superposition between notes might occur, as with the first and second strokes, for example.

As we can see, this subset provides an interesting set of challenges for the problem under study: not only it contains different musicians/instruments combinations, but also many issues (e.g., inter-note interferences, varying note durations, and different tempi, among others) that we will have to face during the step of performance segmentation.

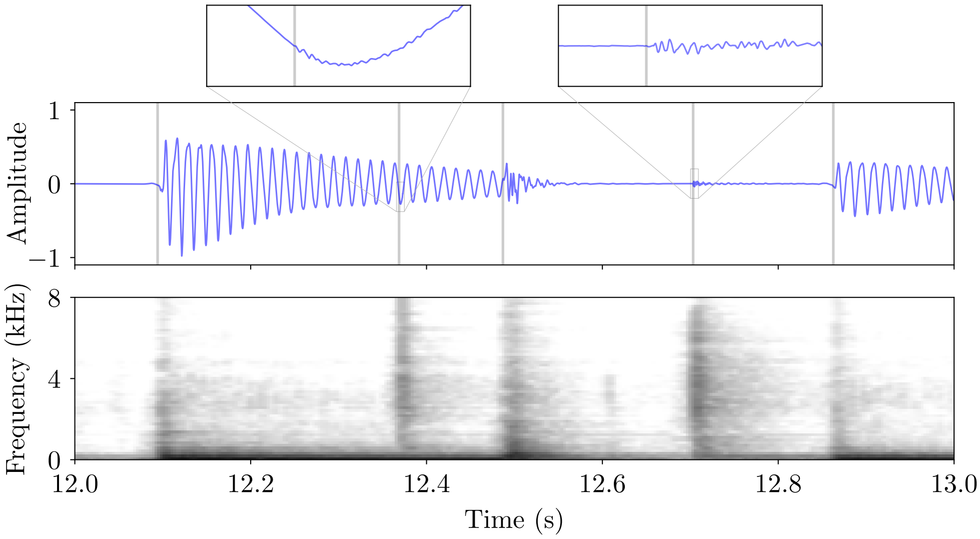


Figure 7.1: Strokes in an excerpt of file `[0357] S3-TT4-05-VSE` (*tantã* solo). From left to right, they correspond to: THUMB, SHELL, FINGERS, SHELL, THUMB. Notice a small burst of energy right after 12.6 s, which cannot be heard, and is possibly caused by the removal of the hand from the drumhead.

7.2 Onset Detection on the Subset

Segmenting an audio signal into its many events (i.e., notes) is a straightforward procedure provided that high-quality event annotations are available. When no annotations are provided, or in real-time/online applications, it is necessary to retrieve these events using some kind of onset detection function (ODF), so called because it assumes high values for time frames that contain an onset, and low values otherwise. ODFs are usually produced by processing the signal in time or frequency domains, but other techniques consider the phase of the Fourier transform, or even a combination of magnitude and phase. Time-domain methods are regarded as an adequate choice for detecting onsets of strongly percussive instruments [98] whereas the spectral flux is usually regarded as a one-size-fits-all solution [71]. Our samples, however, possess some particular characteristics, such as the superposition of note articulations with quite different properties, which are not thoroughly discussed in the onset detection literature. Since we do not know a priori what kind of onset detection function works best for our set of articulations, we choose to experiment with ODFs computed from different signal domains. We briefly describe these methods in the following, but the interested reader is referred to [5, 71, 98] for a full review.

To determine the best performing ODF, we first separate a small sample containing 25% of the recordings of *tantã* (six files) and *repique* (three files), which are selected considering tempo and timbral variety, and then produce ODFs for each of these recordings. Even though we have articulation annotations for all files in the dataset, this procedure (of splitting a fraction of the files for evaluation) mimics that of the real-world application, i.e., when no annotations are available. From the time domain, we compute the derivative of the short-time energy (E), considering only energy increases (half-wave rectification). From the set of spectral-based features, we derive a pair of flux-like functions. The first function uses the L_1-norm of the first-order difference between magnitude spectra (spectral difference, SD), while the second function uses the L_1-norm of the rectified difference (spectral flux, SF). We also compute the high frequency content (HFC) function, which is the frequency-weighted sum of the power spectrogram [98]. Finally, we analyze the four functions from the phase and complex domains [71]: phase deviation (PD), weighted phase deviation (WPD), complex domain (CD), and rectified complex domain (RCD). Phase deviation functions are based on the second derivative of the frame-wise channel phase, whereas complex domain functions work by computing the difference between the m-th frame and a prediction produced with the magnitude and phase derivative of the previous frame. We use the principal argument function to map phase differences to a suitable range [5]. Energy- and spectral-based ODFs are calculated with spectrograms in linear- and mel-frequency scales, in both linearly-scaled and log-compressed amplitudes (which roughly simulates loudness perception [98]). Phase-based and complex domain ODFs are computed with linear-frequency spectrograms only. However, unlike what is commonly done in the literature, we also use log-compressed amplitudes to weight phase deviations and to produce predictions for frame spectra — we believe that the dynamic range reduction could also enhance the recovery of more subtle onsets in the case of features that leverage both magnitude and phase. The log-compressed version of the magnitude time-frequency representation, $X[k, m]$, is given by

$$Y[k, m] = \log(1 + \gamma|X[k, m]|), \tag{7.1}$$

with $\gamma \geq 1$. In all calculations, we use $\gamma = 1000$. Input representations are computed with Hann windows of length 20 ms and 40 ms, and hop length of 10 ms. Mel-spectrograms are computed with 40 mel bands from 0 Hz to the Nyquist frequency. All ODFs (exemplified in Figure 7.2) are subjected to min-max normalization.

The selection of peaks is performed on all ODFs following the heuristic presented by BÖCK et al. [183], which is built upon the verification of three conditions. If $y[m]$ is the value of the ODF at the m-th time frame, we say that m contains an onset if it simultaneously satisfies:

1. the local maximum condition: $y[m]$ is the maximum in a neighborhood of m, i.e.,

$$y[m] = \max\{y[m - m_{\text{pre}}^{\max}], \cdots, y[m + m_{\text{post}}^{\max}]\}, \tag{7.2}$$

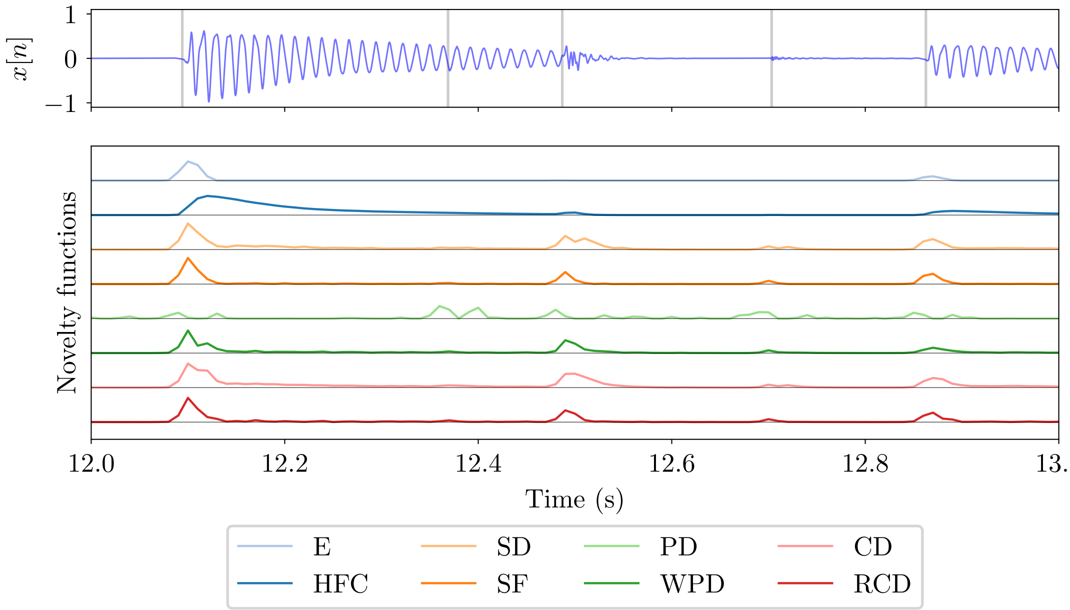


Figure 7.2: ODFs extracted from an excerpt of file `[0357] S3-TT4-05-VSE`. Functions were computed with linear-frequency spectrograms in a linear amplitude scale.

2. the local average condition: $y[m]$ exceeds the combination of the local average and a fixed threshold, i.e.,

$$y[m] \geq \delta + \frac{1}{m_{\text{post}}^{\text{avg}} + m_{\text{pre}}^{\text{avg}} + 1} \sum_{n=m-m_{\text{pre}}^{\text{avg}}}^{m+m_{\text{post}}^{\text{avg}}} y[n], \tag{7.3}$$

3. the waiting condition: m is located a certain time after the last detected onset, i.e.,

$$m - m_{\text{last}} > \Delta m_0. \tag{7.4}$$

The peak picking parameters — $m_{\text{pre}}^{\text{max}}$, $m_{\text{post}}^{\text{max}}$, $m_{\text{pre}}^{\text{avg}}$, $m_{\text{post}}^{\text{avg}}$, δ, Δm_0 — are usually determined according to the signal characteristics. Finding these parameters is a problem that is usually approached with grid search. This is also the approach that we use in this work. For the maximum condition, we look from 10 to 60 ms before the actual frame, in 10-ms steps. For the local average condition, we use from 40 to 120 ms before the actual frame, in 20-ms steps. In both cases, we extract peaks in causal mode ($m_{\text{post}}^{\text{max}} = m_{\text{post}}^{\text{avg}} = 0$, $m_{\text{pre}}^{\text{max}} \neq 0$, $m_{\text{pre}}^{\text{avg}} \neq 0$) and in a strictly symmetric mode ($m_{\text{post}}^{\text{max}} = m_{\text{pre}}^{\text{max}} \neq 0$, $m_{\text{post}}^{\text{avg}} = m_{\text{pre}}^{\text{avg}} \neq 0$). The threshold δ is searched from 0 to 0.07 in steps of 0.01, and the waiting parameter Δm_0 is selected equivalent to $\{0, 10, 20, 30\}$ ms. This leads to a total of 794 880 configurations.

We evaluate the peak picking results with the standard F-measure. An estimated onset is deemed correct if it lies within a time tolerance around a reference onset annotation. Given the number of correct detections (true positives, c), missed detections (false negatives, f^-), and extra detections (false positives, f^+) let us define precision (P) and recall (R) as [71]:

$$\begin{cases} P = \dfrac{c}{c + f^+}, \\ R = \dfrac{c}{c + f^-}, \end{cases} \tag{7.5}$$

respectively, and the F-measure as the harmonic mean between these two figures:

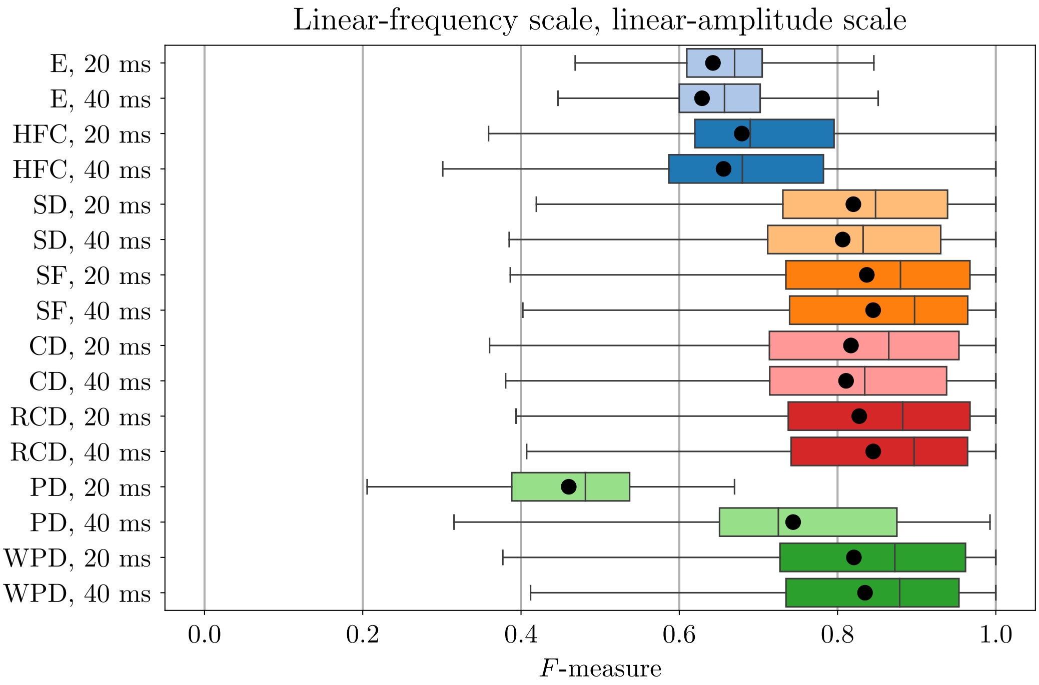
$$F = \left(\frac{P^{-1} + R^{-1}}{2} \right)^{-1} = \frac{2c}{2c + f^+ + f^-}. \tag{7.6}$$

This metric is usually expressed as a percentage.

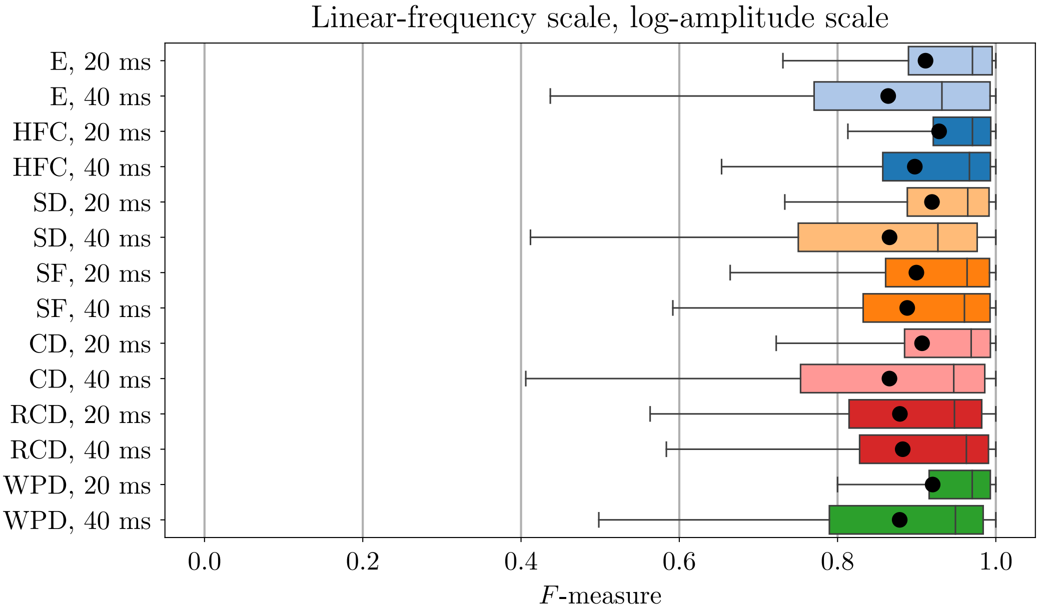
We separately run this experiment with two time error tolerances (detection windows) of ±25 ms and ±50 ms around annotations. Because onset positions are very important for a proper segmentation, we are more interested in the results produced within the narrower window, whereas the more relaxed condition serves as a reference since it is commonly employed in the literature [183].

The results of the grid search are summarized in Figures 7.3 and 7.4 for all configurations (ODF, analysis window, peak picking parameters) with a detection window of 25 ms. Overall, we can observe that using the logarithm to compress ODF input representations yields better onset detection results, which has been extensively reported in the literature. This is also true in the case of our proposed modifications for the complex domain, rectified complex domain, and weighted phase deviation ODFs. We also observe that, irrespective of the peak picking parameters, 20-ms analysis windows consistently yield results equivalent to or superior to those obtained with 40-ms analysis windows. One exception here is the phase deviation ODF, which profits from the larger window. It is worth noting that, generally speaking, grouping frequency bins according to the mel-scale yields slightly better detection. The top-performing ODF for our evaluation data is HFC (20-ms window, mel scale, log-amplitudes), which is less sensitive to the peak picking parameters, as evidenced by the highest mean, highest median, and one of the smallest interquartile ranges (IQR) over all configurations.

We now turn our attention to the results for the top performing configurations (ODF and grid search), which are shown in Table 7.3. For the sake of comparison, we also display configurations for the ±50 ms detection window. We instantly notice that performance figures are overall very similar. The best performing configurations all use as input ODFs computed over the linear-frequency scale STFT with window length of 40 ms and log-compressed amplitude. Unsurprisingly, configura-

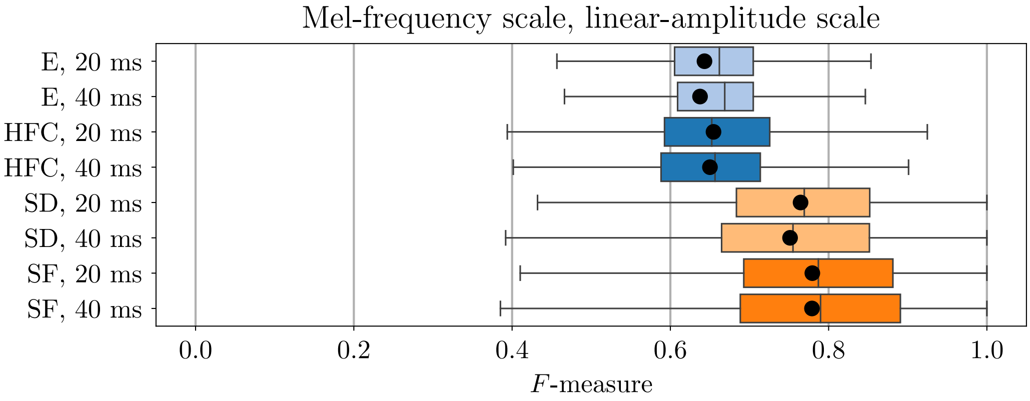


(a)

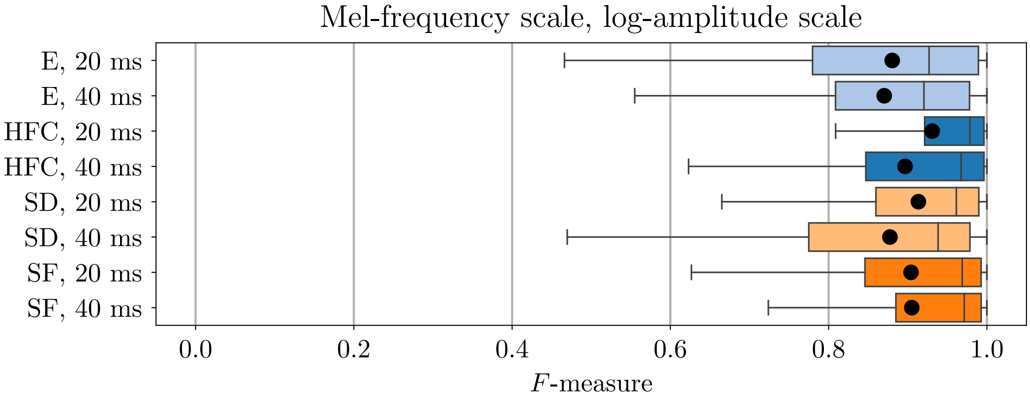


(b)

Figure 7.3: Grid search results for onset detection for ODFs computed with linear-frequency, and (a) linear- and (b) log-amplitude scales. Each box corresponds to interquartile range for F-measures (±25 ms) over the test set with different peak picking parameters. Mean and median values are indicated by • and |, respectively. ODFs: short-time energy (E); high frequency content (HFC); spectral difference (SD); spectral flux (SF); complex domain (CD); rectified complex domain (RCD); phase deviation (PD); weighted phase deviation (WPD).



(a)



(b)

Figure 7.4: Grid search results for onset detection for ODFs computed with mel-frequency, and (a) linear- and (b) log-amplitude scales. ODFs: short-time energy (E); high frequency content (HFC); spectral difference (SD); spectral flux (SF).

Table 7.3: Top five performing configurations in the grid search for each detection window. From left to right, columns correspond to: tolerance; onset detection function; input representation — frequency scale, amplitude scale, and window length; peak picking parameters — maximum condition, local average condition, fixed threshold, and waiting condition; average precision, recall, and F-measure across all files; mean absolute error between annotations and matched detections. X = "don't care" (all values in the grid yield the same performance).

		Input representation			Peak picking parameters				Evaluation			
Tolerance	ODF	Freq.	Ampl.	Win. (ms)	Max. (ms)	Avg. (ms)	Thr.	Wait. (ms)	$\bar{P}$ (%)	$\bar{R}$ (%)	$\bar{F}$ (%)	MAE (ms)
±25 ms	RCD	linear	log	40	±30	±100	0.04	X	99.41	99.41	99.41	3.0
	RCD	linear	log	40	±30	±120	0.03	X	99.22	99.55	99.38	3.1
	RCD	linear	log	40	±20	±100	0.05	0–20	99.44	99.33	99.38	3.0
	RCD	linear	log	40	±10	±100	0.05	20	99.38	99.38	99.37	3.0
	RCD	linear	log	40	±20	±80	0.06	0–20	99.56	99.17	99.36	3.0
±50 ms	E	linear	log	40	±30	±80	0.02	X	99.74	99.45	99.59	9.0
	SF	linear	log	40	±30	±80	0.05	X	99.81	99.36	99.58	6.6
	SF	linear	log	40	±20	±80	0.05	X	99.73	99.41	99.57	6.6
	SF	linear	log	40	±10	±80	0.05	30	99.73	99.41	99.57	6.6
	RCD	linear	log	40	±30	±100	0.04	X	99.56	99.57	99.56	3.0

tions with non-causal peak picking conditions yield better results than their causal counterparts (which did not make it to the top). With the broader ±50-ms detection window, the top ODFs are based on short-time energy, spectral flux, and rectified complex domain. With the shorter window, rectified complex domain remains at the top positions, whereas the formerly best energy-based and spectral flux configurations lose 0.62 and 0.42 (average) percent points on the mean F-measure across all files, respectively. Moreover, we observe that, with a ±25-ms tolerance window, the standard deviation of F-measure for the top performing complex domain configuration is among the smallest (0.54%). The respective standard deviations of those same SF and E configurations average at about 1.41% and 1.94%, respectively, when the short window is used. Even when the broader window is used, RCD presents good recall rates that surpasses those of the top four configurations. Finally, the last column in Table 7.3 displays the mean absolute error (MAE) between the detections produced by each configuration and the corresponding annotations of our evaluation data, which provides an idea of the accuracy of each configuration. To compute the MAE, first we seek a maximum matching between the sequences of true annotated onsets and estimated onsets, i.e., the largest set of correspondences within the given tolerance, such that each annotation and detection is only matched at most once. This means we do not consider neither the false positives nor false negatives for this metric. Then, we take the absolute difference and average the results for all matched pairs. We observe that the top five RCD-based configurations are also the best performing configurations with respect to the MAE. Therefore, despite the higher computational overhead (when compared to other ODFs), we will use RCD as our ODF, with the representation and peak picking parameters listed on the first row of Table 7.3. We note that an energy- or spectral-flux-based ODF might be suitable for our purposes as well, as we discuss in the next section.

7.3 Segmentation of Articulations

The next step in the processing of note articulations is segmentation. This task, like onset detection, must be approached with care. Note events have different durations (cf. Figure 7.1), with most information about the articulation being present in the attack portion of the note. Moreover, overlapping between consecutive articulations increases the difficulty in the segmentation procedure.

First, we use RCD with previously determined peak picking parameters to detect onsets in each file of the entire subset, including the recordings used to evaluate the ODFs. The results of this detection can be seen in Table 7.4, separated by instrument and articulation type. For the sake of comparison, we include the results for the other two ODFs (E and SF), with their corresponding best parameters in

Table 7.4: Onset detection results in the entire subset. F-measure and MAE display compound performance figures for all articulations.

			#true positives			$\bar{F}$ (%)			MAE (ms)		
	Artic.	#	E	SF	RCD	E	SF	RCD	E	SF	RCD
TT	FINGERS	1192	1187	1179	1186	98.84	99.05	98.91	9.4	6.7	3.2
	HAND	1071	1067	1065	1067						
	SHELL	1688	1677	1682	1683						
RP	THUMB	876	875	875	876	99.55	99.46	99.15	8.5	6.2	2.8
	FINGERS	620	620	619	618						
	SHELL	917	910	911	911						

the evaluation set. It is apparent from the table that, once again, detection F-measure performances are very similar for the three ODFs. However, we can try to attribute their minute differences to the way each ODF is built, as well as to how our preliminary experiment was thought out and developed. For example, the energy-based ODF performs slightly worse in the detection of SHELL strokes, which generally carry less energy than the other articulations. Moreover, RCD has higher recall for *tantã* (99.44%) than E and SF (99.11% and 99.03%, respectively), while for *repique* it performs slightly worse than the other ODFs (99.55% against 99.62% and 99.57%). A possible explanation for this may be the fact that our grid search subset features more *tantã* onsets, which results in peak picking parameters that are better suited for this instrument.

It seems reasonable that this slightly worse performance by RCD in *repique* can also be attributed to the frequency content of each instrument. *Repique* has considerably more energy in high frequencies ($\geq$ 5 kHz) than *tantã* and proportionately more phase instabilities in this region of the spectrum, which may have greater effect on phase-based and complex domain ODFs than on other methods. Nevertheless, RCD still produces onset estimates that, on average, are closer to the annotations than the other two methods, as it is evident by the MAE results in Table 7.4. In fact, if instead of looking to the absolute error, we compute the mean and standard deviation of the regular error (estimated onset position minus reference annotation), given in ms, we obtain $(-9.3, 4.0)$, $(-6.5, 4.1)$, and $(-1.5, 4.1)$ for E, SF, and RCD in *tantã* recordings, and $(-8.5, 3.5)$, $(-6.1, 3.6)$, and $(-0.6, 3.4)$, respectively, in *repique*. This indicates that there is a general bias towards estimating onsets earlier than their corresponding annotations,[1] which is considerably smaller in the case of RCD.

[1] More precisely, we can speak of a bias towards perceiving onsets a few milliseconds after they occur, which is reported in literature as the difference between the physical (acoustic) and perceptual onsets. This delay is typically smaller for drums in general, and more so in the case of high sounding drums, such as the snare drum [184, 185], or the *repique* in our case.

Once onset frames (and corresponding timestamps) are estimated, we can finally segment each event. Since we are only interested in the performance of the classifiers with respect to the different features, we again perform a maximum matching between estimates and annotations. We discard the times of detected onsets without a match (false positives), and use only those with a corresponding annotation, which are labeled accordingly with articulation types. Next, we perform the segmentation considering three different starting points: exactly at the estimated onset position; 5 ms before the onset; and a backtracked position — the frame index corresponding to the previous minimum of the ODF. In all cases, we consider a segment length of at most 125 ms or 80% of the duration between each starting point and the next one, whichever is shorter. This criterion is consistent with the average inter-onset interval in our subset, and is around the durations used by VAN STEELANT et al. [99] in the classification of percussive sounds. Moreover, a few of the features we investigate (e.g., modulation features) require a signal analysis over a window that has a minimum size to produce meaningful representations. If the classification scheme we are presenting in this Chapter were to be embedded into a realtime system, the length of the segmentation window and the peak picking parameters (at the previous step) would have to be taken into account for the total latency. We have also attempted to determine the length of each segment by gating the instantaneous power, but resulting segments were inconsistent mostly due to interference between notes.

7.4 Feature Extraction

We approached the classification of segmented notes with features that can be grouped into four different categories:

- temporal features — log attack time, crest factor, decrease, attack/decay ZCR, envelope moments (centroid, spread, skewness, kurtosis), strong decay;
- spectral features — energy and energy band ratios, moments (centroid, spread, skewness, kurtosis), spectral flatness, spectral crest, strong peak, spectral roll-off, spectral contrast and valley;
- cepstral features — MFCC, GFCC, BFCC;
- modulation features — CQT modulation spectrum, scattering coefficients.

All features are described in detail in Chapter 5 along with important parameters. Other parameters were obtained empirically and are defined in the following.

We use the true amplitude envelope for the computation of temporal features. Spectral and cepstral features are calculated with a sliding Hann window of length

23 ms with 75% overlap. Eight bands are used for the computation of the energy distribution (relative to the total energy): 40 to 80 Hz, 80 to 160 Hz, 160 to 320 Hz, 320 to 640 Hz, 640 to 1280 Hz, 1.28 to 4 kHz, 4 to 8 kHz, and 8 to 22.05 kHz. Spectral contrast is computed for seven octave bands, starting at 20 Hz, as formulated by [107]. For MFCCs and BFCCs, we regroup all frequency bins from 0 to 22 050 Hz with 40 analysis bands, keeping only the first 13 coefficients for the corresponding cepstral coefficients. A similar procedure is followed for GFCCs, but with the first filter centered at 40 Hz. Scattering transforms are computed with a maximum log-scale of 10, with $Q_1 = 3$ Morlet wavelets per octave for first order coefficients, and $Q_2 = 1$ wavelet per scale for the second order. We also investigate the power of our CQT modulation spectrum in discriminating note articulations. The first stage is computed at a frame rate of 344 Hz with filters whose center frequencies go from 40 to 8127 Hz, spanning eight octaves with three bins per octave, and the second stage represents modulation frequencies of 12 to 96 Hz. For simplicity, we will refer to both as the first and second stages of the CQT modulation, even though the first stage is a constant-Q spectrum.

The `Essentia` Python package [106] is used for the computation of log attack time, ZCR, temporal moments, spectral, and cepstral descriptors. We use `Kymatio` [186] (Python) to obtain scattering coefficients. For all short-time descriptors, we consider as features both the mean and the variance across frames. For the sake of simplicity, we concatenate the temporal features and treat them all together. We proceed similarly for features that describe the general spectral shape — moments, flatness, crest, strong peak, and roll-off —, which are aggregated as "spectral shape" features. We include delta features (mean and variance of the first derivatives) for energy and band ratios, spectral shape, spectral contrast, spectral valley, and cepstral features. Lastly, we also report classification results with log-compressed versions of the modulation features, as commonly done in the literature [128].

7.5 Classification of Segments

Three questions motivate our investigation:

- Which set of features allow for a good generalization in *tantã* articulations ("FINGERS", "HAND", and "SHELL") when we consider the different instrument variations?

- Which set of features yield good classification accuracy for *repique* articulations ("THUMB", "FINGERS", and "SHELL")?

- Which set of features let us identify archetypal strokes (“OPEN”, “CLOSED”, and “SHELL”) common to both *tantã* and *repique*?

In our experiments, we perform the classification of note segments using SVM classifiers in one-vs-one (OVO) strategy. We have used the implementation available with the `scikit-learn` Python package [187]. Note that, for *tantã* or *repique* (or for the archetypal strokes), we are always dealing with a three-class classification problem. We evaluate SVMs with linear and radial basis function (RBF) kernels, experimenting with different values for the main hyperparameters: the margin/regularization parameter, C, and the parameter γ that controls the width of the RBF kernel $\kappa(\cdot,\cdot)$, such that

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{e}^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2}, \tag{7.7}$$

where $\|\cdot\|$ indicates the Euclidean distance (L_2-norm). Unless otherwise noted, we perform all tests with nested leave-one-group-out cross-validation (LOGOCV). For *tantã*, we first split data into four non-uniform groups, according to the instrument variation of each recording. This is done to avoid any kind of “data leakage” that may happen due to training and testing on recordings from the same variation. For *repique*, since there is a single instrument, we split data into three groups, one for each performer. In either case, we average the results of training over all but one group and testing on the remaining one. We tune all hyperparameters with grid search in an “inner” ten-fold cross-validation loop, which is repeated three times with data being randomly split in each repetition, respecting class distributions. We search for $C \in \{10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ and for $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$, around the recommended values [187]. Training data are scaled to zero mean, unit variance; these scaling parameters are used to scale corresponding validation and testing data. In each run of the tuning procedure, models are fit and then evaluated over validation sets with the macro F-measure, i.e., the arithmetic mean of per-class F-measures. The model with the best mean cross-validated score is chosen and refit on the entire training set before being evaluated over each left-out group.

7.5.1 Articulations of *Tantã* and *Repique*

We investigate the classification performance of each feature set when extracted from segments cut at different starting points — “exact”, “5 ms” before the estimated onset, and “backtrack” (i.e., cutting at the previous minimum of the RCD ODF). This is done separately for *tantã* and for *repique*, and the results displayed on Tables 7.5 and 7.6, respectively. We report averages for both accuracy and macro F-measure over test sets.

One thing that is interesting about the results in these tables is that there seems to be a small advantage in cutting the segment before the onset estimation. This

Table 7.5: Average cross-validation results for the classification of *tantã* articulations. Standard deviations are shown between parentheses. In gray, we highlight the best macro-averaged F-measure among different cutting points; in boldface, we highlight the best macro-averaged F-measure over all features.

		Dim.	Average accuracy			Average F-measure (macro)		
			Exact	5 ms	Backtrack	Exact	5 ms	Backtrack
	Temporal features	10	0.862(0.138)	0.903(0.107)	0.911(0.077)	0.864(0.132)	**0.902(0.104)**	**0.902(0.083)**
Spectral	Energy and band ratios	18	0.743(0.219)	0.809(0.188)	0.760(0.194)	0.721(0.237)	0.795(0.202)	0.732(0.223)
	Energy and band ratios + Δ	36	0.820(0.178)	0.837(0.177)	0.800(0.178)	0.815(0.175)	0.834(0.176)	0.786(0.185)
	Spectral shape	16	0.794(0.239)	0.800(0.256)	0.773(0.200)	0.767(0.280)	0.777(0.292)	0.735(0.243)
	Spectral shape + Δ	32	0.796(0.243)	0.803(0.269)	0.795(0.241)	0.778(0.272)	0.780(0.304)	0.776(0.260)
	Contrast/valley	28	0.817(0.143)	0.793(0.248)	0.764(0.199)	0.784(0.198)	0.778(0.271)	0.746(0.214)
	Contrast/valley + Δ	56	0.807(0.170)	0.827(0.190)	0.794(0.238)	0.780(0.213)	0.823(0.192)	0.778(0.254)
Cepstral	MFCC	26	0.819(0.199)	0.816(0.195)	0.792(0.183)	0.766(0.275)	0.765(0.268)	0.741(0.254)
	MFCC + Δ	52	0.834(0.213)	0.836(0.220)	0.815(0.203)	0.788(0.285)	0.785(0.302)	0.766(0.282)
	BFCC	26	0.764(0.186)	0.773(0.177)	0.768(0.180)	0.692(0.261)	0.710(0.248)	0.713(0.257)
	BFCC + Δ	52	0.842(0.193)	0.831(0.203)	0.811(0.199)	0.806(0.251)	0.785(0.279)	0.764(0.275)
	GFCC	26	0.793(0.188)	0.801(0.197)	0.740(0.190)	0.750(0.258)	0.756(0.272)	0.694(0.258)
	GFCC + Δ	52	0.823(0.172)	0.817(0.194)	0.814(0.162)	0.793(0.220)	0.780(0.256)	0.789(0.195)
Modulation	Time scat., first order	52	0.847(0.196)	0.820(0.204)	0.823(0.182)	0.828(0.221)	0.810(0.212)	0.793(0.223)
	Time scat., first order (log)	52	0.880(0.159)	0.885(0.168)	0.848(0.177)	0.858(0.196)	0.861(0.205)	0.825(0.211)
	Time scat., sec. order	182	0.777(0.119)	0.809(0.122)	0.814(0.144)	0.729(0.165)	0.770(0.177)	0.772(0.200)
	Time scat., sec. order (log)	182	0.784(0.237)	0.842(0.146)	0.841(0.162)	0.778(0.244)	0.816(0.187)	0.816(0.200)
	Time scat., first+sec. order	234	0.883(0.151)	0.867(0.137)	0.843(0.187)	0.862(0.184)	0.858(0.142)	0.812(0.230)
	Time scat., first+sec. order (log)	234	0.857(0.162)	0.850(0.200)	0.825(0.235)	0.835(0.198)	0.828(0.235)	0.808(0.257)
	CQT, first stage	48	0.837(0.210)	0.839(0.210)	0.774(0.261)	0.811(0.250)	0.829(0.221)	0.763(0.264)
	CQT, first stage (log)	48	0.822(0.258)	0.815(0.278)	0.803(0.254)	0.802(0.291)	0.793(0.314)	0.785(0.278)
	CQT, sec. stage	192	0.830(0.207)	0.808(0.216)	0.815(0.222)	0.786(0.272)	0.783(0.239)	0.782(0.267)
	CQT, sec. stage (log)	192	0.867(0.178)	0.871(0.177)	0.877(0.154)	0.837(0.225)	0.839(0.228)	0.840(0.211)
	CQT, first+sec. stages	240	0.859(0.172)	0.856(0.186)	0.839(0.184)	0.834(0.208)	0.851(0.187)	0.823(0.202)
	CQT, first+sec. stages (log)	240	0.874(0.188)	0.874(0.194)	0.878(0.176)	0.846(0.235)	0.846(0.241)	0.848(0.225)

Table 7.6: Average cross-validation results for the classification of *repique* articulations. Standard deviations are shown between parentheses. In gray, we highlight the best macro-averaged F-measure among different cutting points; in boldface, we highlight the best macro-averaged F-measure over all features.

		Dim.	Average accuracy			Average F-measure (macro)		
			Exact	5 ms	Backtrack	Exact	5 ms	Backtrack
	Temporal features	10	0.919(0.015)	0.933(0.014)	0.904(0.047)	0.916(0.013)	0.929(0.018)	0.898(0.050)
Spectral	Energy and band ratios	18	0.833(0.061)	0.829(0.048)	0.796(0.026)	0.813(0.067)	0.808(0.056)	0.786(0.023)
	Energy and band ratios + Δ	36	0.842(0.033)	0.887(0.038)	0.834(0.002)	0.828(0.045)	0.879(0.046)	0.823(0.012)
	Spectral shape	16	0.853(0.056)	0.808(0.045)	0.850(0.019)	0.838(0.058)	0.791(0.052)	0.836(0.024)
	Spectral shape + Δ	32	0.858(0.032)	0.869(0.061)	0.868(0.021)	0.847(0.035)	0.860(0.065)	0.860(0.016)
	Contrast/valley	28	0.818(0.049)	0.891(0.044)	0.891(0.048)	0.797(0.057)	0.881(0.053)	0.882(0.053)
	Contrast/valley + Δ	56	0.794(0.057)	0.886(0.050)	0.872(0.059)	0.769(0.063)	0.873(0.057)	0.864(0.055)
Cepstral	MFCC	26	0.825(0.067)	0.868(0.044)	0.866(0.046)	0.795(0.080)	0.850(0.050)	0.855(0.049)
	MFCC + Δ	52	0.885(0.060)	0.892(0.054)	0.911(0.033)	0.872(0.067)	0.878(0.061)	0.906(0.036)
	BFCC	26	0.817(0.075)	0.828(0.064)	0.852(0.032)	0.797(0.080)	0.806(0.070)	0.839(0.025)
	BFCC + Δ	52	0.862(0.058)	0.875(0.055)	0.880(0.009)	0.846(0.065)	0.862(0.061)	0.864(0.015)
	GFCC	26	0.873(0.021)	0.829(0.056)	0.865(0.028)	0.856(0.024)	0.808(0.060)	0.850(0.028)
	GFCC + Δ	52	0.888(0.043)	0.872(0.051)	0.886(0.032)	0.876(0.048)	0.857(0.059)	0.874(0.032)
Modulation	Time scat., first order	52	0.877(0.091)	0.898(0.079)	0.875(0.100)	0.876(0.087)	0.895(0.076)	0.873(0.096)
	Time scat., first order (log)	52	0.874(0.087)	0.857(0.094)	0.892(0.058)	0.858(0.104)	0.838(0.113)	0.880(0.067)
	Time scat., sec. order	182	0.832(0.083)	0.845(0.104)	0.897(0.053)	0.825(0.077)	0.841(0.098)	0.888(0.061)
	Time scat., sec. order (log)	182	0.851(0.076)	0.840(0.091)	0.871(0.077)	0.841(0.077)	0.826(0.102)	0.858(0.091)
	Time scat., first+sec. order	234	0.827(0.144)	0.832(0.147)	0.931(0.043)	0.825(0.138)	0.830(0.140)	0.926(0.044)
	Time scat., first+sec. order (log)	234	0.891(0.070)	0.840(0.078)	0.859(0.086)	0.877(0.081)	0.827(0.082)	0.849(0.092)
	CQT, first stage	48	0.908(0.045)	0.843(0.143)	0.881(0.072)	0.904(0.047)	0.843(0.137)	0.876(0.072)
	CQT, first stage (log)	48	0.827(0.051)	0.821(0.035)	0.844(0.047)	0.803(0.052)	0.793(0.031)	0.823(0.046)
	CQT, sec. stage	192	0.928(0.038)	0.925(0.039)	0.903(0.058)	0.925(0.040)	0.922(0.040)	0.900(0.059)
	CQT, sec. stage (log)	192	0.953(0.017)	0.938(0.038)	0.917(0.040)	**0.952(0.017)**	0.935(0.038)	0.912(0.044)
	CQT, first+sec. stages	240	0.921(0.049)	0.925(0.046)	0.909(0.053)	0.919(0.050)	0.924(0.046)	0.905(0.055)
	CQT, first+sec. stages (log)	240	0.953(0.020)	0.934(0.031)	0.927(0.035)	0.950(0.022)	0.928(0.033)	0.924(0.036)

is particularly true for spectral-based features in *tantã* recordings, for which the "5 ms" segmentation procedure yielded the best results. Best F-measures for *tantã* articulations are achieved with: temporal (90.2%); first-/second-order time-scattering (86.2%); first-order time-scattering log-coefficients (86.1%), and first-/second-stage CQT modulation (85.1%). For *repique*, a large proportion of the best results involve the use of CQT modulation coefficients. Top-performing are the log-compressed CQT modulation coefficients (95.2%), and the log-compressed version of the aggregation of the CQT spectrum and CQT modulation spectrum (95.0%). Interestingly enough, the CQT modulation alone produces a F-measure of 92.5% — when appending first-stage coefficients this metric drops by 0.1%, probably due to the increase in dimensionality. Temporal features also give good results for *repique* (92.9%). It is somewhat surprising that our proposed modulation features outperform MFCCs and other cepstral representations, which have for long been regarded as state of the art in timbre recognition problems. We observe that adding dynamic features (deltas) consistently improves the classification results. On the other hand, there is no clear evidence that log-compression on modulation coefficients boosts performances.

Closer inspection of the results show that *tantã* classification presents smaller averages and larger standard deviations when compared with the results for *repique* articulations. A likely cause for this might be that *tantã* recordings contain instrument variations with very different characteristics: from the drumhead size (10 to 14 in) to its material (leather and napa/nylon). Since we separate training and testing according to the variation (see Table 7.7), some are more affected by this data split than others. To illustrate this, we present typical cross-validation results for *tantã* in Table 7.7. It becomes immediately evident from this table that models trained on TT1, TT2, and TT3 (leather drumheads) generalize poorly to TT4 (napa/nylon). This ends up lowering the mean and increasing the standard deviation of the results for this instrument overall, something that does not happen in the case of *repique*.

Table 7.7: Example of LOGOCV results with temporal features from *tantã* recordings. We report the best model results on the test group in each case, as well as the average validation F-measures (macro).

Train	Test	Best model			Val. F
		Kernel	Acc.	F	
TT2, TT3, TT4	TT1	RBF ($C = 10^2$, $\gamma = 10^{-1}$)	0.971	0.969	0.969
TT1, TT3, TT4	TT2	RBF ($C = 10^2$, $\gamma = 10^{-1}$)	0.978	0.978	0.968
TT1, TT2, TT4	TT3	RBF ($C = 10^2$, $\gamma = 10^{-1}$)	0.943	0.937	0.971
TT1, TT2, TT3	TT4	RBF ($C = 10^1$, $\gamma = 10^{-1}$)	0.718	0.725	0.986
Test accuracy avg.(std.): 0.903(0.107)					
Test F-meas. avg.(std.): 0.902(0.104)					

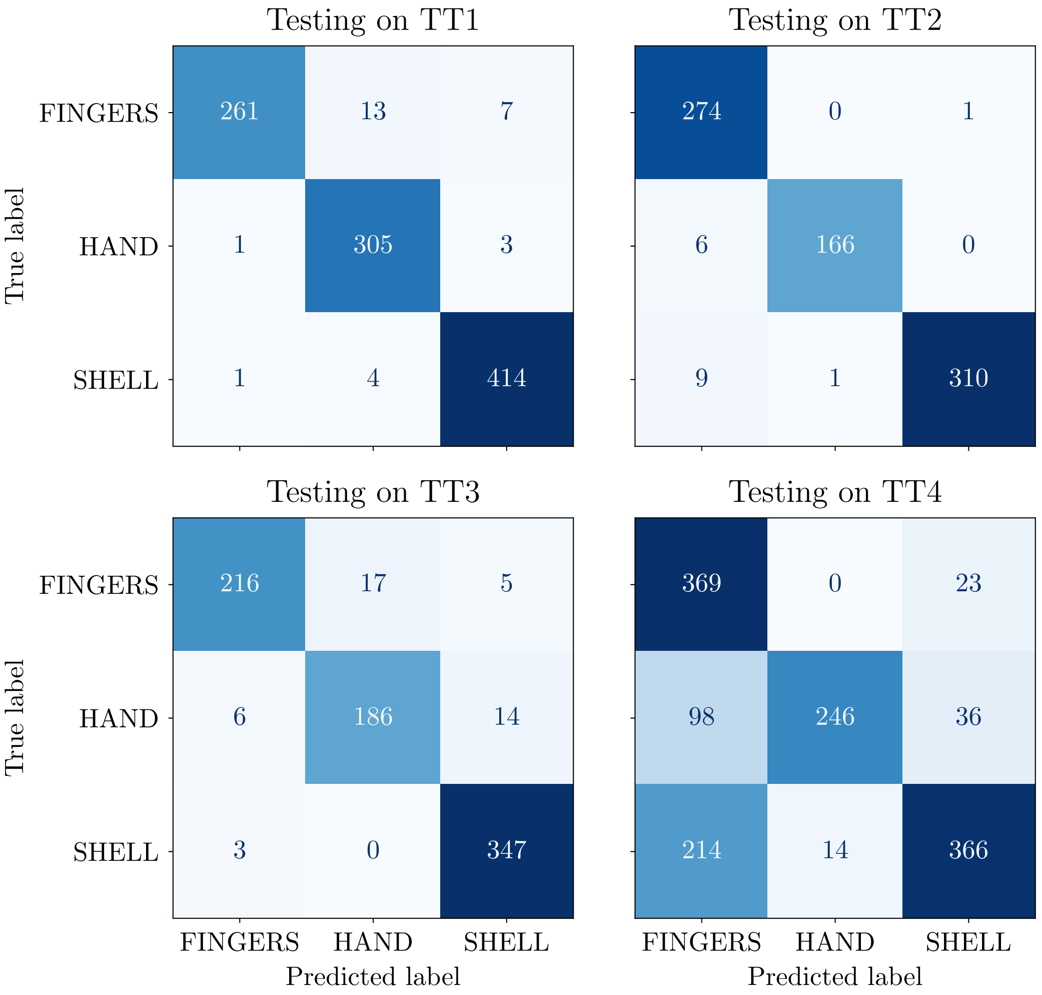


Figure 7.5: Confusion matrices in each of the LOGOCV steps for the best performing model (temporal features, 5 ms) in *tantã* articulations.

Another way to visualize this is through confusion matrices, which are shown for the best performing model configurations in Figures 7.5 and 7.6, for *tantã* and *repique*, respectively. For the model trained on temporal features extracted from *tantã* recordings, we observe in general very good recall and precision rates for each articulation, with most mistakes arising from the confusion between HAND and FINGERS, or HAND and SHELL. However, when training on TT1/TT2/TT3 and testing on TT4, almost a third of the HAND strokes are misclassified as either FINGERS or SHELL, whereas 36% of SHELL strokes are wrongly labeled FINGERS. It is worth noting that, due to its synthetic drumhead, the sounds produced by TT4 are “drier” than from its counterparts, which could explain the overall confusion in this case. The model tested on TT4 also tends to overestimate the number of FINGERS labels. For *repique*, the confusion matrices have well defined diagonals, with few errors due to mixes between open (THUMB) and closed sounds (FINGERS).

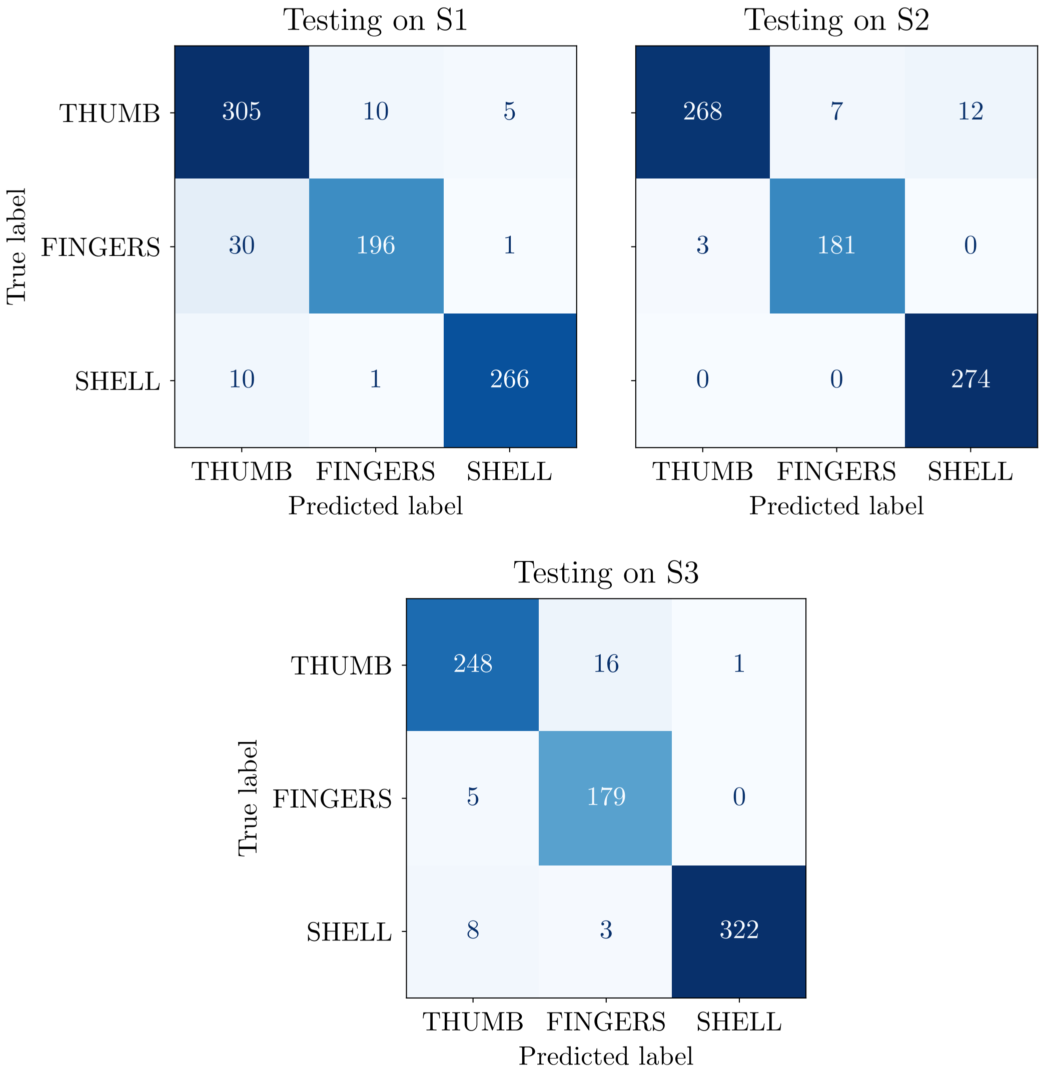


Figure 7.6: Confusion matrices in each of the LOGOCV steps for the best performing model (CQT modulation log-coefficients, exact) in *repique* articulations.

7.5.2 Archetypal Strokes

For this experiment, we group the labels for *tantã* and *repique* segments (observing the sonorous characteristics and musical meaning of each stroke type) and perform the classification of these "archetypal strokes". The labels FINGERS / HAND / SHELL (*tantã*) and THUMB / FINGERS / SHELL (*repique*) are thus mapped to OPEN / CLOSED / SHELL. To simplify our analysis, we classify only basic features (no log-compression or delta features) extracted from segments that start precisely at the onset estimates. This time, we do not perform hyperparameter optimization, instead using a linear SVM and fixing $C = 10$, which is around the average value of the cross-validation procedure from the previous experiment for this kind of kernel. Again, we carry the training/testing split following a LOGO procedure with instrument variations (TT1, TT2, TT3, TT4, and RP1).

Table 7.8: Classification results for archetypal articulations. The highest scores are highlighted in bold.

Feature	Dim.	Accuracy	F-measure
Temporal features	10	0.822(0.126)	0.798(0.157)
Energy and band ratios	18	0.686(0.213)	0.678(0.218)
Spectral shape	16	0.693(0.169)	0.660(0.199)
Contrast/valley	28	0.511(0.161)	0.426(0.149)
MFCC	26	0.739(0.143)	0.686(0.183)
BFCC	26	0.751(0.171)	0.701(0.210)
GFCC	26	0.747(0.188)	0.702(0.239)
Time scat., first order	52	0.788(0.201)	0.761(0.233)
Time scat., sec. order	182	0.781(0.179)	0.736(0.225)
CQT, first stage	48	0.815(0.167)	0.793(0.202)
CQT, sec. stage	192	**0.826(0.193)**	**0.801(0.227)**

Table 7.9: Example of LOGO results with a temporal and modulation features for the classification of archetypal strokes. For each test group, the highest F-measures are highlighted in bold.

Train	Test	Test F-measure (macro)	
		Temporal	CQT mod.
TT2, TT3, TT4, RP1	TT1	0.913	**0.935**
TT1, TT3, TT4, RP1	TT2	0.946	**0.960**
TT1, TT2, TT4, RP1	TT3	0.912	**0.964**
TT1, TT2, TT3, RP1	TT4	**0.647**	0.367
TT1, TT2, TT3, TT4	RP1	0.569	**0.779**

Table 7.8 provides summary statistics for test set accuracies and F-measures from the LOGO procedure. We notice that the CQT modulation coefficients are the most capable in generalizing the classification of archetypal strokes among different instrument variations with 80.1% F-measure. The set of temporal features come in second place, with 79.8% F-measure, but with smaller standard deviation (15.7% against 22.7% for the former). Table 7.9 shows the test results of each LOGO iteration for both feature families. We observe that classification with CQT modulation coefficients actually performs better on RP1 than when temporal features are used, but has a much inferior result on TT4. The performance of the classifier using temporal features on TT4 segments is worse than when other variations (TT1, TT2, TT3) serve as test, but does not drop as much (about 30 percent points against 60) as was the case with CQT modulation features — moreover, its performance is better than that of the classification of RP1 samples with the same feature. These results cannot be simply attributed to imbalances in instrument variations, and require further investigation.

We conclude this experimental section with an attempt to improve the cross-instrumental classification results. One possibility, for example, is to provide the classifier with sets of aggregated features, like we have done with the first- and second-stage scattering coefficients, for example. With multiple viewpoints, in particular, features coming from different domains, the classifier has more options to define separating hyperplanes. The problem we may face sooner than later when aggregating features is that of the “curse of dimensionality” — a high-dimensional feature space requires an even higher number of training samples. Still, the classifier for the problem at hand could probably benefit from more viewpoints. There are several ways of boosting the performance of estimators by selecting meaningful features — these usually require some kind of scoring function in sequential methods such as recursive feature elimination. We refer the reader to the work by ESSID [95] for an application of these techniques to general musical instrument recognition.

In this work, we follow a simple and interpretable approach of combining different features two by two. We follow the same training procedure as in the single-feature case, but this time with each concatenated set of features, i.e., we run the training procedure for all 55 feature combinations. Results for this experiment on the classification of archetypal strokes can be seen in Figure 7.7. The aggregation of temporal and CQT modulation features yields the best classification F-measure (89.0%) over all feature sets. This represents about 8.9-percent-point increase over classifying only with CQT modulation and 9.2 percent points over classifying only with temporal features (cf. Table 7.8). Another interesting combination is that of temporal features with GFCCs, with which the SVM classifier achieves 86.3% F-measure. Overall, we can note that it is very useful to append temporal features to any other feature set, whereas adding features from the same domain does not usually represent a great improvement. Instead, sometimes doing so deteriorates the performance of the classifier. For example, aggregating MFCCs and BFCCs results in an average reduction of 5.5 percent points in F-measure.

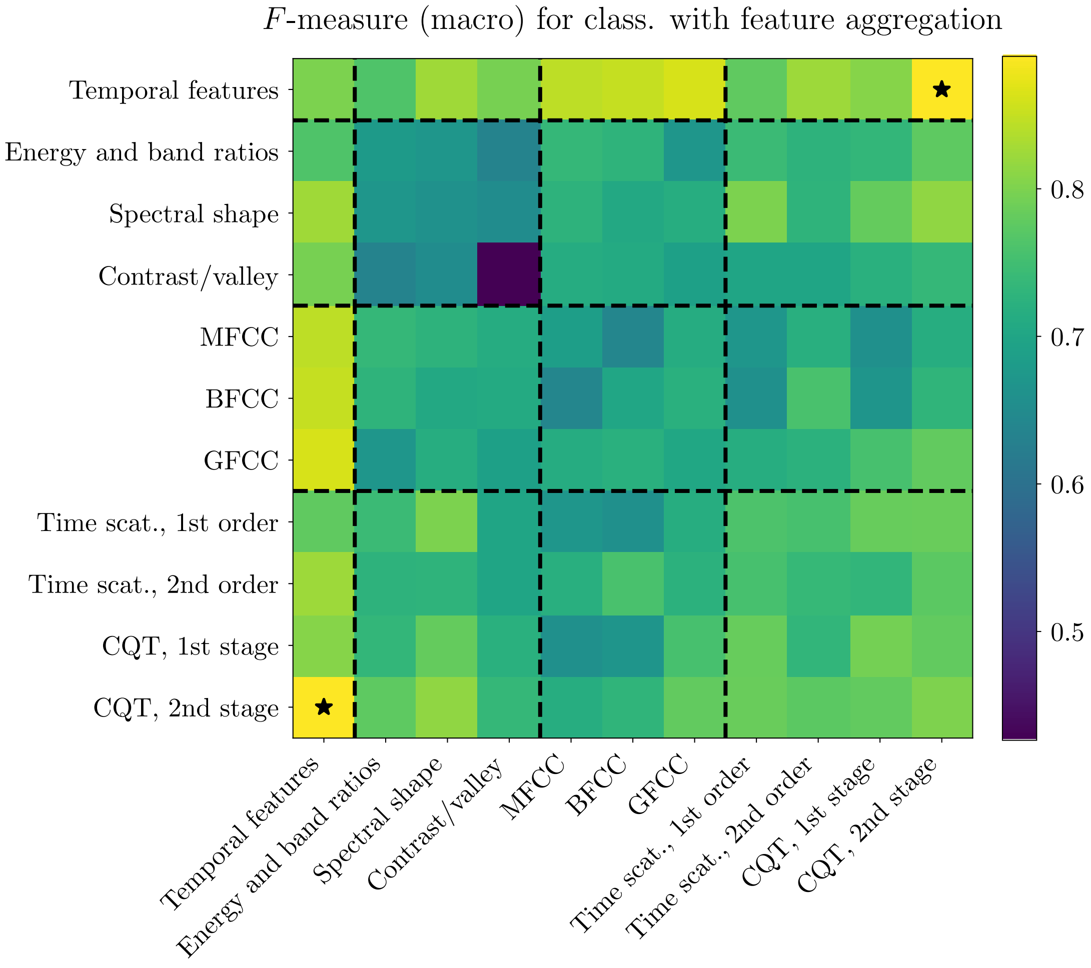


Figure 7.7: Results for classification with aggregated features. Single-feature results are represented along the main diagonal and $\star$ is the best combination.

Part III

Rhythmic Description

Chapter 8

Introduction to Rhythmic Description

In Section 2.4, we have briefly introduced the concept of rhythm and highlighted the notions of timelines, contrametricity, commetricity, and polyrhythm (i.e., the superposition of different rhythmic layers) in *samba*. We also mentioned the phenomenon of *marcialização*, which occurs when the main characteristics of *samba* are lost due to an increase in playing speed. Moreover, we presented some of the most common *batidas* of each instrument in a *bateria*. Now, we will focus on the MIR topic of rhythmic description, which aims to represent time in a "compact and generalizable form" [188]. We can interpret the rhythmic patterns in a music recording as a combination of pulse rate and amplitude modulations [189, 190]. This means that an investigation of those patterns could benefit from periodicity analyses, for example.

In MIR, several methods have been proposed for identifying periodicity patterns and rhythm in music. These methods are sometimes used to detect tempo, infer meter, and characterize small-scale deviations (e.g., swing) in music signals. Additionally, they serve as a preprocessing step in many tasks that rely on similarity, including genre classification and collection retrieval. In fact, according to COCHARRO et al. [15], in the context of machine understanding of rhythm, the research topics of rhythm description and rhythm similarity are closely connected, with similarity models being reliant on the subjacent encoded representation of time patterns [191]. Rhythm description is relevant for various applications, from music recommendation and retrieval, to musicological analyses, composition, and performance [15].

Approaches in the literature vary greatly according to the domain in which the music is presented, i.e., either symbolic or audio. For the analysis of rhythmic patterns (and their similarity) in symbolic domains, we refer to works by TOUSSAINT [191–193]. In the case of audio recordings, which are the subject of this work, rhythmic representations can be categorized into three levels of abstraction — low, middle, and high [15, 194]. At the lowest level, we find the representations obtained via an accent signal, such as the spectral flux and other novelty functions used in Chapter 7. Mid-level representations elicit information comparable to that of a transcription. Examples include histograms of inter-onset intervals and other periodicity features such as the autocorrelation function (ACF). COCHARRO et al. [15] also include, at this level, the representations obtained from tempo estimation and beat tracking. Finally, high-level description is defined at the domain of composition theory or style analysis [15]. These methods can also be classified according to how they compute the similarity of the derived representations: a few examples are the Euclidean and cosine distances, and dynamic programming algorithms [188].

In this chapter, we present a review of the main rhythmic description methods in the MIR literature. We begin, however, by defining a few concepts common to this chapter and the following ones. Lastly, we briefly investigate some techniques from the literature for representing the rhythmic patterns in our datasets and propose a few modifications.

8.1 Musical Concepts for Rhythmic Description

Figure 8.1 presents, in an idealized score-notation format, a simple motivating example containing the concepts used in this and the following chapters. We identify the beat as the main temporal unit perceived in a music piece. This predominant pulsation is performed at a rate per unit of time that defines the piece's tempo. In our example, the beat is related to the quarter note and the metronome marking of 100 bpm indicates the general tempo. However, throughout a piece, tempo is rarely maintained constant: it can vary greatly as a result of compositional or performative choices. We note that music genres can be sometimes characterized by typical tempo [3].

The meter defines a regularly recurring pattern established by the succession of rhythmic pulses and corresponding accents (metrical structure). One such complete pattern is the measure (bar, cycle), whose boundaries lie on so-called strong beats, and are indicated by vertical bars in Western notation. Over these longer time spans, the strong beats are also known as downbeats, whereas the remaining beats are weak in possibly a variety of degrees — these degrees of accentuation are defined in the metrical sense and do not necessarily correspond to note attacks (i.e., onsets) of varying loudness [3]. In Western music, two kinds of meters are commonly found — duple or triple —, depending on how the basic pulsation is subdivided. Our example shows a duple meter, and the time signature $\mathbf{\frac{2}{4}}$ indicates the number of notes of a particular value (in our case, the quarter note) featured in each bar.

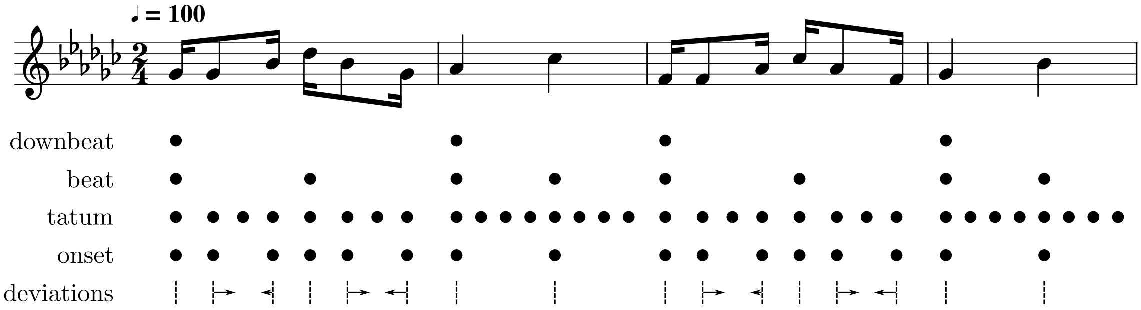


Figure 8.1: Musical concepts concerning rhythm overlaid on the score from Figure 2.13a. Arrows illustrate timing deviations: directions indicate whether notes are played ahead or behind the expected (quantized) positions, and lengths show the magnitude of the deviation.

Meter perception is influenced by the listener's musical training and cultural background [195]. For example, different listeners might perceive "beat" pulsations that lie in different metrical layers, leading, for example, to perceived tempi related by a factor of two [196].

Located at the fastest sub-beat level perceived in music, tatums ("temporal atoms" [171]) are generally dependent on local inter-onset intervals. The tatum rate, which is the tempo analog at this scale, is thus hardly ever constant throughout a piece. Expressive deviations may shape time and shift onset positions at this small scale in ways that Western music notation cannot accurately represent. We note that the systematic use of these deviations, which we will refer to as microtiming, is of structural importance in defining the style of many music genres. This is the case of jazz [197–200], Cuban *rumba* [171], Brazilian *samba* [201, 202], and Uruguayan *candombe* [203].

Finally, as alluded to in Chapter 1, one of the acceptations of the word "rhythm" is the specific patterned configuration of note attacks in time — the "perceivable pattern of temporal space between attacks" [3]. In the example, we can readily identify two phrases that share the same rhythmic pattern (short–long–short–short–long–short–long–long). It is true that the rhythmic cell "short–long–short" (the "characteristic" syncope) is also a relevant pattern in this piece. In our experiments that follow, this particular aspect concerning the duration or extent of the rhythmic pattern will not be taken into account.

8.2 Literature Review

In this section, we will explore various methods proposed in MIR for describing rhythm, with a specific focus on works that emphasize rhythmic similarity. The task of beat tracking, which COCHARRO et al. [15] mentioned in their review, is a crucial aspect of transcription and constitutes an entire field within MIR. Thus, it will receive special attention in the following chapter.

Most early approaches to rhythmic description and similarity involved computing periodicities from features sensitive to musical tempo (e.g., signal similarity matrices, band-wise autocorrelations, inter-onset interval histograms). Since small tempo changes are common not only within music genres but amid performances, due to physical limitations or even expressive intents and conventions, this tempo-dependency of the rhythmic features hinders the performance of most systems with respect to the assessment of similarities. Therefore, some works required prior estimation of the beat frequency and sometimes even of the beat phase (cf. Chapter 9) to avoid the influence of tempo in the final representation. For instance, with knowledge of bar boundaries in a given music recording, we can extract bar-length rhythmic patterns from a quantized representation of the bar interval at a fine temporal grid [38] — the influence of tempo is discarded, allowing for the comparison or bar patterns that were executed at different speeds. To bypass this problem, tempo-robust features were later specially targeted for retrieval and classification problems.

Early Approaches

FOOTE and UCHIHASHI [204] introduced the “beat spectrum”, which was computed by summing the diagonals of the signal similarity matrix at different lags. Its short-time version, the “beat spectrogram”, can be calculated over time using a sliding window. TZANETAKIS and COOK [205] analyzed the audio signal in octave bands using a discrete wavelet transform and extracted a smooth envelope of the rectified signal in each band. Then, they computed an enhanced autocorrelation on the sum of all frequency bands and selected its dominant peaks, which resulted in a feature called “beat histogram”.

DIXON et al. [206] represented rhythmic patterns by extracting, from each measure of a track (known a priori), an RMS envelope at a rate of b samples per bar with a hop length directly proportional to the bar duration. Best results were reported for $b = 72$ and a hop of 50%. This feature worked as the ODFs described in the previous chapter: assuming high values near onsets, low values otherwise. Each piece was then represented by the centroid of the most significant k-means cluster. They achieved a 50.1% classification rate of genres in the Ballroom dataset.

Another early approach, by GOUYON et al. [207], used a total of 73 descriptors to classify Ballroom genres. These included features derived from an inter-onset interval histogram (IOIH) and from a periodicity histogram inspired by [205]. They experimented with different subsets of features, but a classification system using IOIH-based features alone achieved a 51.2% accuracy, whereas 56.7% accuracy was achieved when using a set of features derived from the periodicity histogram. We note that the IOIH-based features required the estimation of the tatum of each piece, yielding a tempo-independent feature. They also reported a maximum genre classification accuracy of 90.1% when global (average) annotated tempo was used as one of the feature dimensions (78.9% with estimated tempo).

To circumvent the influence of tempo, PAULUS and KLAPURI [208] first estimated the metrical structure of the musical signal and then compared rhythmic patterns extracted at that grid resolution using dynamic time warping. This dynamic programming algorithm allowed measuring the similarity between two musical pieces even if they were executed at different speeds by finding an optimal match under certain restrictions.

PEETERS [209] estimated periodicities from an onset strength function by multiplying the magnitude of the DFT with a frequency-mapped ACF. A tempo-independent rhythmic representation was obtained by normalizing the feature by a local tempo estimate or by the annotated global tempo. Peeters exemplified the discriminative power of this feature with a genre classification experiment with regression methods. The system achieved a recognition rate of 80.8% in the Ballroom genres, and 90.4% when tempo information was added. One should note that 78% accuracy was obtained with tempo alone [209].

Avoiding Meter and Tempo Estimation

Other works extracted tempo-robust features without the need for estimating meter or tempo altogether. HOLZAPFEL and STYLIANOU [210] employed a dynamic periodicity warping (DPW) technique as a dissimilarity measure to directly compare periodicity spectra, i.e., the DFT of the onset strength signal (OSS), which is the signal described by an ODF. Later, authors used the scale transform to produce a descriptor that is robust to tempo variations [211, 212]. The choice for the scale transform representation came from the observation that the ACF of a scaled signal (e.g., the same rhythmic pattern at a different tempo) is equal to the autocorrelation of the original signal scaled by the same factor [212]. In other words, the transformation compensates for small scaling factors in the periodicity representation of two similar rhythms with different tempi. To obtain their descriptor, they first computed the local autocorrelation of the OSS. Then, they applied the scale transform to each analysis frame. Finally, features were averaged over time and used for comparisons through the cosine distance. A genre classification accuracy of 91.7% was obtained with an SVM classifier on the Ballroom dataset using this feature, named “scale transform magnitudes” (STM). MARCHAND and PEETERS [213] extended this idea in the “modulation scale spectrum”, which analyzed band-wise periodicities in the scale domain. Authors later included auditory statistics (cross-correlations of energy profiles in different bands) achieving an average recall of 96% in Ballroom classes with SVM [214].

Other approaches obtained similar results to the STM by normalizing for tempo via the transformation of the lag-axis of autocorrelation-based features into the logarithmic domain [215, 216]. In the log-scale, the effect of tempo changes on the autocorrelation function is not scaling, but translation, which can be eliminated by further processing (e.g., cross-correlation, DCT).

Modulation-Based Features

PAMPALK et al. [217, 218] described a set of features known as “fluctuation patterns” (FP), which also allowed for a multi-band analysis of rhythmic periodicities. These were defined by the amplitude modulation of the loudness of an audio signal in each frequency band. The original proposal included a set of psychoacoustic-related transformations such as, for example:

- bundling STFT frequency bins into critical bands (Bark scale);
- emulating masking effects by suppressing spectral components;
- translating sound pressure levels into a perceptive domain (sone scale).

Moreover, modulations were computed for each channel with the DFT and weighted according to a perceptual model of fluctuation strength. Finally, the output was filtered to emphasize beat information and smoothed to improve similarity retrieval. Note that this computation of per-channel periodicities is not robust to large tempo variations, although somewhat insensitive to small-scale tempo changes as a consequence of all the postprocessing steps.

Many modifications have been proposed to this set of features. For example, LIDY and RAUBER [219] aggregated FPs over frequency bins to produce a descriptor known as “rhythm histogram” (RH). POHLE et al. [220] drew inspiration from FPs and defined the “onset patterns” (OP), which used a cent-scale representation as input and mapped the linear-scale modulations onto a log-axis at the output. Through this transformation, tempo differences between recordings are represented by translations in the log-modulation-frequency axis. Authors further processed OPs by applying a two-dimensional DCT and obtained another feature, the “onset coefficients” (OC). The DCT operates by discarding phase information, further improving the robustness to tempo. They show that the classification of Ballroom styles can be improved from 75% with FPs, to 86.7% with OPs, and 87.7% on average with OCs, depending on the number of coefficients of the DCT in the frequency and periodicity dimensions. ABRASSART and DORAS [221] follow the original implementation by PAMPALK et al. [217, 218], but replace the second stage for a CQT to obtain the “constant-Q fluctuation patterns” (CQFP).

FOROUGHMAND and PEETERS [222] introduced a feature denoted "harmonic-constant-Q-modulation", which was computed by extracting a harmonic CQT [223] from a multi-band onset strength signal. This feature was used as the input of a convolutional neural network for global tempo estimation and rhythm pattern classification, achieving modest results in the latter task.

HOLZAPFEL et al. [224] and PANTELI and DIXON [225] presented a comparison of rhythmic descriptors based on the scale transform and fluctuation/onset patterns. In particular, HOLZAPFEL et al. [224] show that, depending on the tempo distribution of a music genre, it can be advantageous to use tempo-robust features or to encode large tempo changes in the representation when computing rhythmic similarity.

8.3 Rhythmic Descriptors

We now describe the rhythmic descriptors that will be used in Chapter 11. While the first descriptor remains mostly unmodified from the STM implementation of HOLZAPFEL and STYLIANOU [212], we propose a few important modifications for the OP descriptor of PAMPALK et al. [217, 218]. Namely, we (1) compress the amplitudes of the base representation and (2) use of a CQT for the computation of periodicities.

We derive our rhythm descriptors from the same log-amplitude mel spectrogram. To obtain this base representation, we first resample the audio tracks to 8000 Hz. Then, a short-time Fourier transform (STFT) of the signal segmented by overlapping sequential 32-ms Hann windows is calculated to produce a 50-Hz frame rate spectrogram. Finally, we map the frequency bins to a 40-band mel scale and take the logarithm to represent amplitudes in the dB scale.

Scale Transform Magnitudes (STM)

To extract this tempo-robust descriptor we follow the original proposal of [212]. First, we compute a spectral flux from the mel-scaled spectrogram. This is possible via first-order differentiation and half-wave rectification of each mel band, followed by the aggregation of all bands. We detrend the resulting OSS by removing the local average with the following difference equation

$$y[n] - 0.99y[n-1] = x[n] - x[n-1], \tag{8.1}$$

where $y[n]$ is the "detrended" OSS. We also smooth the resulting signal with a one-dimensional Gaussian filter with a standard deviation of 20 ms. Then, we determine the short-term autocorrelation of the OSS with a moving rectangular window of length 8 s and 0.5 s hop. Each frame of the ACF is transformed into the scale domain by the direct scale transform [90] using a resolution of [212]

$$\Delta c = \frac{\pi}{\ln \frac{T_{\mathrm{up}}+T_{\mathrm{s}}}{T_{\mathrm{s}}}} = 0.52, \tag{8.2}$$

where $T_{\mathrm{up}} = 8$ s is the maximum retained lag time and $T_{\mathrm{s}} = 0.02$ s the sampling period in our case. We limit our representation to the first 400 scale coefficients (up to scale $C = 400\Delta c = 208$). At the final step, we average this feature over time, achieving a dimension of 400 for each track.

Onset Patterns Histogram (OPH)

The other feature we use to compare audio excerpts is our tempo-sensitive descriptor that draws mostly from [217, 218], but also [220, 224–226]. We will refer to this simply as the “onset pattern histogram”. To extract our OPH descriptor, we first subtract, from each mel channel, the moving average computed with a normalized rectangular window of length 0.25 s, and half-wave rectify the result. This “unsharp mask” also has an effect of amplitude normalization, since the spectrogram is represented in dB [226].[1] At the second stage, instead of using the FFT to obtain per-channel modulations and mapping them to a log-frequency scale [220, 224], we compute a CQT of the signal in each channel and take its magnitude as in [221]. Like previous works, we define a minimum modulation frequency of 0.5 Hz (30 bpm). Periodicities are described in 25 bins, at five bins per octave, up to 14 Hz. Similar to [219, 225], we average the periodicities over all channels and take the mean feature across all time frames. This results in a descriptor with a dimension of 25.

8.4 Distance Metric

Several metrics have been proposed for evaluating rhythmic similarity with the descriptors from the STM and modulation-based families: Euclidean distance [217, 218, 220, 224], cosine distance [211, 212], correlation [225], Mahalanobis distance [225], Jensen–Shannon divergence [220].

In particular, for the STM descriptor, HOLZAPFEL and STYLIANOU [212] discuss how the Euclidean distance is not really applicable due to an unknown energy normalization factor $\sqrt{a}$ in the scale magnitude that is different for each recording. The angle between STM representations should be used to compare representations instead, which can be expressed by the cosine distance. If $\mathbf{x}$ and $\mathbf{y}$ are two vectors,

[1] POHLE et al. [220] and HOLZAPFEL et al. [224] apply the logarithmic scaling on amplitudes after the unsharp mask, while PANTELI and DIXON [225] skip this step.

the cosine similarity between them is computed as

$$s_{\cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|}, \tag{8.3}$$

where $\cdot$ is the dot product, and the cosine distance is given by

$$d_{\cos}(\mathbf{x}, \mathbf{y}) = 1 - s_{\cos}(\mathbf{x}, \mathbf{y}). \tag{8.4}$$

Following the recommendation of [212], we will use this metric for computing the distance between rhythmic representations derived from the STM feature. The cosine distance will also be used for comparing OPHs.

8.5 Qualitative Analysis of Rhythmic Descriptors

In this section, we briefly demonstrate the properties of the STM and OPH as rhythmic descriptors through a few examples.

Synthetic Patterns

For the first example, we use the rhythmic cell of *teleco-teco* (Figure 2.19c) as a basis pattern. We consider only the pattern generated by the drumstick striking the drumhead, discarding the soft notes played with a finger. We artificially synthesize six different versions of this pattern, the first five using a *tamborim* note sample:

- $x_{\text{ref}}[n]$ (reference signal), the pattern synthesized at a tempo of 100 bpm;
- $x_{\text{a}}[n]$, the pattern synthesized at a tempo of 120 bpm;
- $x_{\text{b}}[n]$, a simplified version of the pattern, with some notes removed — this pattern corresponds to a different subdivision of note groupings;
- $x_{\text{c}}[n]$, a version of the pattern with added microtiming deviations, ranging from 1% to 11% of the inter-beat interval (IBI);
- $x_{\text{d}}[n]$, a version of the pattern that includes dynamic variations;
- $x_{\text{e}}[n]$, a version of the pattern synthesized with a different instrument (using an *agogô* note sample).

All synthesized versions share a duration of 30 s and a sampling rate of 44 100 Hz. Figure 8.2 shows the first few seconds of the reference signal and its alternate versions.

We then compute the STM and OPH for each of these signals, following the procedure of Section 8.3. The results are shown in Figure 8.3, up to the 200th scale

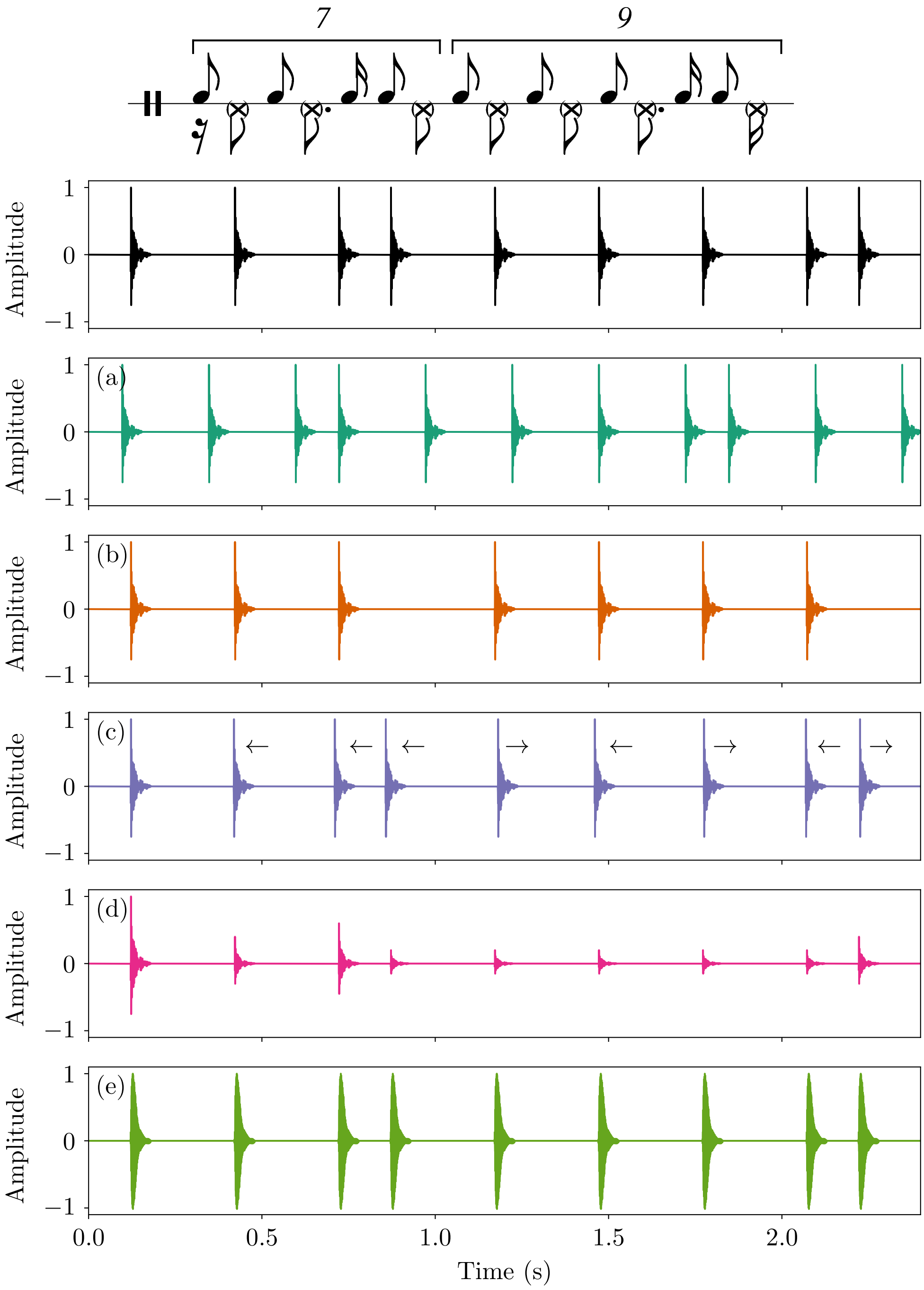


Figure 8.2: Signals derived from the *teleco-teco* timeline, from top to bottom: $x_{\mathrm{ref}}[n]$, the reference signal (tempo = 100 bpm); (a) $x_{\mathrm{a}}[n]$, at a tempo of 120 bpm; (b) $x_{\mathrm{b}}[n]$, a simplified version (missing onsets); (c) $x_{\mathrm{c}}[n]$, with microtiming deviations; (d) $x_{\mathrm{d}}[n]$, including dynamic variations; and (e) $x_{\mathrm{e}}[n]$, synthesized from a different sample (*agogô*).

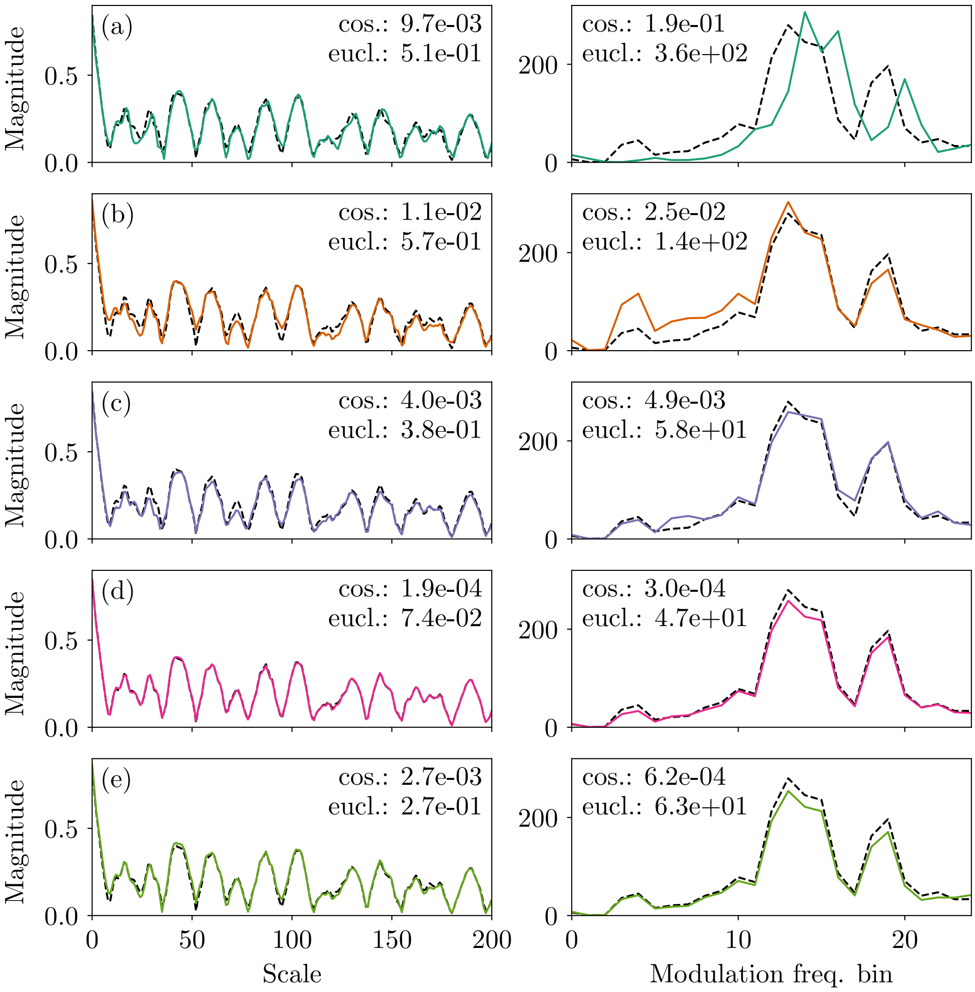


Figure 8.3: STM (left) and OPH (right) coefficients for the signals in Figure 8.2. Representations of the reference signal $x_{\mathrm{ref}}[n]$ (black, dashed) are displayed against the representations for other variants (in solid colors), from top to bottom: (a) $x_{\mathrm{a}}[n]$ (tempo = 120 bpm); (b) $x_{\mathrm{b}}[n]$ (simplified); (c) $x_{\mathrm{c}}[n]$ (w/ microtiming); (d) $x_{\mathrm{d}}[n]$ (w/ dynamic); and (e) $x_{\mathrm{e}}[n]$ (*agogô*). For each representation, we also display the cosine and Euclidean distances between the representation for $x_{\mathrm{ref}}[n]$ and that of the corresponding variant.

coefficient for STM and all modulation frequency bins for OPH. We also display Euclidean and cosine distances between the representations for the reference signal and its different versions. Our comparative analysis is based on the cosine distance.

Comparing the representations for $x_{\mathrm{ref}}[n]$ and $x_{\mathrm{a}}[n]$, we can see how the different descriptors deal with tempo differences. STM is insensitive, as expected, whereas OPH displays the greatest cosine distance to the reference across all versions. Moreover, we observe that peaks in the OPH representation shift right following the increase in bpm. For instance, the rightmost peak lies at bin 19 (∼6.96 Hz) for the reference signal and at bin 20 (8 Hz) for the scaled version. These correspond to the smallest IOI in the synthesized signals, 0.15 s and 0.125 s, respectively.

We now turn our attention to the representations for $x_{\mathrm{b}}[n]$. Its STM representation shows the greatest distance (in cosine sense) to the reference. By removing that specific onset from the original pattern, we have also removed a set of peaks from the autocorrelation function, particularly the one that has the smallest lag. The difference in the representations is more significant at higher scale values (above the 200th coefficient). OPH, on the other hand, is less affected by this variation, but we can observe an overall increase in the energy of the lower modulation frequencies; this matches our intuition about the nature of the periodicities after the removal: more weight is transferred to lower frequency components.

Microtiming deviations (c) had little effect on either representation. In the case of the STM descriptor, this is by design and can be explained by the smoothing operation (convolution with a Gaussian kernel). The Gaussian has a standard deviation of 20 ms, which is greater than the maximum shift of 11% of the IBI (6.6 ms at a tempo of 100 bpm) that onsets have been subjected to. In OPH, we note that energy has shifted slightly between neighboring bins in this version of the pattern. Similarly, neither dynamic variation (d) nor the timbre change (e) had a noticeable effect on the representations — this indicates that only the temporal distribution of the pattern is being analyzed.

Number of STM Coefficients

In the second example, we use the STM descriptor to represent all solo tracks in the BRID dataset. First, we evaluate the influence of the number of scale coefficients in a subgenre classification task, exploiting the fact that we have different subgenres among BRID solos (e.g., *samba*, *samba-enredo*, *partido-alto*). To simplify this problem, we aggregate tracks containing *viradas* (e.g., labeled as VSA) in the main subgenres (e.g., SA). We vary the number of scale coefficients in $\{3, 10, 50, 100, 200, 300, 400\}$ and perform the classification with a linear SVM, choosing the weights inversely proportional to class frequencies due to natural imbalances in the subset. With a linear SVM we can have an idea of how linearly separable

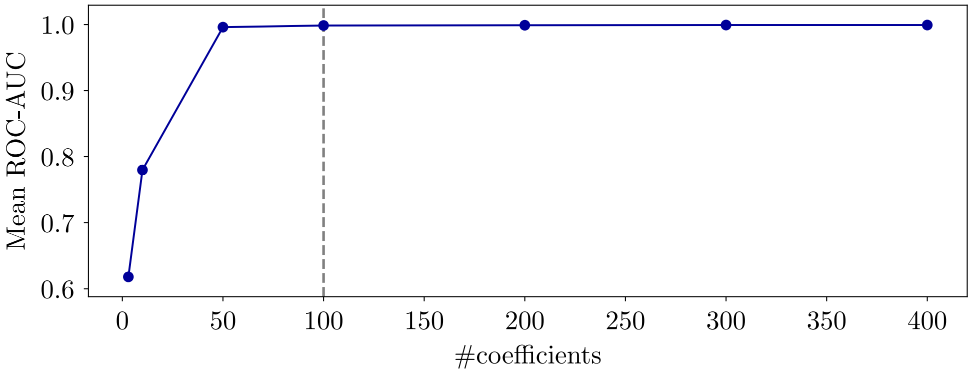


Figure 8.4: Average ROC-AUC scores for the cross-validated classification of genres with a varying number of coefficients.

the classes are in the high-dimensional feature space. We perform this classification process three times using a stratified cross-validation scheme with 10 folds. A different randomization is used in each repetition. The average ROC-AUC score is reported, with classes being compared in a one-vs-rest (OVR) strategy for simplicity. This score is determined by calculating the area under the receiver operating characteristic (ROC) curve, which shows the relation between true positive and false positive rates at various binary classifier thresholds. The closer the value is to 1, the higher the performance of the classifier. We note that there is some data "leakage" in this approach to the classification problem, as train and test sets might contain neighboring frames of the representation from the same recording, with a lot of overlap. Nevertheless, our objective is to evaluate how the similarity is affected by the number of coefficients of the STM. The results for this classification can be seen in Figure 8.4. We observe from this graph that, for this subset, the classification is barely improved when the representation has more than 100 coefficients, where the ROC-AUC score reaches 0.999. We will use this number of coefficients (up to $C = 52$, at the optimal resolution) in the following example.

UMAP Visualization

As regards the last example, to visualize the feature space generated by the STM in BRID solos, we used a dimensionality reduction technique called "uniform manifold approximation and projection" (UMAP) [227]. Unlike principal component analysis, UMAP does not expect a linear relationship between the different dimensions, instead focusing on learning the complex structure of the data. Moreover, a major advantage of UMAP over other commonly used dimensionality reduction approaches like t-distributed stochastic neighbor embedding (t-SNE) is that it can better represent global data structure while preserving local neighborhoods [228]. Once again we incorporate *viradas* into the main subgenres. Embedding results are shown in

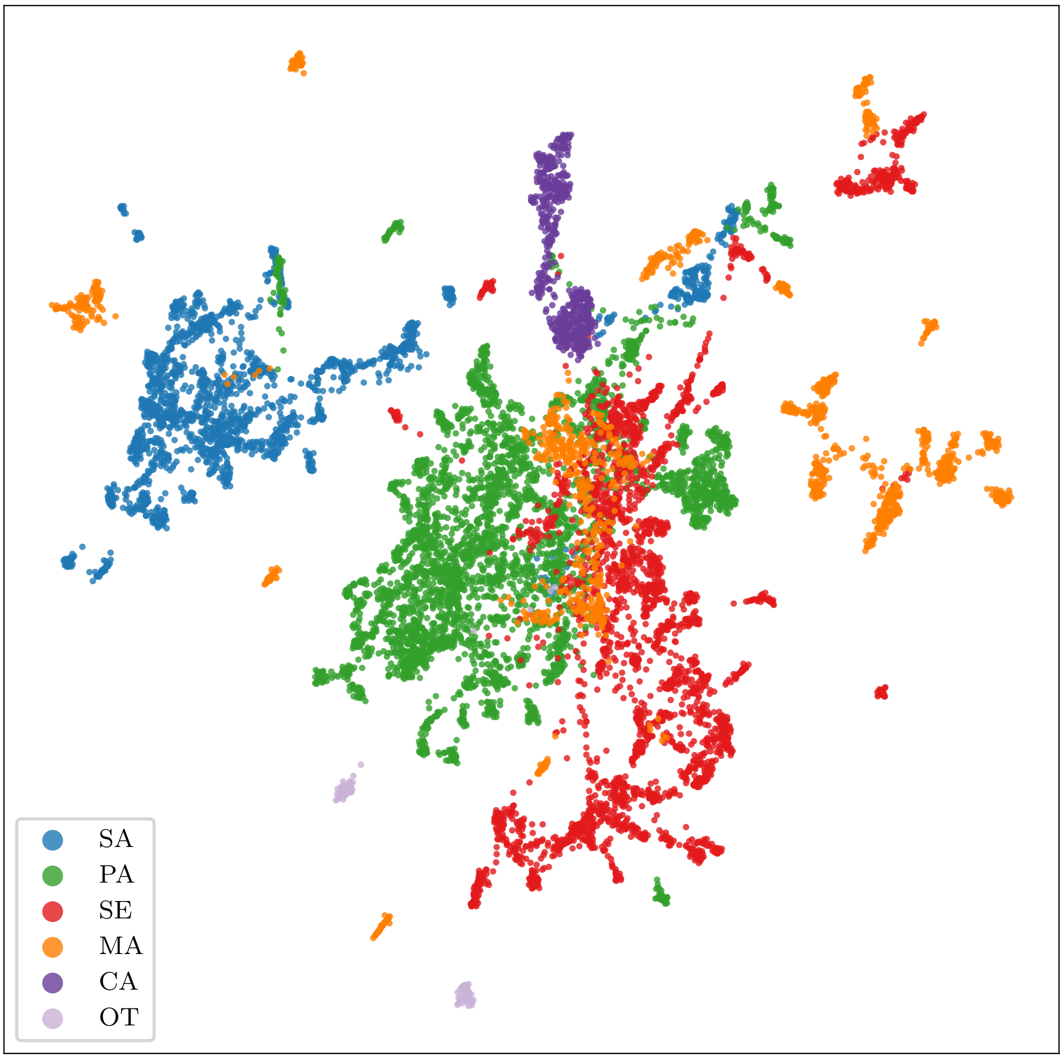


Figure 8.5: UMAP projection of BRID solos with the STM descriptor (cosine metric, n-neighbors $= 100$, min-dist $= 0.5$). Colored by genre.

Figures 8.5 and 8.6, with different colorings revealing genres and instruments, respectively. Instead of averaging the results along the time axis, we have represented each frame of the STM feature as a point in the embedding. We observe many interactions in the subset, but a fine structure also seems to emerge from the manifold visualization of Figure 8.5. The different subgenres have some particularities in common to most of their tracks, as evidenced by the large clusters of datapoints from SA and CA, for example. The two OT tracks, corresponding to patterns typically seen in *baião* and *maxixe* genres, are also separated in the embedding, as outliers. In the central area of the plot, there are many similar rhythmic cells from MA, PA, and SE — these mostly correspond to "fuller" cycles, with mostly all tatums articulated. Since the feature is somewhat insensitive to small-scale timing deviations, we cannot determine if this intersection of *samba-enredo*/*partido-alto* (which are performed at a high tempo, >100 bpm) and *marcha* can be attributed to the phenomenon of

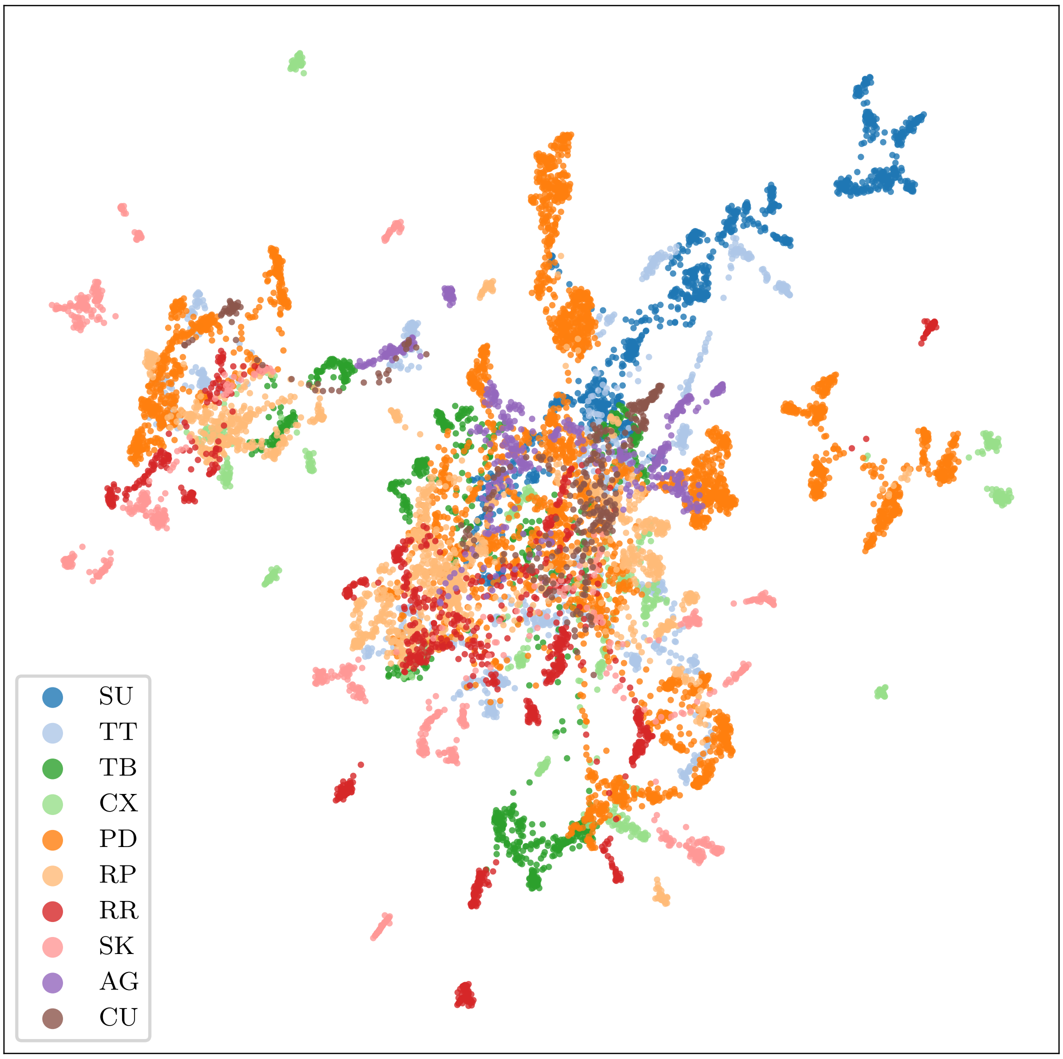


Figure 8.6: UMAP projection of BRID solos with the STM descriptor (cosine metric, n-neighbors = 100, min-dist = 0.5). Colored by instrument.

marcialização, i.e., performers are not able to impress the characteristic microtiming on the former genres.

By directly comparing this figure with Figure 8.6, we see that similar patterns are played in different instruments; this points to the existence of genre-specific cross-instrumental patterns. It is worth noting that most of the frames from *surdo* (SU) and *tantã* (TT) are located in close proximity to each other in the northeastern region of the graph. As we have mentioned before, these two instruments typically play similar parts, one in the *bateria* and the other in the *roda*. At the northeasternmost part of the figure, we mainly have recordings from these instruments playing beats "1" or "2". As we get closer to the central "mass" of the embedding, however, the patterns played by these instruments get increasingly complex — there we find most of *surdos de terceira* recordings as well as *tantã* parts with numerous embellishments.

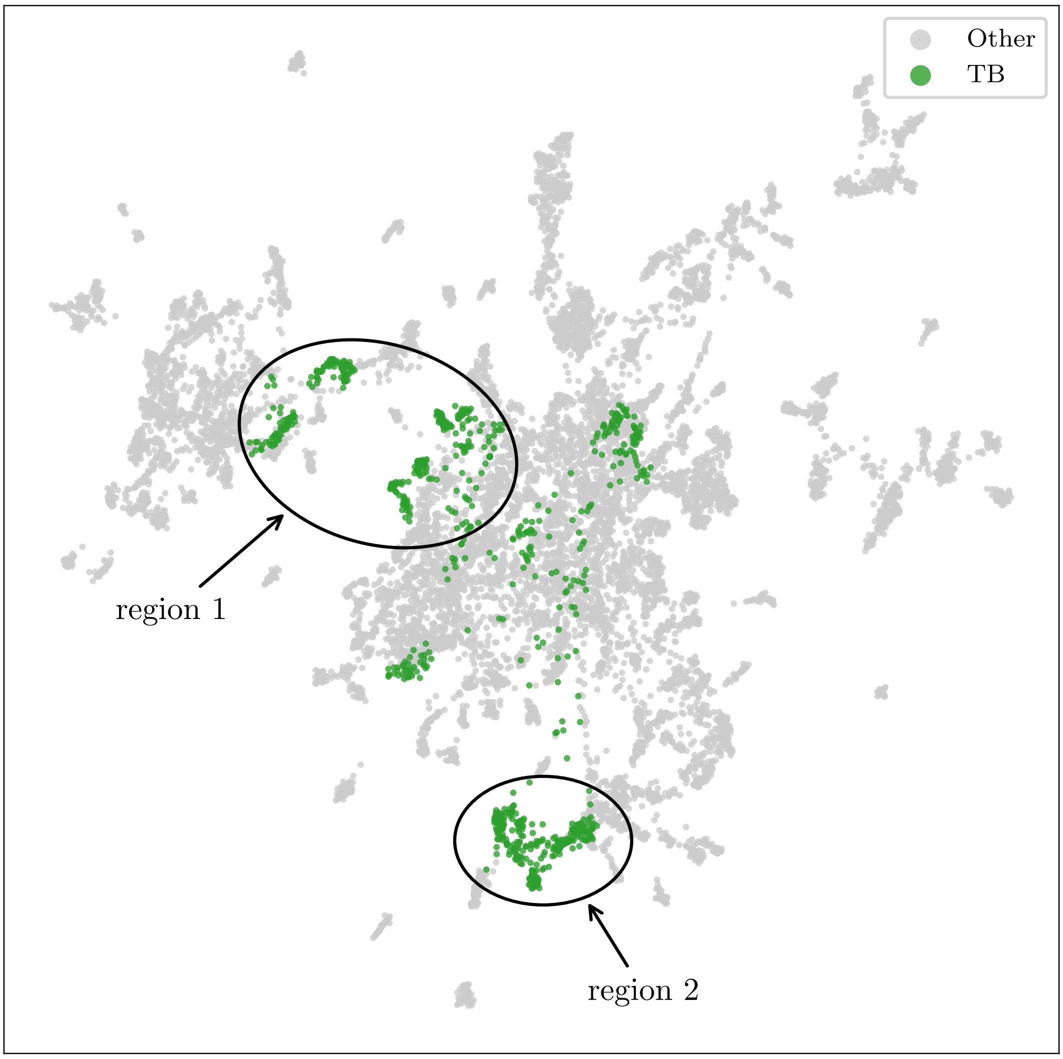


Figure 8.7: UMAP projection of BRID solos with the STM descriptor (cosine metric, n-neighbors $= 100$, min-dist $= 0.5$). Highlighting *tamborim*.

Figure 8.7 brings the same embedding plot, now highlighting STM frames from *tamborim* recordings. We call attention to two regions in this figure. Region 1 contains STM frames from recordings `[0129]`, `[0130]`, `[0134]`, `[0135]`, `[0214]`, `[0215]`, `[0302]`, `[0303]`, which display the styles of *samba* and *partido-alto*. The most prevalent pattern in these recordings is the *teleco-teco*. After listening to all of these examples, we have found out that the only distinguishing factor between these *teleco-teco* versions is that in SA tracks the soft note produced with a finger on the underside of the drumhead is quite prominent and widespread, whereas in PA tracks it is either absent or less noticeable. Region 2 represents the tracks `[0131]`, `[0132]`, `[0136]`, `[0216]`, `[0218]`, `[0304]`, and `[0306]`, which are all recordings of *carreteiro* cycles. We refer the reader to Figure 2.19 for the main differences between these two cycles. Other TB points scattered around the embedding mostly come from tracks that document more complex *viradas* (`[0217]`, `[0219]`, and `[0317]`).

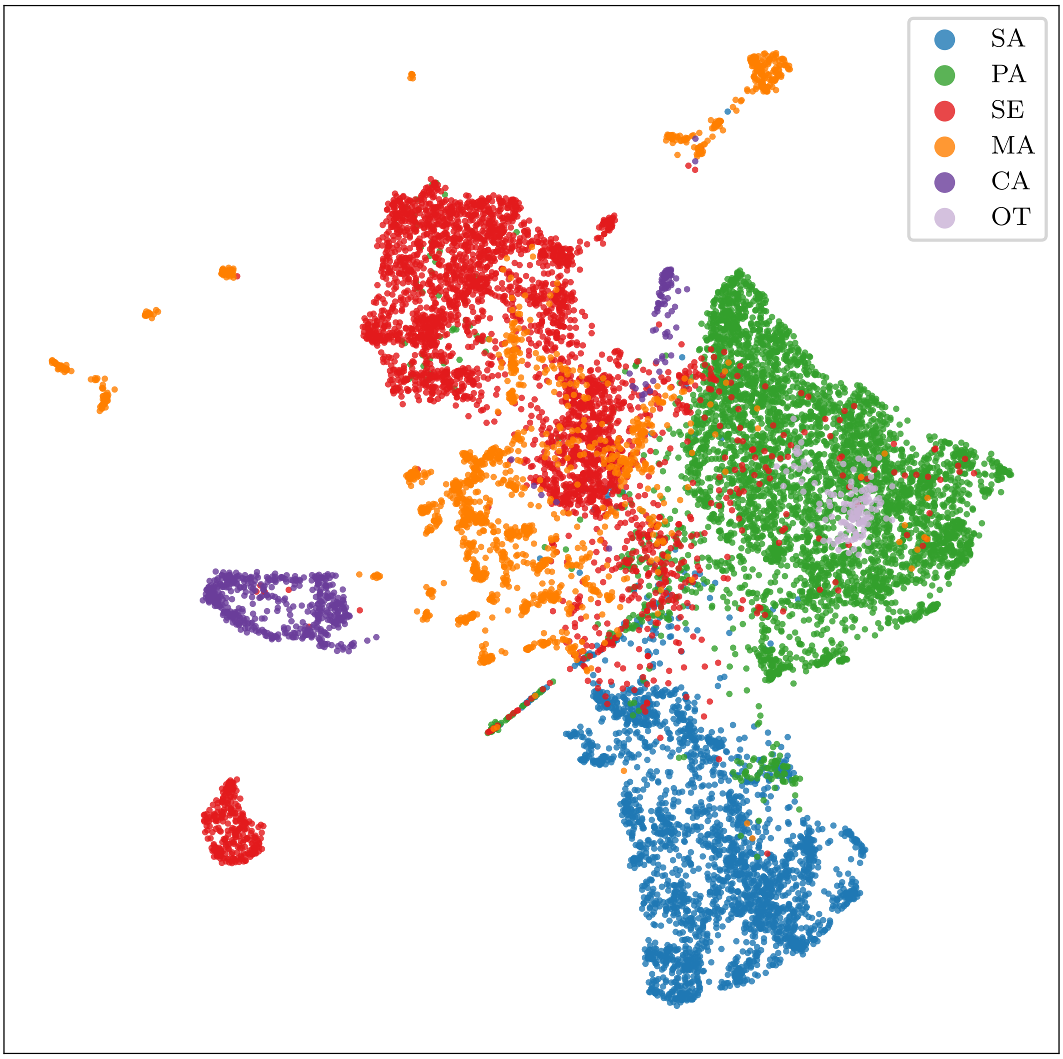


Figure 8.8: UMAP projection of BRID solos with the OPH descriptor (cosine metric, n-neighbors $= 50$, min-dist $= 0.3$). Colored by genre.

In Figure 8.8, we display an embedding for the framewise OPH feature of the same subset, colored by subgenre. Arguably, the subgenres are more separable in this representation. However, considering that OPH encodes tempo information, it provides a great clue for genre recognition in this subset, which was built with a different target bpm for each genre (except *viradas*). Finally, Figure 8.9 shows the projection using PCA instead of UMAP. As mentioned before, PCA is unable to capture the more complex structures in the rhythmic feature data.

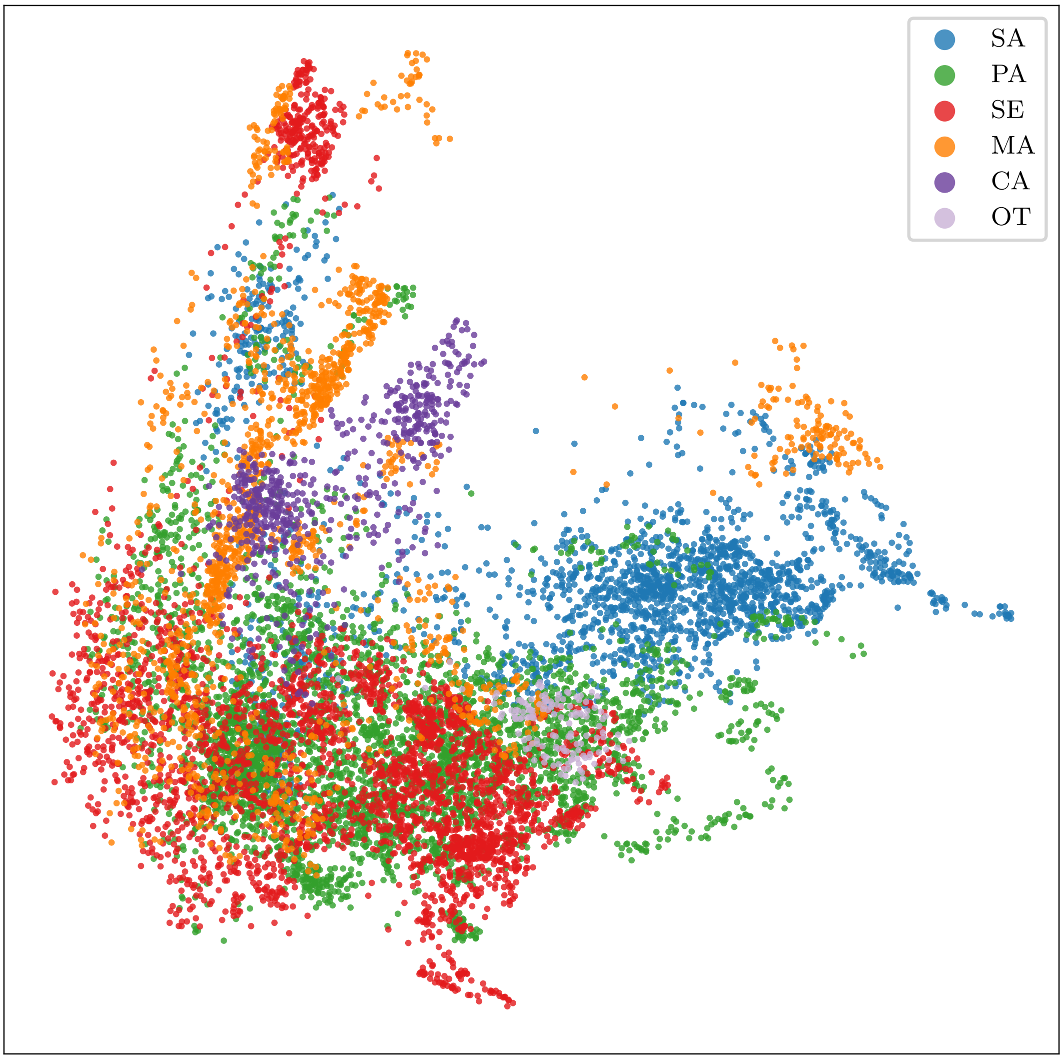


Figure 8.9: PCA projection of BRID solos with the STM descriptor. Colored by genre.

Chapter 9

Metrical Structure and Microtiming

In the previous chapter, we discussed some useful tools for describing rhythm in a compact form [188]. We will now delve into two related topics: beat tracking and microtiming analysis. Beat tracking is of particular importance in MIR due to the number of applications that benefit from the knowledge of beat positions or from beat-synchronous features. For instance, proper beat synchronization is of utmost importance in automatic accompaniment, score alignment, music transcription, DJ'ing (synchronization between stems/tracks), expressiveness transformations, music similarity analysis, and structural segmentation, among others. These two topics were not covered in Chapter 8; instead, they are addressed here in a dedicated chapter where we can better present the substantial amount of research devoted to them.

Beat tracking algorithms were first introduced in the 1980s [229]. This task involves the automatic detection of the most salient metrical level (tactus [12]), which is the pulse where a person would tap along with the music. Beat tracking is closely related to other rhythmic tracking tasks, such as tempo, meter, rhythmic pattern, and sub-beat/super-beat structure (i.e., faster/slower pulsations). Some works have attempted to track beat positions and these related properties, either jointly or separately. For instance, few works have simultaneously tracked the sub-beat level of the tatum, which is the fastest regular pulse that can be inferred from the music. Other levels such as sections and parts have been featured less often, but can also be automatically estimated [230]. More commonly, though, systems have attempted to jointly track beats and downbeats, which are indicators of measure boundaries. Downbeat tracking has also been carried out as a task in itself, i.e., independent of beat estimation, but relying on a different subdivision of the bar (usually the tatum [231–234]).

A simplifying assumption in many early approaches to beat tracking is that the tempo of a piece is constant or varies slowly, and that beat locations tend to coincide with note onsets. However, in expressive music, beats are rarely distributed isochronously, as interpreters shape tempo to convey different musical meanings (e.g., *rubato*, *notes inégales* [235]). Beyond this issue of timing, estimating the phase and period of beats and downbeats becomes more complex in music genres with syncopation, triplets, or even swing [76], which are elements common to many non-Western music genres, for example.

In the particular case of challenging expressive pieces and of recordings from non-Western music traditions, microtiming analyses are of great interest. As mentioned in Section 8.1, microtiming refers to small-scale temporal deviations that are systematic and genre-defining. These deviations occur at the onset level and have a composite origin, arising due to the player's own mechanical limitations and expressive intentions, including the aforementioned genre-specific characteristics. These analyses are also crucial when dealing with music synthesis, in humanization algorithms aiming at creating natural-sounding performances [236–238].

In this chapter, we will review the literature on both tasks to set the stage for further investigations and contributions in the remainder of this thesis.

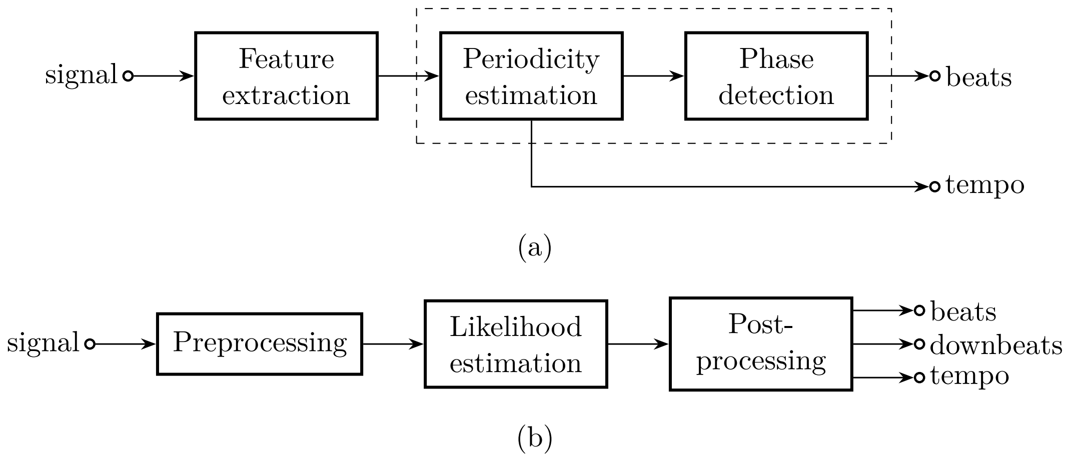
9.1 Approaches to Beat and Downbeat Tracking

This section reviews the main architectural choices for tracking beats and downbeats in musical audio signals. We also discuss current trends, and detail a few examples of algorithms tailored for this task. For more thorough reviews, we refer the reader to [5, 10, 229, 239–242] which either discuss early techniques in more detail or shine light on the more recent deep learning-based systems. In particular, the tutorial of DAVIES et al. [241] provides a good hands-on approach to tempo, beat, and downbeat estimation.

Early Approaches to Beat Tracking

Most of the early approaches have treated beat tracking as a problem of finding the frequency and phase of a time-varying signal [229]. The instantaneous frequency of this signal is proportional to the local tempo, whereas the phase helps determining beat positions. A solution for this problem has usually been obtained after a few common steps: first, low- or mid-level features are extracted from the audio signal; the periodicities are analyzed within the selected feature sequence; and, finally, the beat times are properly estimated. This general scheme of traditional beat tracking methods is presented in Figure 9.1a.

In many instances, little musical knowledge has been embedded into the beat-tracking model. One common assumption is that beats occur synchronously with



(a)

(b)

Figure 9.1: General (simplified) schemes for traditional beat tracking. Adapted from [243].

significant changes in the music (e.g., percussive onsets, harmonic changes) that should somehow be represented in the feature sequence. Thus, the choice of feature representation usually depends on the type of music at the input (e.g., when drums are present, an amplitude envelope could be beneficial), but also on the subsequent estimation steps [243]. Examples of low-level features include amplitude/energy envelopes, spectral features (e.g., spectral flux [244, 245], complex spectral difference [246]), and phase features (e.g., group delay [247]). Other works have leveraged mid-level features which comprise indications of chord changes [248] and lists of onset times [249, 250], among others.

For the analysis of periodicities in the input feature, most works have used techniques based on autocorrelation functions [208], comb filter banks [244, 246], beat histograms [251], and predominant local pulse functions [252]. The final step of detecting pulse phases has also been carried through in a variety of forms, including, but not limited to: multiple agent hypotheses [248], dynamic programming (DP) [245, 246, 252], and inference via HMMs [244, 247].

Downbeat Tracking and the Incorporation of High-Level Information

The problem of downbeat phase estimation has received far less attention in the early days of rhythm tracking systems. Few works have turned to the exploitation of preference rules, such as the fact that harmonic changes usually correlate well with measure boundaries in Western pop music [248]. Due to the more complex nature of downbeat tracking, other methods have incorporated even higher-level musical information. For instance, some works have attempted to match known bar-length patterns to the feature sequence [244, 253], or to hierarchically select downbeat positions from previously estimated beat data [254]. One hierarchical model, by DURAND et al. [255], leverages time signature estimation and musicallyinspired features (chord and pattern changes, bass-to-snare-drum ratio, and accents) with an SVM to classify annotated beats into downbeats. In contrast, JEHAN [256] reported a semi-automatic model in which an SVM classifies candidate events into downbeats using prior information learned from a human listener, without tempo or beat annotations.

Beat tracking has also been improved by introducing high-level knowledge into the models. Oftentimes, this has been done via a probabilistic formulation, such as an HMM, a dynamic Bayesian network (DBN), or a conditional random field (CRF) that integrates style-specific information (e.g., genre-specific tempo distribution or rhythm patterns) as hidden variables, and through which the beat period and phase are generally jointly estimated [76, 257–260].[1] This "joint model" is represented in Figure 9.1a by the dashed line. In one such model, known as the "dynamic bar-pointer model", the Bayesian formulation has also allowed to jointly track other rhythmic parameters — downbeats, meter, tempo, and patterns —, which were all integrated into the state space [76, 262, 263]. This paradigm has been originally presented by WHITELEY et al. [259] and later adapted by many authors [76, 196, 263–269]. Most of these systems are data-driven and capable of learning parameters (conditional distributions) from annotated audio input. This learning step has been achieved, for example, by modeling beat/downbeat likelihoods with a Gaussian mixture model (GMM) [76, 263, 265] or a different unsupervised clustering technique (e.g., k-means [63]) over a spectral-flux-like input. Whenever exact inference is feasible (e.g., in a coarsely discretized state space), the maximum posterior probability can be computed with the Viterbi algorithm [76]. Sometimes, however, exact inference is not possible or is too costly (e.g., in large state spaces), requiring either the simplification of variable dependencies or approximate inference using particle filters (PF) [196, 263, 267, 268]. We further investigate this model in Section 9.1.1.

A Change in Paradigm: Neural Networks

A new state of the art was inaugurated for both beat and downbeat tracking with the introduction of models based on deep neural networks, an approach first proposed by BÖCK and SCHEDL [243] for beat tracking. These networks allowed transferring the bulk of beat/downbeat determination to a preliminary step of likelihood estimation (feature learning), which can then be followed by a less complicated post-processing step [243]. The model architecture of this new paradigm is better illustrated by the diagram shown in Figure 9.1b.

At the input, neural-network-based systems use chroma features, spectral

[1] This has also been achieved outside a Bayesian framework under more restrictive (i.e., fixed tempo) assumptions [261].

flux [231–234, 270], or other representations such as timbre-related features (MFCCs) [231]. However, most commonly an audio input is preprocessed to obtain a time-frequency representation, which then passes through the network. These can be difference [243] or linear [231] spectrograms, or other log-scaled time-frequency representations — STFTs, CQTs, mel-spectrograms [231–234, 240, 271–276] — that are possibly presented to the network in different time-frequency resolutions [240, 243, 271]. In some cases, the representation is further processed to generate a tatum-synchronous feature sequence [231–234]; or beat-synchronous features [270] are used, when the model is only concerned with tracking downbeats. Networks can receive a single type of features [273, 274] or multiple features representing a set of musical attributes [231–234, 270]. Furthermore, the model can also be developed to learn hierarchically-related feature representations [275].

At the likelihood estimation stage, many different neural network topologies have been used, including multi-layer perceptrons [231], CNNs [232–234, 272, 277], recurrent neural networks (RNNs) [243, 265, 267, 270, 271], and convolutional recurrent neural networks (CRNNs) [239, 268, 269]. The state-of-the-art beat and downbeat performance has been achieved by models based on temporal convolutional neural networks (TCNs) [273–276] — an architecture that processes sequential data using dilated convolutions — and Transformers [278, 279] — an architecture that uses self-attention mechanisms to process entire sequences in parallel —, or a combination of both [278]. A TCN-based state-of-the-art tracking model is further reviewed in Section 9.1.2.

Finally, for the majority of the works that depend on neural networks, the post-processing stage, i.e., the selection of the most likely pulse candidates, has been tackled with probabilistic graphical models: either HMMs [231, 232, 234, 269, 272, 277], DBNs [239, 240, 270, 271, 274, 275, 278, 279], or CRFs [233, 280]. These act as constrained versions of the bar-pointer model, where global optima can be obtained through exact or approximate inference schemes (e.g., particle filters [267, 268]). Alternatively, few works have resorted to simple peak picking algorithms [243, 272], which typically yield inferior results. An interesting investigation, by CHEN and SU [276], showed that, if the network's architecture and loss function are carefully reformulated, the performance of a joint beat and downbeat tracking scheme can be improved even when no post-processing stage is used.

We end this section highlighting a couple of works that manage likelihood estimation in interesting ways. First, we comment on the multi-model of BÖCK et al. [240], in which multiple RNN models were trained and specialized on different genres. Each recurrent network consisted of a concatenation of three bi-directional long short-term memory (BLSTM) hidden layers with 25 units per layer. At runtime, the model with the most adequate activation function was selected to produce a beat likelihood. The adequacy of each model was computed as the mean square difference to the output of a reference model trained on the whole training set. Tempo and beat phase were determined using a DBN. A joint beat and downbeat tracking system was later presented by BÖCK et al. [271]. The authors used three different magnitude spectrograms and their first order difference as input representations, in order to help the networks capture features precisely in both time and frequency. These representations were fed into a cascade of three fully-connected BLSTMs, obtaining activation functions for beat and downbeat as output. Subsequently, a highly constrained DBN was used for inferring the metrical structure. In another work, by KREBS et al. [270], a downbeat tracking system was proposed using as input two beat-synchronous features, which represented the percussive and harmonic content of the audio signal. These representations, based on spectral flux and chroma, were then fed into two independent bi-directional gated recurrent units (BGRU), a different type of RNN similar to LSTMs, whose output was averaged to obtain the downbeat likelihood. Once again, the inference over downbeat candidates relied on a constrained DBN. Lastly, in the work presented by DI GIORGI et al. [277], a different CNN architecture, exploiting scale-invariant convolutional layers, was used for downbeat tracking. These layers learn temporal patterns from the data without caring for their scale. The model was shown to achieve tempo invariance, thus generalizing well to unseen tempi during training. This is a particularly important contribution, since conventional CNN-based models might suffer from learning bias due to uneven tempo distributions in the annotated data [277].

9.1.1 Dynamic Bar-Pointer Model

The dynamic bar-pointer model is a Bayesian formulation that allows tracking not only the beat pulsation of a music track, but also more complex rhythmic parameters (tempo and meter). These parameters are treated as a sequence $\mathbf{x}_{1:K}$ of hidden states in a latent space inference problem, which generates the observed sequence of features $\mathbf{y}_{1:K}$ from the audio data with length equivalent to K frames. In this model, which is generally expressed as a DBN, the joint distribution of hidden and observed variables, $P(\mathbf{y}_{1:K}, \mathbf{x}_{0:K})$, can be factorized as

$$P(\mathbf{y}_{1:K}, \mathbf{x}_{0:K}) = P(\mathbf{x}_0) \prod_{k=1}^{K} P(\mathbf{x}_k|\mathbf{x}_{k-1})\, P(\mathbf{y}_k|\mathbf{x}_k), \tag{9.1}$$

where $P(\mathbf{x}_0)$ is the initial state distribution, and $P(\mathbf{x}_k|\mathbf{x}_{k-1})$ and $P(\mathbf{y}_k|\mathbf{x}_k)$ are the transition and the observation models, respectively. We note that $P(\mathbf{x}_0)$ is usually assumed to be uniform, although it can be learned from data or manually set using a priori knowledge about the music under study.

The main idea behind this Bayesian formulation of the beat tracking problem is that, by explicitly modeling these three concepts (beats, tempo, and meter), one can resolve the ambiguities in the musical structure, and adequately deal with changes in the underlying rhythmic pattern [259]. In the following, we present a general definition of the bar-pointer model inspired by both [196] and [266]. Other versions with discretized variables or other dependency relations can be seen in [76, 263–265, 267–269]

Variable Definition

The bar-pointer model tracks the dynamics, $\mathbf{x}_k$, of a hypothetical "bar pointer", which traverses each measure (or cycle) of the music at a velocity proportional to the local tempo. Here we define the hidden variable $\mathbf{x}_k = [\phi_k, \dot{\phi}_k, r_k]$ at time frame k, where:

- $\phi_k \in [0, M_{\max})$ represents the current position in the bar. It increases from 0 to $M_{\max}$ and goes back to 0 at the start of a new bar.
- $\dot{\phi}_k \in [\dot{\phi}_{\min}(r_k), \dot{\phi}_{\max}(r_k)]$ is the velocity at which the pointer progresses through the bar, i.e., instantaneous tempo, given in bar positions per time frame (or sometimes in bpm). Tempo limits $\dot{\phi}_{\min}(r_k)$ and $\dot{\phi}_{\max}(r_k)$ can be dependent on the rhythmic pattern and learned from data, or simply set by the user.
- $r_k \in \{1, \ldots, R\}$ is the index of the current bar-length rhythmic pattern, and indicates which of the R observation models learned from data or defined by the user is being traversed. Different rhythmic patterns might have different lengths, M_r — it is common to set $M_{\max}$ with the length of the longest pattern and scale other lengths accordingly.

Transition Model

The factorization of the transition model is contingent upon the conditional independence relations between hidden variables. For example, given the model presented in [76] (see Figure 9.2), we can write

$$P(\mathbf{x}_k|\mathbf{x}_{k-1}) = P(\phi_k|\phi_{k-1}, \dot{\phi}_{k-1}, r_{k-1}) \times P(\dot{\phi}_k|\dot{\phi}_{k-1}, r_{k-1}) \times P(r_k|r_{k-1}). \tag{9.2}$$

Simplified variable dependencies have been proposed in other works, including [263].

The position transition is defined so that the bar position increases consistently, at a rate given by the instantaneous tempo at the previous position, $\dot{\phi}_{k-1}$. We could express this as

$$P(\phi_k|\phi_{k-1}, \dot{\phi}_{k-1}, r_{k-1}) = \mathbb{1}_\phi, \tag{9.3}$$

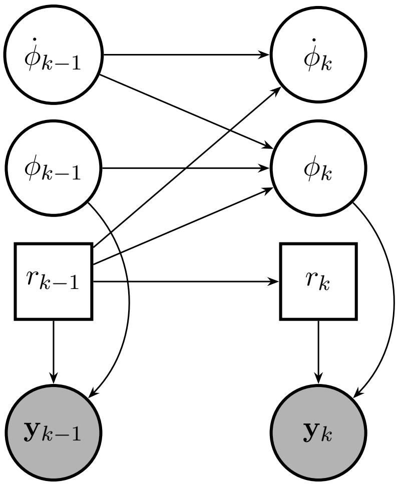


Figure 9.2: DBN of the bar-pointer model. Squares and circles designate discrete and continuous variables, respectively. Gray nodes are the observed variables, and white nodes are part of the hidden state. Arrows indicate direct dependency relations. Adapted from [76].

where this indicator function equals one when $\phi_k = (\phi_{k-1} + \dot{\phi}_{k-1}) \bmod M_{r_k}$, for instance. The modulo operator is used to guarantee that ϕ_k is reset when it exceeds M_{r_k} (i.e., the bar pointer crosses a bar boundary).

The tempo transition can be specified as a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean of $\dot{\phi}_{k-1}$ and standard deviation $\sigma_{\dot{\phi}}$, which is possibly dependent on $\dot{\phi}_{k-1}$, such that [196, 263, 266]

$$P(\dot{\phi}_k|\dot{\phi}_{k-1}, r_{k-1}) \propto \mathcal{N}(\dot{\phi}_{k-1}, \sigma^2_{\dot{\phi}}) \times \mathbb{1}_{\dot{\phi}}, \tag{9.4}$$

where this indicator function guarantees $\dot{\phi}_k$ stays inside the allowed tempo range. Other transition forms are presented in works dealing with discrete tempo variables. For example, complementary probabilities can be assigned for the three disjoint cases when the bar pointer remains at the same velocity, is accelerated, or is decelerated [76, 259, 264]. Some works have presented the tempo transition as a Laplace distribution, also symmetrical about $\dot{\phi}_{k-1}$ [265, 272],

$$P(\dot{\phi}_k|\dot{\phi}_{k-1}, r_{k-1}) \propto \mathrm{e}^{-\lambda\left|\frac{\dot{\phi}_k}{\dot{\phi}_{k-1}}-1\right|}, \tag{9.5}$$

which is controlled by the parameter $\lambda \in \mathbb{Z}^*$ and only allowed at estimated beats positions. We note that λ controls the steepness of the distribution [265], with $\lambda = 0$ indicating tempo transitions are all equiprobable.

Finally, in the presented model, it is assumed that the rhythmic pattern is fixed for a single track, such that

$$P(r_k|r_{k-1}) = \mathbb{1}_r, \tag{9.6}$$

with the indicator function being one when $r_k = r_{k-1}$ and zero otherwise. Different model constructions might allow for change at bar boundaries. This modifies the factorization and produces a transition function like [263]

$$P(r_k|r_{k-1}, \phi_k, \phi_{k-1}) = \begin{cases} A_{r_{k-1}, r_k}, & \phi_k < \phi_{k-1}, \\ \mathbb{1}_r, & \text{otherwise} \end{cases}, \tag{9.7}$$

where $\mathbf{A}$ is a homogeneous transition matrix for rhythmic patterns. In this latter case, the transition probabilities are usually learned from data [263, 264, 266].

Observation Model

From the derivation of Figure 9.2, we can express the observation model as

$$P(\mathbf{y}_k|\mathbf{x}_k) = P(\mathbf{y}_k|\phi_k, r_k), \tag{9.8}$$

which means the probability of observing a feature at frame k depends only on the bar position and rhythmic pattern variables [196]. In most works, the observation features $\mathbf{y}_k$ are obtained after the computation of a spectral-flux-like function in two channels, typically above and below 250 Hz. The low frequency channel, in particular, provides good representation of rhythmic patterns in the audio [76]. One common approach is to assign a rhythmic pattern to each bar in the training data, using either manual annotations or clustering techniques (e.g., GMM, k-means). For instance, bars can be discretized into 64th note cells, for example, and a GMM can be fitted for each point in this temporal grid and for each rhythmic pattern. This manifests the tempo independence of observations probabilities, and allows for a slow change only at every 64th note step (e.g., 25 bar positions when $M_{\max} = 1600$). Other works have exploited activations produced by neural networks as observation sequences [265, 269, 271].

Inference

The objective of the bar-pointer model is to identify the most probable state trajectory $\mathbf{x}^*_{1:K}$ given the observations, i.e., the maximum a posteriori (MAP) sequence that maximizes the posterior probability,

$$\mathbf{x}^*_{1:K} = \underset{\mathbf{x}_1,\ldots,\mathbf{x}_K}{\arg\max}\, P(\mathbf{x}_{1:K}|\mathbf{y}_{1:K}). \tag{9.9}$$

For a discrete space state, exact inference is possible with the Viterbi algorithm. Results can be approximated with particle filters [196, 263, 266–268].

9.1.2 TCN-Based Multi-Task Approach

Early deep-learning models mainly used RNNs (BLSTMs) at the likelihood estimation step. RNNs are widely known to be expensive to train and have several limitations such as vanishing gradients, low interpretability, and non-parallelizability, despite being naturally suitable for modelling sequential data [274].

The current state of the art has been achieved by replacing the BLSTM network for a TCN, first for beat tracking [274], then tempo and beat [273], and finally tempo, beat, and downbeat [275]. A TCN is a special type of CNN where temporal structures in the data are learned using dilated convolutions, which are convolutions across subsampled input representations [275]. The TCN-based model presented by DAVIES and BÖCK [274] was inspired by the WaveNet generative model [281]. When compared to the traditional BLSTM formulations, the TCN was shown to perform at the same level, while requiring less trainable parameters and being much more efficient to train due to the parallelization property of CNNs. In particular, [273] and [275] showed that multi-task formulations improve the quality of the beat tracking. This section reviews the components of the multi-task architecture in [275], which can be seen in Figure 9.3.

Input Representation

The input representation of the TCN model is a single log magnitude spectrogram obtained from a mono audio input signal (sampled at 44.1 kHz) with a hop size of 10 ms (i.e., a frame rate of 100 Hz) and a window size of 46.4 ms (2048 samples) [274]. The frequency axis is represented in the log-scale by grouping frequency bins with a set of 12 overlapping triangular filters with 12 bands per octave. This yields an input representation with 81 logarithmically spaced frequency bins from 30 Hz to 17 kHz [274].

Convolutional and Pooling Layers

Prior to being passed through the TCN, the input spectrogram is first processed by three convolutional layers. Each layer has 20 filters of sizes 3×3, 1×12, and 3×3, respectively, and is followed by a 1×3 max-pooling layer that samples the largest value along the frequency direction. This means that overlapping spectrogram slices of 5 frames in length are reduced from 5×81 down to 3×26, 3×5, and finally to a single dimension with 20 features. Each convolutional layer has an exponential linear unit (ELUs) as the activation function and a dropout rate of 0.1 is applied after pooling.

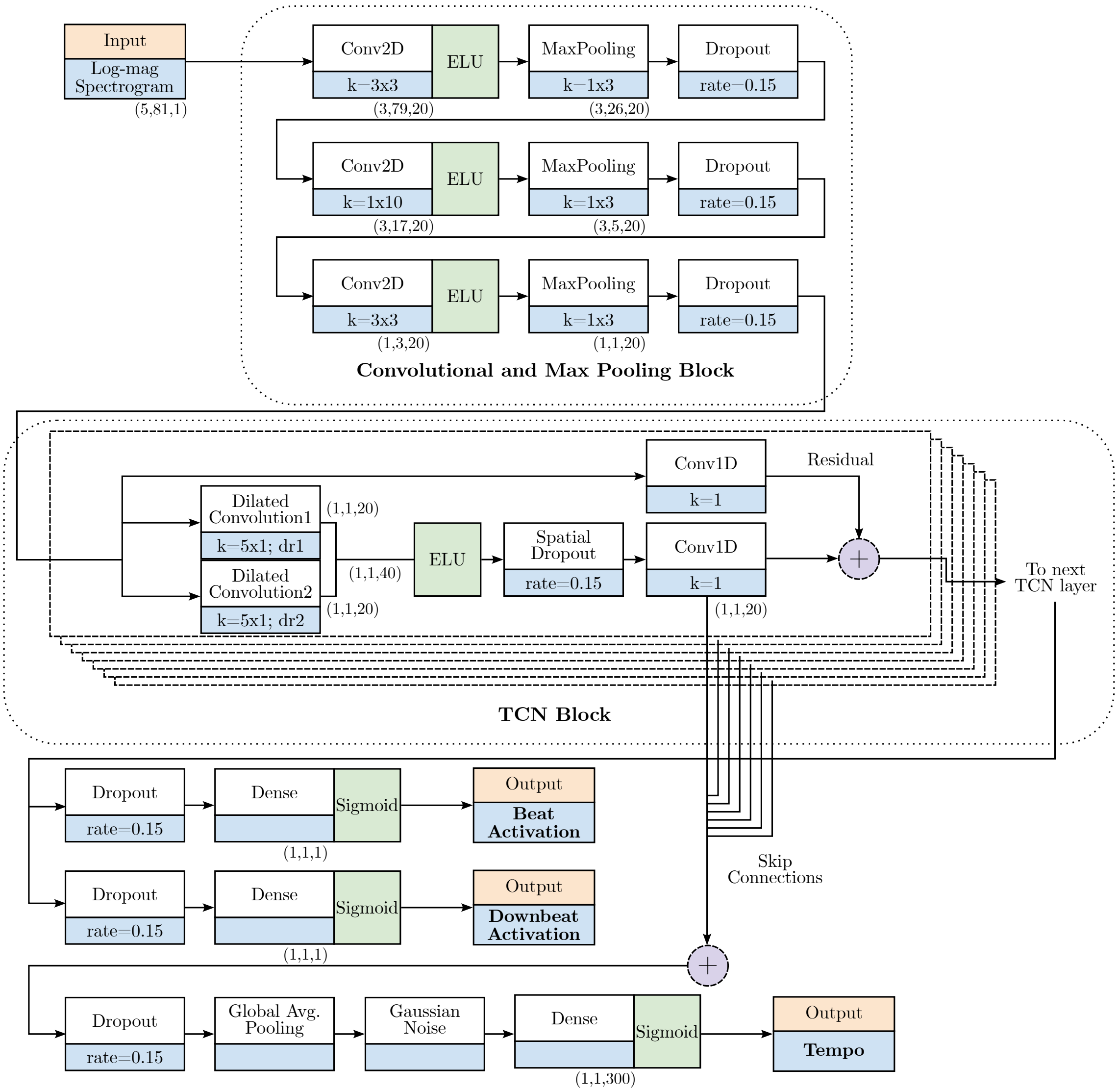


Figure 9.3: Architecture of the TCN model. Adapted from [17].

TCN Layers

The 20-dimensional feature output by the convolutional block is then fed to the TCN block, which captures temporal structure using filters with dilated convolutions. While regular convolutions can only handle contexts in the sequential data whose sizes vary linearly with the depth of the network [282], dilated convolutions enable exponentially large receptive fields [241] and much wider contexts. Moreover, the TCN beat tracking model uses non-causal filters, which allow capturing information both forwards and backwards in time. Figure 9.4 illustrates this processing. Authors have used 11 layers with one-dimensional filters of size 5 and geometrically increasing dilations from 2^0 to 2^{10} frames. A second feature map obtained from a second dilated convolution (at a doubled dilation rate) is concatenated with the first feature map, before spatial dropout (with a rate of 0.15) and ELU activation. The feature maps are combined by a 1×1 convolution with 20 filters, added to the identity path of the layer,[2] before being forwarded to the next layer.

Tempo, Beat, and Downbeat Estimations

The TCN uses skip connections that are branched from the parameterized paths of residual blocks to predict tempo in a linear range from 0 to 300 bpm. These skip connections are aggregated by summation and averaged over time, leading to a single 20-dimensional feature vector for classification in a dense layer (with softmax activation). Quadratic interpolation over targets is used to find the exact tempo of the piece.

The main TCN output is used in two different binary classification problems, where the network is asked to predict if any given frame is a pulse (beat or downbeat) or not. To accomplish this, the authors implemented two branches, one for beats and the other for downbeats, each with its own dropout layer, dense layer, and sigmoid activation. These branches are trained to indicate the likelihood of a frame containing the desired pulse. Both likelihood sequences then pass through a post-processing stage with a DBN, where inference can be done jointly or separately. In the latter case, the beats are first detected and then, based on the beat predictions, downbeats are estimated in a second inference step.

Tempo, beat, and downbeats targets are widened such that direct neighbouring frames and the ±2 bpm values are also positive, but with lower weights than the annotation.

[2] A residual block contains two branches that are added to form the output of the block: a “shortcut” or “identity” connection directly taken from the input; and the transformed version of the input (a parameterized “residual” mapping). This is done for the ease of training, as layers are then trained to learn modifications to the identity signal, rather than full transformations [282].

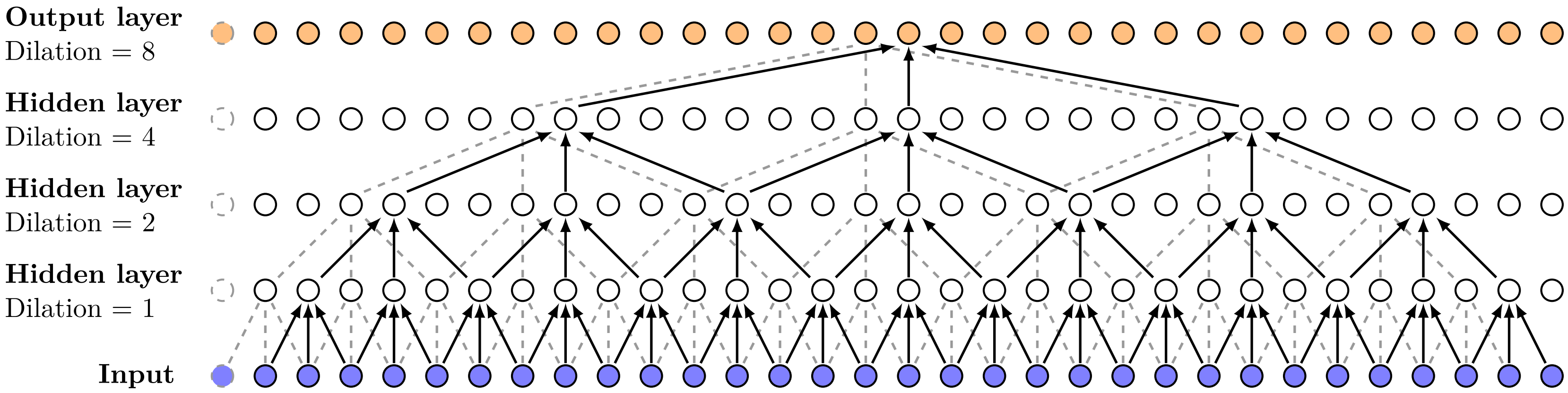


Figure 9.4: Overview of the TCN structure. Example of a non-causal network with a depth of four layers, and geometrically increasing dilation factors ($d = 1, 2, 4, 8$). Gray dashed lines show the network connections shifted back one time step. Adapted from [274].

Data Augmentation

We alluded before to the fact that CNN-based systems can suffer from tempo bias, i.e., when the model is unable to generalize well to unseen tempi. With this in mind, BÖCK and DAVIES [275] introduced a simple data augmentation strategy to extrapolate information from well-covered regions of the training data to sparser ones. Instead of applying transformations (e.g., time stretching, pitch shifting, sample rate conversion) to the audio signal, they change the hop parameter of the STFT sampling from a normal distribution with 5% standard deviation from the annotated tempo (updating the targets accordingly). These representations produced by different overlaps are effectively interpreted by the network as differing in tempo.

9.1.3 Adaptive Beat Tracking

Much of the previously addressed research on beat tracking with data-driven strategies has focused on developing "universal" models trained on large amounts of annotated data. Due to the nature of these state-of-the-art solutions, which typically depend on deep learning methods, high accuracy scores can usually be achieved given a sufficiently large pool of quality data annotations [242, 275, 283]. However, this good performance cannot be guaranteed when models are used to estimate beats from challenging or unseen music, e.g., music with highly expressive timing [284] or from culturally specific traditions that were not present during training [21, 22].

In recent years, there has been an increasing amount of literature on this real-world problem, i.e., when an end user wants to apply state-of-the-art models to a limited subset of examples with unseen rhythmic characteristics. For example, FIOCCHI et al. [283] proposed an inductive transfer learning approach: a beat tracking system, built with a three-layer LSTM network, was first trained on popular music (including some songs from the Ballroom dataset) and later adapted to work on a smaller target dataset of Greek folk music. The adaptation was achieved by freezing the lower two layers of the network and running more training cycles on the previously unseen data. As in other similar works, a DBN was used for the inference of beat positions. The main model was compared against a baseline RNN-based network trained only on the popular music dataset and another BLSTM-based network trained only on Greek music. In all cases, the same 40% of the Greek music dataset was left out for testing. Authors reported several metrics, with modest results — a gain of 6 percent points on average (e.g., F-measure of 57.2% in the baseline network, 58.4% in the "specialist" network, and 64.0% with the transfer learning approach). In this work, recurrent networks provide an interesting approach for genre adaptation but are computationally expensive, making them of concern for real-world applications.

Other systems have leveraged the subjective nature of beat induction to allow a user to guide and improve the tracking process with a limited amount of annotation. Unlike FIOCCHI et al. [283], PINTO et al. [17] restricted the network adaptation problem to a single music piece. They used as baseline the multi-task TCN-based model presented in Section 9.1.2, which was trained on six large datasets (over 26 h of music material): Ballroom, Beatles [285], Hainsworth [273, 286], HJDB [271, 287], Simac [288], and SMC [289]. The fine-tuned model used the same architecture as the baseline, and in both cases a DBN was used to process the outputs. The network was fed with a 10-second excerpt of an unseen music piece — the first 5 s were used for training and the remaining 5 s for validation — and required to predict beats for the remaining of the piece. The adaptation was performed allowing all layers to be updated at a smaller learning rate than (one fifth of) the one used in the baseline. They reported promising results across common datasets. More interestingly, they exemplified the improvements obtained when using the proposed approach with two challenging pieces: an *a cappella* performance and a guitar piece with a lot of *rubato*. This system could be used in the real-world applications to enhance the annotation workflow, providing an end-user some flexibility in the selection of which small portion (10-second region) of the audio to annotate, and ultimately leading to more accurate and efficient annotations.

Finally, YAMAMOTO [290] presented a different solution in the form of an interactive beat-tracking interface. His system used an architecture very similar to that of the TCN model, but with delta MFCCs as input representation, which was trained on a local context (∼5 s) with the following datasets: GTZAN [205, 235, 291], RWC popular and genre [292, 293], and an in-house dataset of 400 musical pieces of various genres. Beat positions were inferred with a hidden semi-Markov model (HSMM) and the Viterbi algorithm. Once the user has loaded a music signal, the system estimated beats for the entire piece with the pre-trained network. A derived network architecture including an adaptive self-attention mechanism was used at runtime. Every time the user has made a correction on a single beat, the system adapted to this local modification, and reflected changes over the global context. His experiments showed that, by adapting to both user and piece, the framework dramatically reduces the effort for manual corrections during the annotation process.

9.2 Approaches to Microtiming Analysis

In this section, we discuss various works related to micro-rhythmic characterization of music, both from musicological and computational perspectives. We specifically focus on the computational analyses performed on Brazilian *samba*, but investigations on jazz are also presented. FUENTES [294] has classified computational methods for microtiming description into two categories: those based on grouping of rhythmic patterns and those based on the autocorrelation function. We will also follow this classification in this section.

Musicological Aspects

As we mentioned in Chapter 2, Brazilian music and dance practices have been significantly influenced by various African cultures. This influence can be perceived in rhythm, melody, instrumentation, and overall organization of the musical phenomenon. For instance, the rhythmic pattern of the "characteristic" syncope is believed to have been either brought to or developed in the continent by enslaved Africans from the diaspora, possibly after the contact with the triplets commonly found in Iberian music [50]. The "texture" of *samba* is itself generated by the superposition of different rhythmic and timbral cycles, much akin to percussive ensembles found on the other side of the Atlantic [24]. In many types of African music, and in West African dance music traditions in particular, bells are played in an *ostinato* that provides a structural matrix around which the entire performance is organized [295]; different high-pitched instruments (or even hands clapping) have the same responsibility in *samba de roda* [296] and *partido-alto*.

Moreover, musicologists report how there exists, especially in West African and Afro-Atlantic music, a systematic approach to rhythmic expression that is delivered at a very fine scale [199]. Practically speaking, this corresponds to performing small timing deviations, i.e., articulating notes a little earlier/later than what would be expected in an equidistant division of the cycles and beats (isochrony), in order to elicit the sensation of "groove". Examples of these non-isochronous rhythms include Cuban *rumba* [171], Malian *djembe* music [14], and *candombe* [38, 203]. This phenomenon is also featured in different Afro-Brazilian genres [296, 297]. GERISHER [297] investigated these micro-rhythmic properties in *samba baiano* using steady sixteenth notes, i.e., four-beat cycles where each beat contains four fast pulses. She found a consistent microtiming pattern in each beat across different instruments (*reco-reco*, *triângulo*, and *pandeiro*), expressed by the inter-onset intervals following an order given by: medium–short–medium–long. A similar pattern was later verified in *samba carioca* by GRAEFF [296], which postulated that these deviations emanate from the performer's body movements, configuring the "acoustic–motional structure" [298] of the performance. GRAEFF [296] explained that these micro-rhythmic accents must be related to the stress accents on pulses 1 and 4, which in turn correspond to wider motions of hand or arm made by the player. Specifically for pulse 4, GERISHER [297] noted that the extension of its duration is greater when it is followed by a strong accented beat (e.g., beat "2" in the $\frac{2}{4}$ meter typical of *samba*), for which it "prepares" the dancer and the listener. We could say

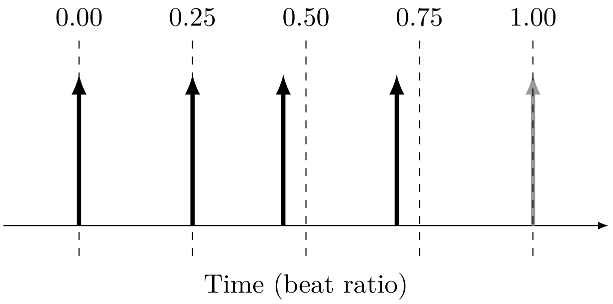


Figure 9.5: Microtiming profile found in some Afro-Brazilian genres. Beats (at positions 0.00 and 1.00) and other sixteenth notes are displayed as arrows. Dashed lines correspond to expected positions in an isochronous grid.

that the faster pulses of the *ostinato* patterns lie between an even subdivision of the beat and the triplet [201], due to this microtiming "compression". Figure 9.5 illustrates the microtiming profile of a single beat as measured by GERISHER [297] and GRAEFF [296]; notice how the medium–short–medium–long structure corresponds to playing pulses 3 and 4 ahead of their nominal positions in an isochronous subdivision of the beat interval (positions 0.5 and 0.75). More recently, HAUGEN and DANIELSEN [299] studied these non-isochronous duration patterns in *samba* with a *pandeiro*. They reported a medium long–short–medium short–long pattern irrespective of tempo.

Computational Approaches

Computational approaches for the analysis of microtiming can be generally divided into grouping- or statistical-based and those that exploit ACF representations. In methods of the first type, the audio signals are commonly preprocessed to obtain a feature vector at a slower frame rate; bar patterns (or patterns of another size) are length-normalized, making the representation independent of local tempo; and the multiple patterns in the piece are stacked together for a statistical analysis. A similar bar-length feature vector was proposed by DIXON et al. [206] (see Chapter 8).

GOUYON [300] conducted an analysis of microtiming in *samba de roda carioca*. In total, this study used 49 audio excerpts taken from commercial CDs (44.1 Hz, mono) and ranging between 10 to 30 s in length. The excerpts featured instruments like *cavaquinho*, *pandeiro*, and others. The signal was first analyzed and a complex domain ODF was computed with a frame size of 23.2 ms and 50% hop, resulting in a feature rate of 86.1 Hz. Beat positions were obtained with a semi-automatic software and manually corrected to align with the nearest maxima in the ODF. Beat-length patterns were resampled to 40 points per segment and clustered with k-means. Most of the reported patterns present local maxima at the sixteenth-note level, which corresponds to the fast rhythm usually set by the *cavaquinho* and *pandeiro* in *samba de roda.* Gouyon also confirms the compression in time observed by GRAEFF [296], GERISHER [297] of the third and fourth sixteenth notes, which are played ahead of their corresponding quantized positions by about $\frac{1}{40}$ of the IBI (e.g., 20 ms at a tempo of 90 bpm).

NAVEDA et al. [201, 301] performed a similar investigation over 106 excerpts (median duration of 33 s) from commercial recordings of *samba carioca*, *samba-enredo*, *partido-alto*, and *samba de roda baiano.* Instead of resorting to an automatic beat tracker, beats and downbeats were manually annotated by three Brazilian musicians with the Sonic Visualiser application [302]. Features were extracted with the following process. First, the audio data was preprocessed by a model of the auditory system, yielding loudness curves at a rate of 200 Hz in 44 channels distributed over 22 critical bands (centered from 70 Hz to 10 843 Hz). After aggregating the channels into three spectral regions (low, mid, high) according to the "spectral signatures" of the different instruments, authors used the annotations to segment the recordings at three metrical levels: one-beat, bar (two-beat), four-beat. Then, for each excerpt–metrical level combination, they interpolated an isochronous grid at the sixteenth-note level (using the annotated IBI); obtained a refined position of the first beat of the pattern with the average peak position of the loudness across all spectral regions; retrieved the position of the highest peak around each sixteenth note in each spectral region. The refined position of the first sixteenth note is the reference relative to which all IBIs and microtiming were computed. Then, for each sixteenth note, they recorded the position of its peak (relative to the IBI) and its intensity, all stored in a $(3 \times 2 \times n)$-dimensional vector, where n is the number of sixteenth notes at the metrical level. Confirming previous studies, they observed significant anticipations of the third and fourth sixteenth notes with respect to the quantized positions at the one-beat level. Separated by spectral region (low, mid, high), these anticipations were valued at -0.026, -0.031, -0.032 beats and -0.028, -0.018, -0.027 beats for the third and fourth sixteenth notes, respectively. This results are very consistent with the findings by GOUYON [300]. They also clustered the data to analyze the relations between microtiming, metrical level, intensity and spectral region. This final analysis revealed a small delay of the instruments at the lower part of the spectrum on the first sixteenth note of each beat, which is more expressive on the second beat of a bar. Moreover, they also detected a rhythmic device akin to *accelerando* and *ritardando* at the microtiming level.

Other works that use similar techniques to analyze drumming recordings include [171] (*rumba*), [38, 203] (*candombe*), and [14] (*djembe*).

We end this section highlighting two approaches that describe micro-rhythmic elements with the help of an autocorrelation function. The object of study in these works is the "swing" at the eighth-note level that is employed in jazz music, but also in other genres as blues or rock [235]. This deliberate micro-rhythmic variation occurs when the performer modifies the beat subdivision between two consecutive eighth notes. As a result, the first note is lengthened while the second is shortened. The swing ratio is precisely defined as the ratio between the durations of the long and short notes [235]. This is a continuous scale, usually ranging from “straight eighths” (1:1), “triplet feel” (2:1), to “dotted eighths” (3:1), and occasionally more extreme ratios [198].

The system by MARCHAND and PEETERS [235] started the estimation of the swing ratio the same way as the other approaches, i.e., with the computation of an ODF. They used a frame size of 16 s in length, with a hop of 1 s for the analysis. They noticed that, when there was no swing, the ACF of the ODF showed one peak representing the tactus (quarter note) and one peak representing the duration of the eighth note. However, when swing was present, the eighth note peak was split in two for the durations of the “short” and “long” eighth notes. To determine the corresponding durations of the eighth notes, the system first found the tactus with a probabilistic framework, from which it estimated the duration of the eighth note, δ_e; then it performed two Gaussian fits, one in the interval $[\frac{\delta_\mathrm{e}}{2}, \delta_\mathrm{e}]$ and the other in $[\delta_\mathrm{e}, \frac{3\delta_\mathrm{e}}{2}]$. The mean of each Gaussian (μ_s and μ_l) and their standard deviations were used in a series of heuristics to determine if there was swing or not in each frame. Where swing was present, the swing ratio could be computed as

$$s_\mathrm{r} = \frac{\mu_\mathrm{l}}{\mu_\mathrm{s}}. \tag{9.10}$$

Authors reported a mean recall of 74% for the recognition of swing/no swing in a subset of the GTZAN dataset annotated at the eighth-note level. When annotated tempo was also used (and the beat tracking stage was bypassed), mean recall improved to 91%.

DITTMAR et al. [197, 198] explored the same idea, but instead of the regular ACF, they used the log-lag ACF (LLACF), which, as mentioned in 8.2, is a tempo-insensitive representation where tempo changes become translations in a logarithmically-warped lag axis. Then, they constructed a dictionary containing the equivalent representations of an idealized ride cymbal pattern with different swing ratios in the range $1 \leq s_\mathrm{r} \leq 4$. Swing ratios were estimated from audio recordings by matching against the dictionary prototypes with either the Pearson correlation coefficient [198] or the inner product of the DFT magnitudes [197]. The latter is particularly efficient, and works by ignoring the translation of the LLACF representation when the phase information is discarded. In [197], authors also presented the swingogram, which can be interpreted as a spectrogram with a swing-ratio axis instead of frequency. Each bin (m, s_r) in the swingogram is equivalent to the similarity score between the m-th audio frame and the prototype corresponding to s_{r}. This representation allowed the tracking of the swing ratio along the performance.

Chapter 10

Investigation on Beat and Downbeat Tracking

In this chapter, our focus is on the challenge of producing high-quality beat and downbeat annotations for our datasets. We described in Chapter 3 how BRID was manually annotated. Manual annotations are labor-intensive and costly, so it would be beneficial to evaluate the suitability of different automatic algorithms in our specific musical context. However, it is important to note that most of the beat and downbeat trackers presented in the literature have either been tailored for or developed with Western music in mind. Therefore, we approach this problem of annotation cautiously, as these out-of-the-box tools might be ineffective if directly applied to music from different cultural traditions.

Before presenting our investigation, we introduce a few evaluation metrics that allow the interpretation of large-scale experiments. To simplify, we explain all metrics as they are used in beat tracking, but their extension to downbeat tracking is evident. Some of the information described here was originally published in [19, 20], and we maintain nearly the same structure and results of these documents, with some additional text and experiments that connect the publications in this work. Sections 10.4 and 10.5 go into detail on the annotation of beats and downbeats in SAMBASET, respectively.

10.1 Evaluation Metrics

Several evaluation metrics are used in beat tracking without consensus among researchers [285, 289]. Among such metrics, in objective methods, a list of I tracked beat times $\{b_i\}$, $i \in \{1, ..., I\}$, is compared against one or more ground truth annotated beat times, e.g., a sequence $\{a_j\}$ of J elements, where $j \in \{1, ..., J\}$. This kind of evaluation is analogous to that of onset detection (see Section 7.2).

As the hand-labelled ground-truth data (beat and downbeat times) are bound to carry uncertainties, which typically revolve around 50 ms [285], and it is virtually impossible that estimates exactly match them, evaluation metrics usually set error tolerance intervals in either absolute (e.g., ±70 ms) or relative (e.g., 20% of the inter-annotation interval, IAI) time [285]. We present in the following a few usual metrics, explaining their functionality and limitations. The reader is referred to [285] for other metrics and a more thorough discussion about the subjective nature of beat (and downbeat) annotations, and the relation between the actual perception and the delay in human response. We note that in our experiments all figures of merit are computed with standard settings of the `mir_eval` Python package [303] (v0.7).

F-Measure

First, there is the F-measure [250], which is obtained via the harmonic mean between precision and recall, as defined for onset detection in Equation 7.6. An estimated beat position b_i is considered correctly detected if it lies inside a window centered at annotation a_j. As mentioned before, the tolerance is usually set to ±70 ms [285].

F-measure values go from 0% to 100%, where the former can only occur when no beat times fall within any of the tolerance windows. If beats in the estimated and annotated sequences are well aligned, but express different metrical levels related by a factor of two, we have what is known as an "octave error" [5]. In this case, the F-measure drops from 100% to 66.7% [289]. Completely unrelated sequences typically perform around 25% [289].

Continuity-Based Measures

Continuity-based evaluation [286] considers the ability of the beat tracking system to correctly and continuously track the meter in different regions of the music, which contrasts with the simpler evaluation scheme of the F-measure. Continuity is assessed by three conditions where, within a tempo-dependent tolerance factor ($\xi = 17.5\%$) of the current inter-annotation interval $\Delta a_j = a_{j+1} - a_j$, a beat detection b_i is said to have been correctly estimated if [285]:

1. it falls within a tolerance window around the closest annotation, i.e.,

$$|b_i - a_j| < \xi \Delta a_j; \tag{10.1}$$

2. the same holds for the previous estimated beat, i.e.,

$$|b_{i-1} - a_{j-1}| < \xi \Delta a_{j-1}; \tag{10.2}$$

3. the estimated inter-beat interval is locally consistent with inter-annotation interval, i.e.,
$$|\Delta b_i - \Delta a_j| < \xi \Delta a_j. \tag{10.3}$$

The proportion of detected events that satisfy these conditions and the total number of annotations defines the total number of correct beats at the correct metrical level (CMLt). If, by resampling $\{a_j\}$, we allow for detections at double or half the correct metrical level (octave errors), we obtain the AMLt (allowed metrical level).

A score of 0% can only occur for these metrics if no two consecutive beats fall within any tolerance windows, e.g., when the sequences are related by unspecified metrical relations like 3:2 [289]. For completely unrelated sequences, AMLt reports scores around 18%, on the account of coincidences [289].

Information Gain

Finally, the information gain [304] is defined as the Kullback–Leibler divergence between two distributions (approximated by histograms). The first one is the observed beat error distribution, which considers the normalized timing errors of all estimated beats within a beat-length window around the annotations. The second is a uniform distribution that models the error of a pair of unrelated (estimated and annotated) beat sequences. Therefore, the information gain measures the distance between the empirical beat error distribution and the theoretically worst beat tracker.

If we define the set of beats within a one-beat window around a_j as $\{b_q | a_j - \Delta a_{j-1}/2 \leq b_q \leq a_j + \Delta a_j/2\}$; and the normalized timing error as

$$\zeta_{b|a}(q) = \begin{cases} \dfrac{b_q - a_j}{\Delta a_{j-1}/2}, & b_q \leq a_j \\ \dfrac{b_q - a_j}{\Delta a_j/2}, & b_q > a_j \end{cases}; \tag{10.4}$$

then we can construct, from the sequence of errors ζ, a K-bin histogram $p_\zeta(z_k)$, which represents the estimated probability of bin $k \in \{1, \ldots, K\}$ centered at the beat error value $z_k \in [-0.5, 0.5]$ beat, such that $\sum_{k=1}^{K} p_\zeta(z_k) = 1$.

Finally, we can determine the information gain as [285]

$$\begin{aligned} D_\zeta &= \sum_{k=1}^{K} p_\zeta(z_k) \log_2 \left(\frac{p_\zeta(z_k)}{\frac{1}{K}} \right) \\ &= \sum_{k=1}^{K} p_\zeta(z_k) \log_2 p_\zeta(z_k) + \log_2 K \\ &= \log_2 K - H(p_\zeta(z_k)), \end{aligned} \tag{10.5}$$

where

$$H(p_\zeta(z_k)) = -\sum_{k=1}^{K} p_\zeta(z_k) \log_2 p_\zeta(z_k) \tag{10.6}$$

is the entropy of the estimated beat error distribution. Since the entropy is lower and upper bounded by 0 (single-valued probability distribution) and $\log_2 K$ (uniform distribution), respectively, the information gain also spans this range $[0, \log_2(K)]$ bit. In practice, the largest entropy of the forward and backward sequences $\zeta_{b|a}$ and $\zeta_{a|b}$ is selected. An empirically determined good choice for the number of bins is $K = 40$ [304] or 41 [303]. This parameter controls the quantization of the beat error: if it is too small, the shape of the distribution is not accurately captured, whereas if it is too large, the histogram becomes too sparse [304].

This metric is insensitive to consistent beat-relative offsets (phase errors) and not very sensitive to octave errors,[1] while at the same time providing a true zero value for unrelated sequences.

10.2 Typical Beat and Downbeat Tracking Errors

In this preliminary test, we estimate beat and downbeat positions on a few samples from the BRID datasets (solos and mixtures) using three different deep-learning-based systems that are available as out-of-the-box tools. We then perform an analysis of the results, discussing the limitations of these models with respect to their application on non-Western datasets through an analysis of typical errors.

We select a subset of eight audio files that are representative of the content of the dataset. This subset comprises four solo and four mixture tracks, involving different rhythms (*samba*, *samba-enredo* and *partido-alto*), tempi, and ensembles, as summarized in Table 10.1.

Table 10.1: Recordings in the selected subset.

	Filename	Instruments	Genre
Solos	`[0218] S2-TB3-01-SE`	Tamborim	Samba-enredo
	`[0229] S2-CX2-02-PA`	Caixa	Partido-alto
	`[0258] S2-SK2-02-PA`	Chocalho	Partido-alto
	`[0280] S2-SU2-05-SE`	Surdo	Samba-enredo
Mixtures	`[0013] M4-13-SE`	Cuíca, Caixa, Tamborim, Surdo	Samba-enredo
	`[0039] M3-20-SE`	Caixa, Tamborim, Tantã	Samba-enredo
	`[0047] M3-28-SE`	Caixa, Surdo, Surdo	Samba-enredo
	`[0051] M2-03-SA`	Tantã, Surdo	Samba

[1] Tapping at different metrical levels creates multiple modes (peaks) in the error distribution, which will still be far from uniform, and therefore yield high information gain. Provided that the tapping tempo is close to the actual tempo, octave errors do not seriously harm the metric [304].

Finally, we adopt three deep-learning-based tracking systems that were briefly discussed in the previous chapter: the multi-model beat tracking system of BÖCK et al. [240] (BO1), the joint beat and downbeat model of BÖCK et al. [271] (BO2), and the downbeat model of KREBS et al. [270] (KRE). We use the implementations available in the `madmom` package (v0.16) and score detections with the F-measure against our manual annotations.

Analysis of the Selected Solos

As we have stated many times, the instruments in *samba*, such as *tamborim* or *chocalho*, are usually played in an *ostinato*, i.e., a repeating rhythmic pattern (see Figure 10.1). This phenomenon was captured in the BRID dataset. Due to this cyclic performance, it can be very difficult to establish the location of beats and even more so of downbeats in solo tracks without any further references — this was also a

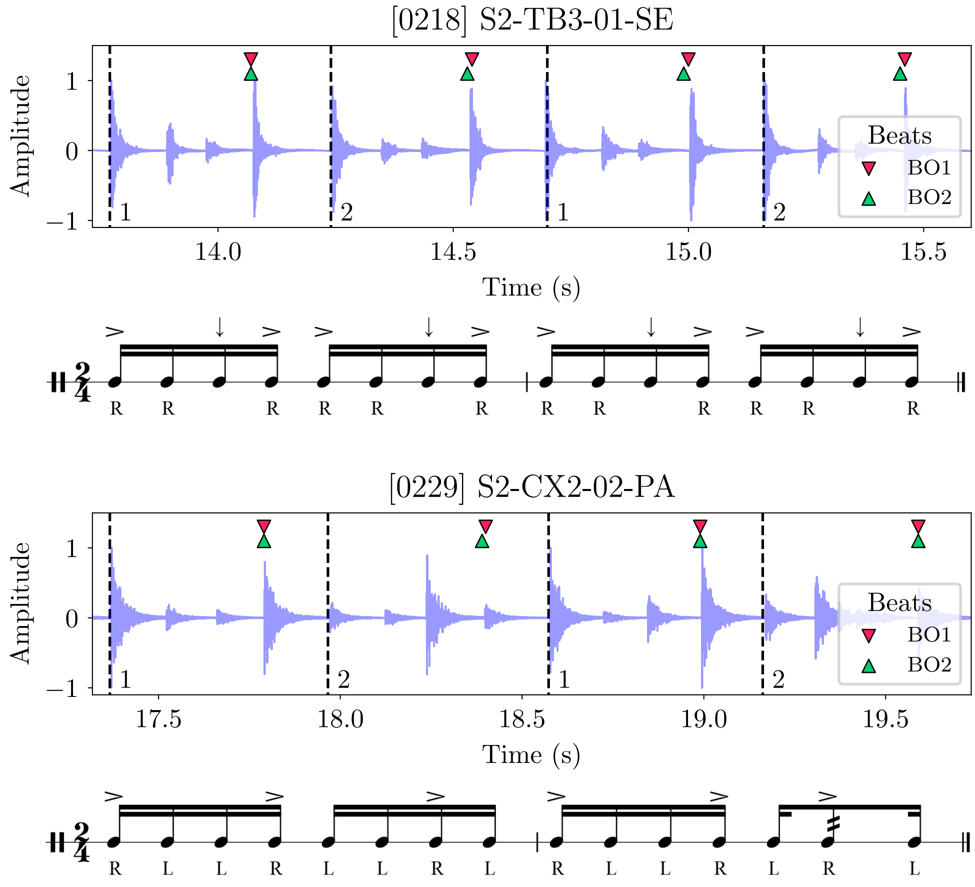


Figure 10.1: Beat tracking for the selected solo track examples. Instruments: *tamborim* and *caixa*. The waveforms show two bars of the rhythmic patterns, with dashed lines indicating annotated beats. Other markers depict the position of beat estimates with BO1 and BO2. Rhythmic patterns are schematically represented in music notation (below) and roughly aligned. (Continued on the following page.)

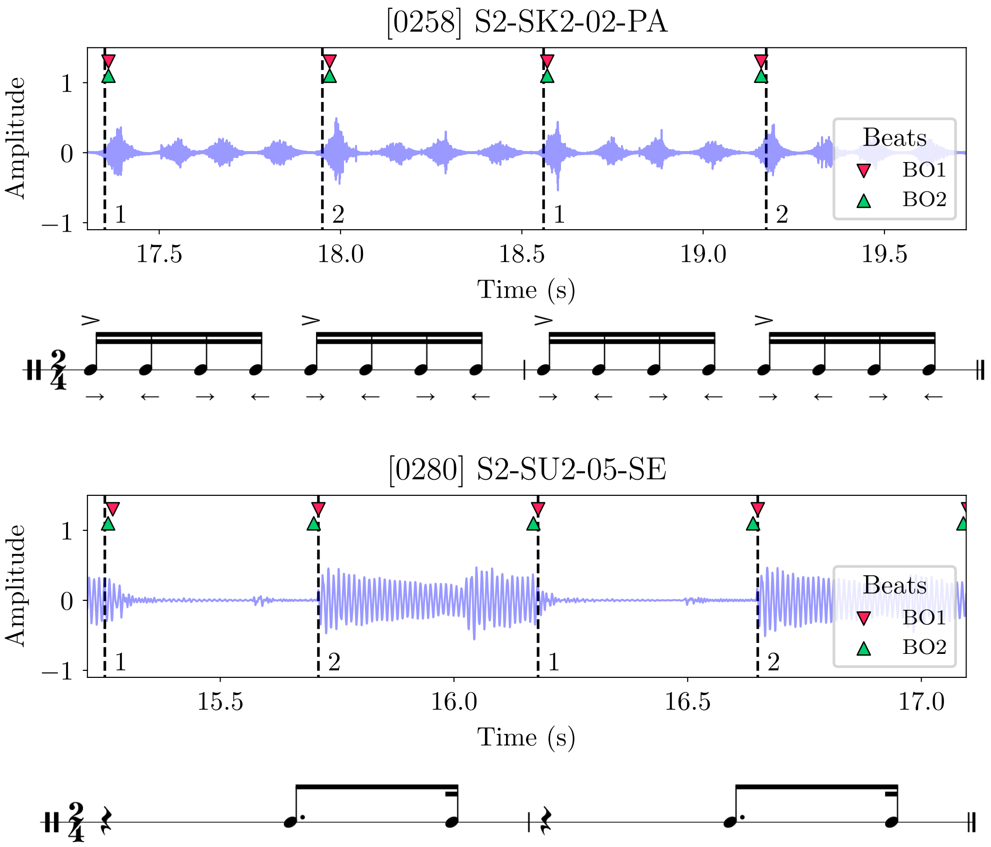


Figure 10.1: (Continued from previous page.) Instruments: *chocalho* and *surdo*.

challenge during the manual annotation process. Of course, the dynamic evolution of the pattern (succession of strong and weak pulses) as well as the microtiming profile give important clues for the annotator. However, this specific musicological knowledge is not available to most tracking models. For this reason, we restrict ourselves to only investigating beat tracking in solo tracks in this section.

The beat positions for each of the four solo track are estimated using the BO1 and BO2 algorithms and results are presented in Table 10.2. A two-bar length excerpt of each audio file is shown in Figure 10.1, which also depicts the annotated beat positions, the beat estimates for each algorithm, and a roughly-aligned notation of the rhythmic pattern. Although we are not interested in downbeats at this time, we indicate in this figure the beat number of the annotations considering a $\frac{2}{4}$ meter (i.e., “1” and “2”), for reference.

We see from these results that the algorithms perform very similarly: both miss the phase of the beat in two of the files (`[0218]` and `[0229]`) and correctly track the other two tracks (`[0258]` and `[0280]`). A detailed inspection of Figure 10.1 makes it clear that the troublesome rhythmic patterns, i.e., from *tamborim* and *caixa*, have strong accents displaced with respect to the metric structure. Conversely, the *chocalho* pattern accentuates every beat more than other pulses. Finally, in the

Table 10.2: Beat F-measure for different models on selected solos.

	Beat F (%)	
Track	BO1	BO2
`[0218] S2-TB3-01-SE`	0.0	0.0
`[0229] S2-CX2-02-PA`	0.0	0.0
`[0258] S2-SK2-02-PA`	100.0	100.0
`[0280] S2-SU2-05-SE`	96.5	100.0

case of the *surdo* track, which records a *surdo de terceira*, there are actually several different rhythmic patterns played. Nevertheless, most of the time, the second beat of the bar is clearly articulated. We mentioned before that this accentuation of beat “2” is a distinctive trait of *samba.* While advantageous for beat tracking, it proves to be very challenging for downbeat estimation, as will be shown next.

Analysis of the Selected Mixtures

We track both beats and downbeats in the four recordings of acoustic mixtures, as the ambiguity in the rhythmic cycles can be better resolved in these cases due to extra information provided by the superposition of different instruments. Beats are estimated using BO1 and BO2, while downbeats are inferred with KRE and BO2. Since all mixtures are in $\mathbf{\frac{2}{4}}$ meter, we set the search-space of the downbeat-tracking DBN to bar lengths of $\{2, 4\}$ beats. Results are displayed in Table 10.3.

Table 10.3: Beat and downbeat F-measures for different models on selected mixtures.

	F-measure (%)			
	Beat		Downbeat	
Track	BO1	BO2	KRE	BO2
`[0013] M4-13-SE`	99.1	100.0	0.0	0.0
`[0039] M3-20-SE`	98.1	100.0	0.0	0.0
`[0047] M3-28-SE`	97.9	100.0	0.0	0.0
`[0051] M2-03-SA`	56.2	39.5	0.0	0.0

Beat tracking in the mixtures is apparently met with ease, except for file `[0051]`, for which half of the estimates are out of phase probably due to the presence of an anacrusis.[2] On the other hand, downbeat tracking algorithms are completely unsuccessful; both fail to correctly track the downbeats for all the recordings. We observe that the downbeat estimates tend to follow not the first, but the second beat,

[2] In music, an anacrusis is the presence of one or more notes that precede the first metrically strong beat of a phrase [3].

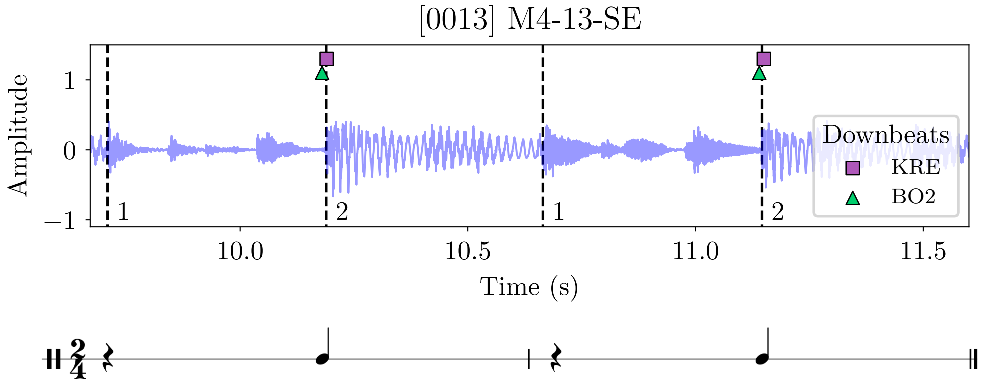


Figure 10.2: Downbeat tracking for one of the selected mixture track examples. The waveform shows two bars, with dashed lines indicating the annotated beats. Markers depict the position of downbeat estimates with KRE and BO2. We have notated (below) the *surdo* part, which we believe is mostly responsible for misleading the downbeat detection process.

what suggests that *samba*'s characteristic accent is misleading the trackers. This can be mostly attributed to the patterns executed by *surdo* and *tantã*, which most clearly articulate this beat. Figure 10.2 illustrates this problem when estimating downbeats for track `[0013]`.

10.3 Beat Tracking on BRID Mixtures

In the previous investigation, we have seen the limitations of deep-learning-based trackers, which represent the state-of-the-art in beat estimation for many datasets, in producing reliable estimations for solo (beat) and mixture (beat/downbeat) tracks of BRID. Given the apparent success of these models in tracking beats on the acoustic mixtures, we now perform a more in-depth investigation of available beat trackers in this specific task. We expand the number of algorithms with two other systems: BeatRoot, by DIXON [250] (DIX), which is based on a multiple agent architecture, and available in Java (v0.5.8); and the dynamic programming system of [245] (ELL), available in the `librosa` [305] Python package (v0.6.2). Since we are only dealing with beat tracking, we do not consider KRE in this case, and investigate only the BO1 and BO2 algorithms. The beat estimation results for BRID mixtures is summarized in Table 10.4, including the continuity metrics and information gain.

Contrary to our expectations, this time the average beat tracking performance is much lower, with a mean F-measure of 60.5% for BO1 and 53.5% for BO2 (cf. Table 10.3). Moreover, we observe that both deep-learning-based systems have been outperformed by DIX (considering F-measure and AMLt) and ELL (under all metrics). In fact, ELL presents itself as the most accurate in this scenario.

Table 10.4: Performances of different beat tracking models on BRID mixtures.

Model	CMLt (%)	AMLt (%)	*F*-meas. (%)	Inf. gain (bits)
DIX	42.8	83.6	72.8	3.48
ELL	**82.3**	**84.6**	**85.4**	**3.66**
BO1	45.1	66.8	60.5	3.51
BO2	37.0	62.5	53.5	3.32

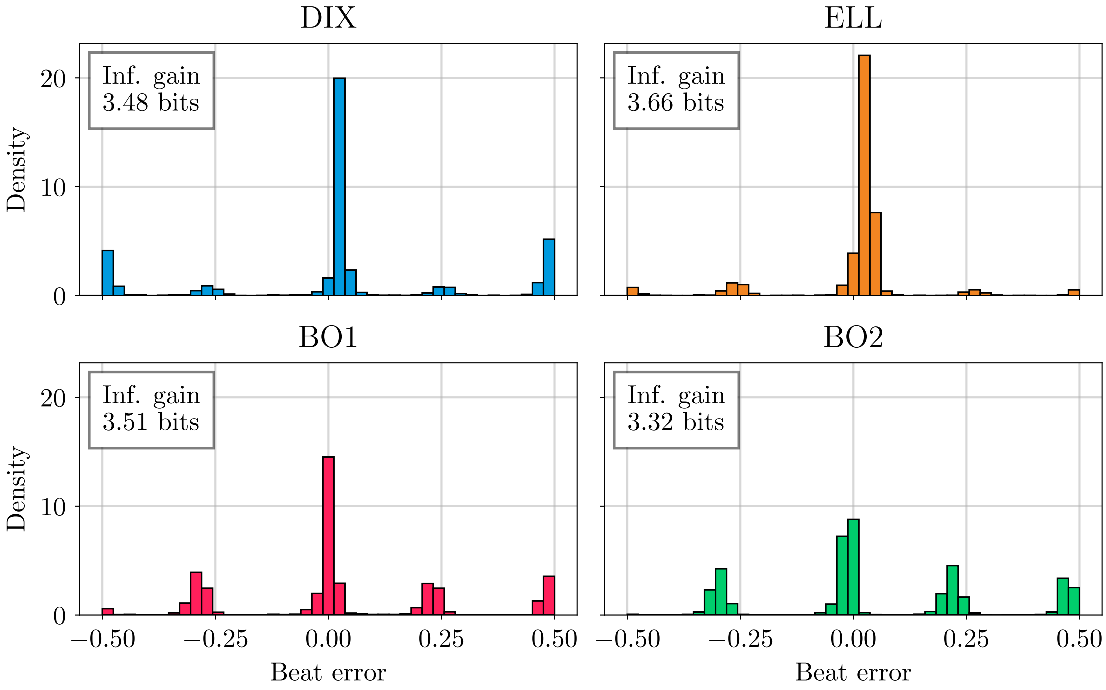


Figure 10.3: Histograms for beat detection errors across all mixture tracks.

The large differences between the CMLt and AMLt for DIX, BO1 and BO2 point towards the occurrence of octave errors in the metrical level of several tracks. This can be better understood with the aid of Figure 10.3, which presents the estimated error histograms for each model. In all histograms, we notice a high peak centered close to 0 and smaller masses in the neighborhoods of −0.5, −0.25, 0.25, 0.5 (i.e., $\pm\frac{1}{2}$ and $\pm\frac{1}{4}$ beat), which display higher densities for DIX, BO1, and BO2. While the relative errors of ±0.25 can be attributed to simple phase errors, errors of half a beat can also appear as a result of octave errors (i.e., methods estimated double/half tempo). These errors are ignored by the AMLt, but absorbed by the CMLt metric. The global beat tracking information gain quantifies these differences: the beat error distribution for BO2 is more spread out (less related to the actual annotations), whereas the distribution for ELL is more concentrated around a single value.

Figure 10.4 illustrates two interesting situations. We show an excerpt of file `[0009] M4-09-SA`, which records an ensemble made by *tantã*, *pandeiro*, *agogô*, and

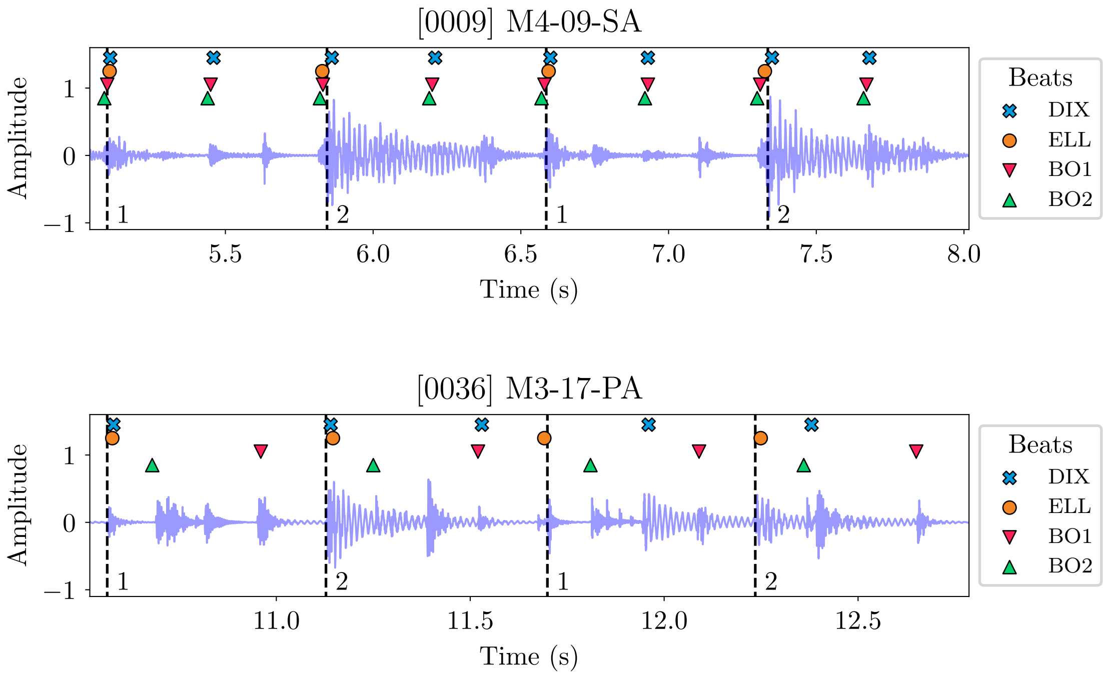


Figure 10.4: Beat tracking for two mixture tracks. The waveforms show two bars of the rhythmic patterns, with dashed lines indicating annotated beats. Other markers depict the position of beat estimates with DIX, ELL, BO1, and BO2.

surdo. ELL correctly predicts all beats (F-measure = 98.6%, CMLt = AMLt = 97.3%, inf. gain = 3.60 bits), whereas the other algorithms have doubled the meter (F-measure ≈ 66.7%, CMLt = 0.0%, AMLt ≈ 98.6%, and inf. gain ≈ 3.00 bits). We also display a typical case of phase error in track `[0036] M3-17-PA` (*repique*, *tamborim*, *tantã*). In this case, BO2 and BO1 have correctly estimated the tempo of the piece, but the beat phase is incorrect. The former has settled with the second sixteenth note of each measure, whereas the latter constantly tracks the fourth one. All metrics penalize this, except the information gain (3.78 bits for BO1 and 2.92 bits for BO2), which, as explained in Section 10.1, is insensitive to consistent phase errors. DIX shows a drifting behavior and, consequently, low values across all metrics (F-measure = 48.6%, CMLt = AMLt = 28.9%, inf. gain = 2.54 bits).

Since model BO2 jointly tracks beats and downbeats, we report here its downbeat tracking performance for the sake of curiosity: an F-measure of 1.6% and a CMLt of 4.6%. As in the previous investigation, what happens with almost the entirety of this subset of multiple-instrument recordings is that the BO2 model ends up consistently tracking the characteristic strong accent in the second beat of each measure, which is usually more heavily stressed by the *surdo* or the *tantã*. The consistency of this phase error in the estimation results in a high global downbeat tracking information gain of 4.0 bits, out of an approximate maximum of 5.4 bits.

10.4 Annotation of Beats in SAMBASET

Next, we consider the problem of beat annotation and estimation in SAMBASET. When compared to the purely percussive BRID, this dataset contains a lot more cues (e.g., harmonic changes) for the annotators and algorithms. This investigation was originally carried through when no beat annotations were available for SAMBASET. In order to measure the degree of challenge presented by this dataset in regard to beat perception, we have exploited a selective sampling technique, inspired by [289, 306, 307], to extract a series of recordings sampled at different levels of "difficulty". This difficulty is measured by the agreement between members of a committee of beat tracking systems. Notice that this procedure does not require annotations, just the beat time estimates produced by each algorithm. As a by-product, this analysis also gives us information about which methods are good candidates for estimating beats in the dataset. Initial beat estimates can then be produced with the best algorithm, and manually corrected by a human annotator afterwards.

We start this section by reviewing the procedure of [289, 306, 307].

10.4.1 Selective Sampling for Beat Tracking Evaluation

HOLZAPFEL et al. [289, 306] presented a method for selecting challenging music examples for the beat tracking task without ground-truth annotations using a query-by-committee approach [308] with a set of beat tracking algorithms.

When annotations a_j are available, the procedure for quantifying the difficulty of a piece is straightforward [306]. Authors first estimated, with each beat tracker B_k, a sequence of beat positions b_i^k for the piece. Then, they scored each sequence with a measure that takes the ground truth into consideration. Finally, they averaged all scores $S(b_i^k, a_j)$ to obtain the mean ground truth performance (MGP). In this regard, a piece was considered difficult if its MGP was low.

However, when no ground truth is available, the MGP cannot be used to infer the difficulty of the piece. The idea in [289, 306] is that, leveraging the beat committee, a piece can be considered "informative" or "interesting" by comparing the different estimated beat time sequences. First, they defined the mutual agreement between two estimated beat sequences $\{b_i^{k_1}, b_j^{k_2}\}$ output by two beat tracking systems, B_{k_1} and B_{k_2}, for the same piece as

$$\mathrm{MA}_{k_1,k_2} = D_{\zeta_{1|2}}, \quad k_1 \neq k_2, \tag{10.7}$$

where, in a simplified notation, $D_{\zeta_{1|2}}$ is the information gain (in bits) of the empirical beat error sequence of $b_i^{k_1}$ given $b_j^{k_2}$. For a committee of N beat trackers, they calculated the $N(N-1)/2$ different mutual agreements and averaged them all to obtain the mean mutual agreement (MMA) for that piece. HOLZAPFEL et al. [289] showed a correlation between low MMA and perceptual/musical properties that make tapping difficult for humans, inferring MMA can be used as a measure of the difficulty of the piece. They build a challenging dataset by selecting examples with MMA < 1 bit, given a committee of five beat trackers.

In a related paper [307], they showed that the MaxMA, i.e., the algorithm whose output presents the maximum mutual agreement with the rest of the committee, provides the most reliable estimation for a given music example. Using the same committee of five beat trackers, they conducted subjective listening tests to determine a perceptual threshold for acceptable quality of this chosen output given the corresponding MMA. They proposed an MMA threshold for the committee of 1.5 bits: an MMA ≥ 1.5 bits indicated that automatic beat tracking should be perceptually acceptable; whereas an MMA < 1.5 bits suggested inaccurate beat tracking.

10.4.2 Selection of SAMBASET Excerpts

In this thesis, we follow a similar approach to select samples of various difficulties for the committee of state-of-the-art algorithms. We have collected the implementations of 14 beat tracking systems, removing eight that were featured in the original work [289] but were now unavailable, and adding six others that were presented after that publication, most notably those provided in the `madmom` package [72].

The algorithms are implemented in different programming languages and, in a few cases, require different operating systems. We used the Python implementations of AUB (v0.4.9), ELL (`librosa` [305] v0.6.2), DEG and MFT (`Essentia` package [106] v2.1-beta5-dev), BO0, BO1, and BO2 (`madmom` package [72] v0.16.1); and the available releases of DIX (Java, v0.5.8), DAV (Vamp plugin in conjunction with the Sonic Annotator [309]), IB1 and IB2 (v1.0 binaries). Finally, the C++ implementation of KLA was kindly run by Martín Rocamora from a copy of the code provided by the author.

To obtain our selection of excerpts with different difficulty levels, we first extract 30-second audio segments from the middle of all the different *sambas-enredo* in SAMBASET. Then, from each of the 493 excerpts, we compute the MMA between the beat estimates produced by the committee of beat trackers. As the different collections in SAMBASET (HES, ESE, and SDE) have different characteristics (as discussed in Chapter 3), we perform the following analysis separately.

We sort the excerpts from each collection in ascending order by mean MMA, as presented in Figure 10.5. Then, for each collection, we determine P evenly spaced MMA values (including the maximum and minimum points), and select the excerpts

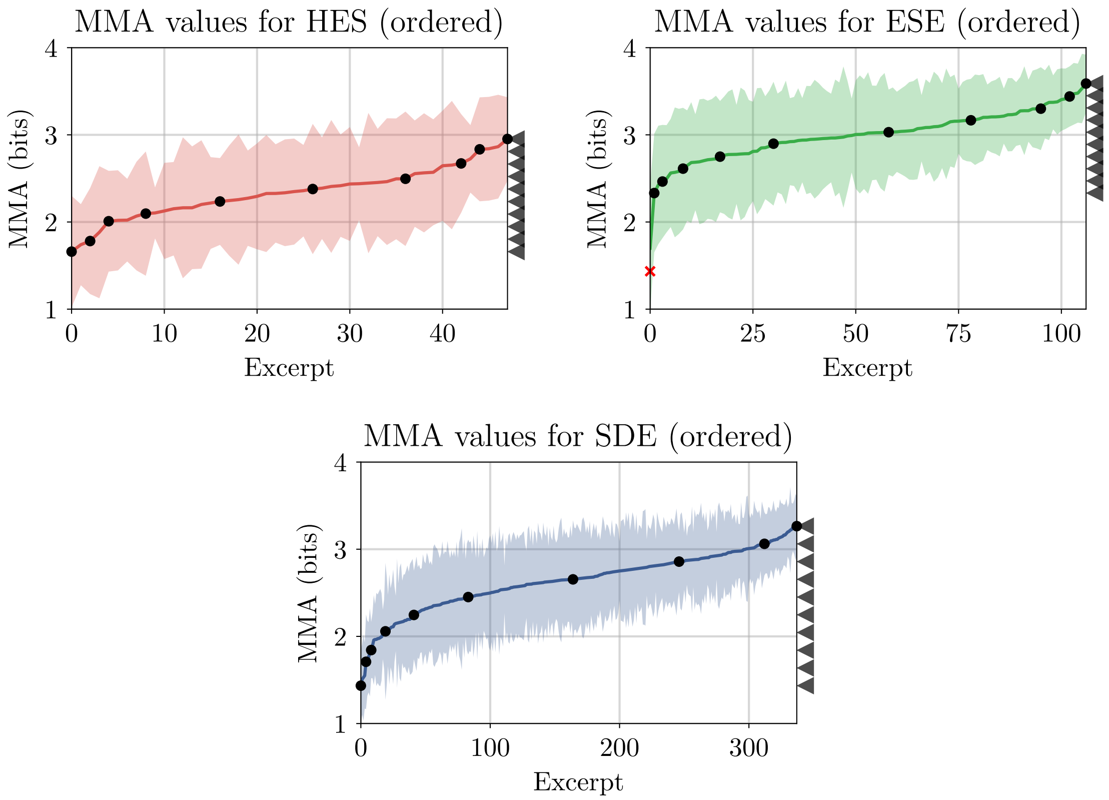


Figure 10.5: Collections sorted by mean MMA (solid line), with standard deviation (shaded region). Annotated samples (solid circles) were chosen as the closest to ten evenly spaced MMA values (solid triangles). One sample was treated as an outlier (cross) in ESE.

closest to each of these values. These are the excerpts that we manually annotate. There are two reasons for this procedure. The first one is that, by selecting the same number of samples from each collection, we compensate for the large imbalance between them (e.g., in SDE there are nearly seven times more excerpts than in HES), while ensuring that their unique characteristics are equally represented in the annotated subset (we recall that the expression and variability in HES is much higher than that in ESE, or SDE, as mentioned in Section 3.2.1). The second reason is that, this way, we guarantee that the beat tracking algorithms are compared within a group of samples where they have varying levels of agreement (and that would possibly provide a human annotator with a gamut of challenges). We use $P = 10$ and, in total, manually annotate 30 files, totalling just over 1900 beats. It should be noted that a moderate number of annotated samples is sufficient, since we are dealing with a single music genre, which considerably limits the range of variations between them.

10.4.3 Discussion of the Results

We see in Figure 10.5 that, in general, the 14 beat tracking systems show more agreement in beat time estimates for tracks in the ESE collection, followed by those in SDE, with HES in last. In fact, for over 50% of the tracks in ESE, the algorithms presented an MMA $>$ 3 bits, against slightly under 12% for SDE tracks and 0% in HES tracks in the same conditions. Considering a threshold at 2.5 bits, those percentages grow to 95%, 70% and 23%, respectively. This agrees with our overall impression that the HES collection is the most "flavorful", whereas ESE is less expressive (cf. Section 3.2.1).

With the manual annotations for the 30 excerpts, we can also estimate the individual performances of each algorithm, which are reported in Table 10.5 along with the mean across all methods. We observe that seven beat trackers perform better than the mean in all metrics, some of them outperforming the others by a large margin. For our dataset, the four best algorithms are BO1, BO0, DAV, and BO2.

Table 10.5: Performance of the beat tracking algorithms on the selected SAMBASET excerpts. The best performance for each metric is highlighted in bold. The five-member committee proposed in [289] is indicated by an asterisk.

Beat tracking model	CMLt (%)	AMLt (%)	*F*-meas. (%)	Inf. gain (bits)
Aubio (AUB) [310]	59.4	65.6	61.9	2.30
BayesBeat-HMM (KR1) [76, 265]	42.7	65.6	67.6	2.27
BayesBeat-PF (KR2) [76, 196]	47.6	52.9	58.0	2.25
*BeatRoot (DIX) [250]	79.4	82.8	86.4	3.15
Davies (DAV) [246]	97.2	97.2	97.5	3.66
*Degara (DEG) [311]	88.3	91.2	89.7	3.40
*Ellis (ELL) [245]	76.9	76.9	78.7	3.35
IBT causal (IB1) [312]	83.4	83.4	86.2	2.45
*IBT non-causal (IB2) [312]	51.1	90.8	80.0	2.49
*Klapuri (KLA) [244]	61.3	63.7	63.1	3.09
BeatTracker (BO0) [243, 313]	98.1	98.1	98.6	3.78
DBNBeatTracker (BO1) [240, 265]	**99.5**	**99.5**	**99.5**	**3.80**
DBNDownBeatTracker (BO2) [271]	94.0	97.3	97.1	3.68
MultiFeature (MFT) [314]	86.4	86.4	86.8	3.55
Mean	76.1	82.2	82.2	3.09

For the sake of comparison, we also evaluated the MMAs of the 493 excerpts with the five-member committee proposed in [289] and used in [307]: for 98.6% of the dataset the committee yields an MMA $>$ 1.5 bits; a single excerpt has MMA $<$ 1 bit. This indicates that, overall, SAMBASET excerpts are not very challenging to the algorithms in this reduced committee, which would provide a good number of acceptable estimates or a good level of confidence in the MaxMA estimation. This

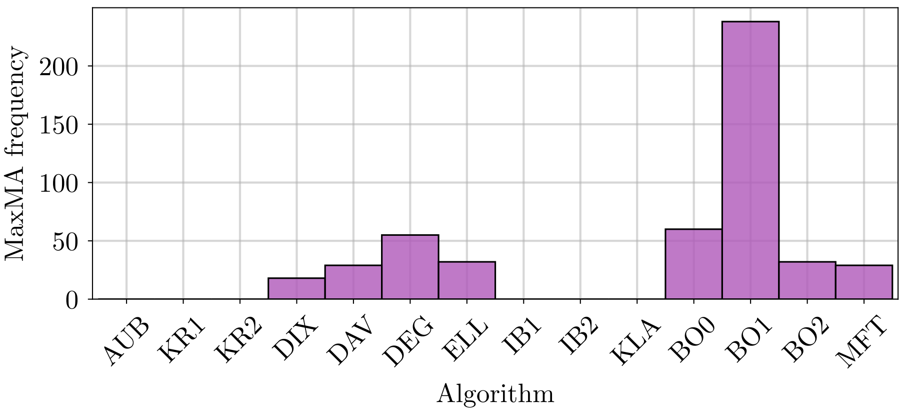


Figure 10.6: Histogram of the frequencies of each algorithm as MaxMA.

analysis of state-of-the-art algorithms indicates a safe approach to semiautomatically annotating beats in this dataset.

Instead of obtaining the MaxMA from the reduced committee of HOLZAPFEL et al. [289], we perform this analysis in our expanded committee. Figure 10.6 shows that BO1 is the MaxMA for over 54% of the excerpts in the dataset. We have therefore used this algorithm for producing initial guesses for beat positions in the entire tracks of SAMBASET, which have then passed through manual corrections before being stored as actual annotations.

10.4.4 Musicological Insights

Here we investigate the evolution of average tempo in *samba-enredo* recordings across the years as represented in the SDE collection. For each excerpt, we use the automatically detected beats provided by BO1, and compute the average tempo in bpm as the inverse of the mean IBI. Figure 10.7 shows the average tempo for every track in SDE, plotted against the release year.

Although no clear trend is apparent from the whole data, we can readily verify the existence of local trends in three different regions of the graph. The first region accounts for the years of 1994 through 1998, and corresponds to the end of an era of "live" recordings in the *Teatro de Lona* (Barra da Tijuca), a large circus-like tent. As Moehn reveals on his essay "The Disc is not the Avenue" [315], by then the recordings were being made with a large number of musicians from each *escola* (around sixty) as well as large choirs from the respective community.

A radical change took place in the production of the 1999 disc: the entire process was moved to the studio and the number of *escola* members was reduced, not only to cut costs, but also to regain control over the sound organization [315]. Producers wanted the disc to sound "clear" and, thus, constrained the creative liberties of the *bateria*'s directors (e.g., they were not allowed to choose the tempo of the

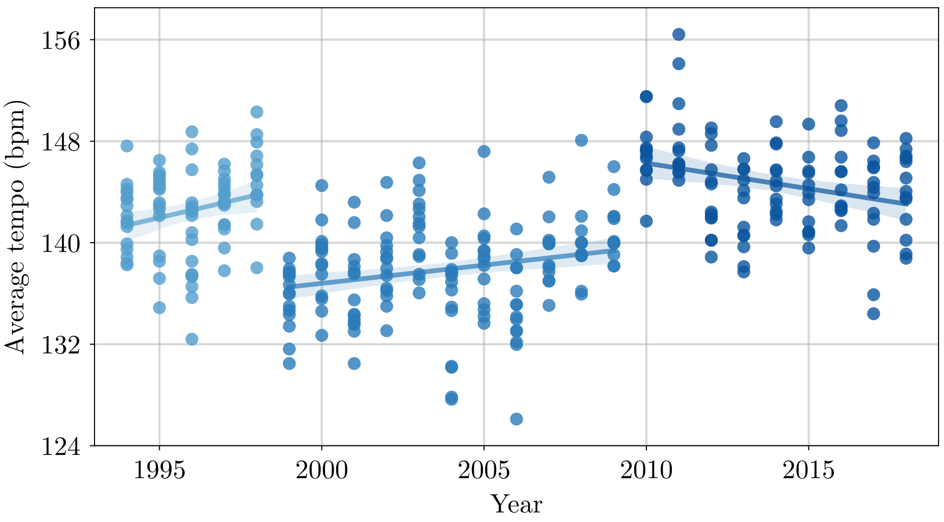


Figure 10.7: Average tempo across the SDE collection. Trend lines are shown for three distinct regions, along with the respective confidence intervals (shaded areas).

performance or to follow certain musical conventions that are common in live performances). This was an attempt to recover the disc's marketability (sales had been dropping in previous years), despite distancing it from the actual phenomenon of *samba-enredo* [315]. In 2010, "live" recordings were resumed, this time in the *Cidade do Samba* (Gamboa neighborhood). Producers retreated in their interference on the soundscape creation, and the *escolas* were able to reclaim the final saying in some aspects of the recording, such as the tempo. The larger space provided by the *Cidade do Samba* also lead to an increase in the number of musicians taking part in the recordings: more than 8 000 for the 2014 CD against 1 500 in the 1998 recording [315].

Therefore, we can say that the first and third regions of Figure 10.7 more closely represent actual *samba-enredo* performances. In particular, notice that the average bpm in the third region is above the averages in the other two regions. This can be seen as a direct translation to the digital media of the decisions to accelerate the live performances (and the marching pace), so that the *escolas* satisfy changes in parading time limits, as reported by many specialists [28, 58, 316].

10.5 Annotation of Downbeats in SAMBASET

In this final section, we present the results for downbeat tracking on the previously selected and annotated SAMBASET excerpts. We estimated downbeats with four systems. Three of them jointly estimate beats and downbeats, and were also featured in Section 10.4: KR1, KR2, and BO2. The other one is KRE, which solely estimates downbeats, and was used before in Section 10.2.

We report the average CMLt and F-measure, and also two alternate metrics that represent an evaluation with inverted targets, i.e., exchanging the labels of beats "1" and "2" in the annotations. These alternate versions of the original metrics are indicated as "starred metrics". Instead of looking at simple phase problems or octave errors, these metrics show us whether the downbeat trackers are wrongly treating beat "2" as beat "1", which is highly probable considering the strong contrametric accent found in *samba*. If this were the only problem with the downbeat tracking algorithm, it would be simple to manipulate the results to obtain accurate annotations: this would only require the knowledge of the beat phases and a relabeling procedure exchanging the labels "back".

We show the results for downbeat tracking on the 30 excerpts in Table 10.6. What can be seen from the table is that downbeat tracking performance is overall poor, but this cannot be entirely attributed to the models simply exchanging beats "1" and "2". This is only part of the issue, as the starred metrics show a moderate improvement (except for KRE) over the regular ones. In fact, looking into the results, we observe that finding the exact phase of the downbeat is challenging for all models, and in most cases we end up with phase errors of fractions of a beat. For this reason, we manually annotated downbeats in SAMBASET by selecting which estimated beats (after manual correction) corresponded to the labels "1" and "2".

Table 10.6: Downbeat tracking on the selected SAMBASET excerpts.

Downbeat tracking model	CMLt (%)	F-meas. (%)	CMLt* (%)	F-meas.* (%)
BayesBeat-HMM (KR1) [76, 265]	5.8	12.6	20.6	45.0
BayesBeat-PF (KR2) [76, 196]	3.2	9.6	3.8	32.2
BarTracker (KRE) [270]	0.0	41.7	0.0	28.1
DBNDownBeatTracker (BO2) [271]	0.0	19.4	3.3	47.2
Mean	2.2	20.8	6.9	38.1

Chapter 11

Contributions to Beat and Downbeat Tracking with Few Data

We presented in Chapter 9 a brief overview of the evolution of meter tracking models and, in particular, the recent developments in the state of the art with the introduction of deep learning methods. The data-driven supervised nature of these approaches conflicts with a current trend of expanding the frontiers of MIR and shifting its focus from Western music to other underrepresented music traditions [6, 63, 64, 317–319]. On one hand, these traditions often lack annotated data, which are costly to produce and usually require culturally-aware expertise. On the other hand, off-the-shelf general-purpose models typically underperform in these music genres, since they are unrepresented in the datasets used for training. We have exemplified this latter issue in Chapter 10, where we showcased the poor results of beat/downbeat tracking on BRID solos, beat tracking on BRID mixtures, and downbeat tracking on SAMBASET excerpts.

In this chapter, we present some contributions to beat and downbeat tracking that were originally reported in [22, 23]. Instead of proposing a novel architecture, we approach the problem of beat tracking from the perspective of an end user, i.e., someone that wishes to use state-of-the-art algorithms to annotate a dataset. We assume this user is not an expert in the music genre featured in the dataset or, at least, is not willing to manually annotate a large amount of data. Moreover, for this user, generating labels is not an end in itself; instead, it is a first step towards solving more complex, perhaps musicologically-inspired, questions that require prior knowledge of the temporal organization of music. If the object of study, i.e., the dataset, is composed of Western music that does not contain much expression (e.g., large tempo changes, *rubato*, swing), we expect that the beat and downbeat induction can be performed with out-of-the-box models and will be largely successful — after all, this is precisely the kind of music on which most of the state-of-the-art methods are trained. However, when dealing with styles that differ much from this kind of musical material, e.g., non-Western music with challenging rhythmic properties (microtiming or displaced phenomenological accents, among others), the user is bound to find some difficulties.

Assuming the user is willing to annotate a reduced number of audio examples that can then be used to train a state-of-the-art model, our solution is to refine and adapt this model to work well in that specific context, i.e., the particular music genre. The model should be trained on the small subset and work on the remaining data, facilitating the annotation process. This goes in hand with a series of recent efforts that, instead of developing "universal" models capable of performing equally well across various music genres (requiring large quantities of labeled data), have shifted towards adapting preexisting models to succeed on a subset of interest [283, 320], which can be as restricted as a single musical piece [17, 284, 290]. These efforts were discussed previously in Section 9.1.3.

Of course, many questions can be raised at this point, especially regarding the aforementioned "hungry" nature of deep-learning-based models set against the small annotated subset the user is asked to provide. We argue that, if the dataset is sufficiently homogeneous (in some senses), then there is no need for being wary of an "overfitted" solution — in fact, it is precisely what we require. Additionally, one could ask whether there is an appropriate way of selecting the subset to annotate, while at the same time minimizing the annotation effort and generalizing to the remaining samples in the dataset. We also tackle this problem here.

We start this chapter by quickly presenting two machine learning concepts that are relevant to our discussion. Then, we demonstrate our main premise, i.e., that it is possible to train meter tracking models with small quantities of data for "cohesive" music traditions and achieve a high performance comparable to that of traditional training schemes (which require much more data). We explore this adaptability in terms of data, performance, and computational cost. Lastly, we present a scheme for selecting informative samples from the dataset using the rhythmic features presented in Chapter 8 and a set of selection techniques based on representativeness and diversity. We show the results for applying this scheme in the task of beat tracking.

11.1 Active and Few-Shot Learning

The discussion presented in this chapter has parallels with two machine learning subareas. The first is the technique of few-shot learning [321–323], whose main idea is to train a model that is able to generalize to unseen classes at inference time by exploiting a few examples. Then, there is the concept of active learning, in which it is posited that a supervised learning algorithm performs better and with less data if it is allowed to choose its training samples [324]. These instances are selected from the most informative ones of the unlabeled dataset and sent to an oracle, typically a human user, that annotates them and forms a labeled training set, which in turn is used to update the model. Both paradigms have been increasingly used in audio and music-related tasks, most notably sound event detection [325–327], drum transcription [328], musical source separation [329], and music emotion recognition [330].

The most commonly used active learning sampling methods are uncertainty sampling and query-by-committee. In particular, we described in Chapter 10 how HOLZAPFEL et al. [289] explored this latter concept to create a committee of beat trackers that allow the determination of difficult-to-annotate examples. And, of course, we ourselves have exploited this notion to select a gamut from "easy" to "challenging" excerpts in SAMBASET. Other sampling methods take advantage of the internal structure of the input data distribution, either by analyzing local densities or by trying to construct a diverse labeled dataset. For example, SHUYANG et al. [325] used a k-medoids clustering technique to annotate and classify sound events. Similar to k-means, the k-medoids algorithm attempts to minimize the distance of points in a cluster to a referential data point (medoid) using a custom dissimilarity measure. SHUYANG et al. [325] leveraged it to cluster unlabeled sound segments and propagate labels from the medoids.

Moving away from MIR, an influential work by SU et al. [331] investigates the performance of several selective annotation methods as a first step before retrieving prompts for in-context learning of large language models. Authors included confidence-based selection as well as methods that promoted representativeness, diversity, and both.

Section 11.2 deals with minimizing the amount of data seen by a meter tracking model, similar to few-shot approaches, but at training time, consequently reducing the annotation cost. In Section 11.3, we build on the idea that a set of informative samples can be extracted from the input data distribution and that the small annotation effort is better employed over these data, much akin to the active learning paradigm.

11.2 Training with Few Data

We stand by the intuition that BRID is somewhat homogeneous with respect to timbre and rhythm, which is supported by our previous exposition of the main characteristics of *samba* in Sections 2.4 and 2.5, and of the dataset in Chapter 3. To further validate our findings, we will include, at this point, an analysis of *candombe*. We assume that the same intuition applies to datasets representing this music genre, which contains strong phenomenological accents and well-described rhythmic properties [38, 203].

Our objective in this section is to understand, using these music traditions — *samba* and *candombe*, whether it is possible to train meter tracking models with small quantities of data to obtain good tracking performances, and if so, how much data is actually needed. We train the models with increasing amounts of annotated data, ranging from less than a minute up to nearly 40 min, and compare the performance and computational cost of each configuration against the others. We contrast three different training strategies:

1. Training the model from scratch with either *candombe* or *samba* snippets;
2. Fine-tuning a model previously trained with 38 h of data from diverse datasets of Western music to work on either *candombe* or *samba*;
3. Same as the previous two, but training the models with data augmentation to artificially increase the "small data" input.

We use a state-of-the-art TCN-based model [275] for our experiments, as it presents an interesting compromise between performance and computational cost. We contrast the TCN against off-the-shelf models trained in/developed for Western music. To understand the adaptability and computational cost of deep-learning-based methods, we compare the TCN against another simple yet effective baseline, a Bayesian model (BayesBeat) [76], trained on the same non-Western data. In the following, we explain our methodology in detail.

11.2.1 Datasets

As aforementioned, we have selected datasets of two different Afro-rooted Latin American music traditions for our experiments. Due to the challenge it posed to both beat and downbeat tracking tasks in Chapter 10, we use in our experiments the acoustic mixtures subset of BRID. Just as a reminder, this corresponds to 93 short tracks (about 30 s each) of musicians playing together rhythm patterns found in *samba* and two of its subgenres (*samba de enredo* and *partido-alto*). To represent *candombe*, we use the Candombe dataset [63, 74], which consists of 35 recordings of *candombe* drumming, for a total of nearly 2.5 h. As discussed in Chapter 3, Candombe tracks have been segmented into non-overlapping 30-second excerpts to allow comparison with BRID, and, in each experiment repetition, we use a random sample of 93 *candombe* excerpts.

We also use six datasets to train a baseline TCN model: Ballroom [75, 76], Beatles [285], GTZAN [205, 235, 291], and RWC (Classical, Popular, Jazz) [292, 332]. Together, these correspond to over 38 h of audio data. We note that Ballroom and

GTZAN datasets comprise many diverse music genres (e.g., waltz, tango, rumba, rock, pop, country, etc.). We used the loaders from the `mirdata` Python package (v0.3.6) [333], except for a custom loader used with Ballroom.

11.2.2 Handling Small-Sized Datasets

For our experiments, in all cases, we first separate train and test data (80% and 20% of 93 excerpts respectively) to ensure a fair assessment of the models. Then, we divide the training data into six subsets, spanning {4, 9, 18, 37, 55, 74} 30-second tracks. We want to determine how differently the models adapt to small quantities of data, so we followed a similar approach to that of PINTO et al. [17] (see Section 9.1.3) to define the amount of data to be used for training. We select short 10-second temporal regions at the beginning of the audio excerpts, along with the corresponding beat and downbeat annotations, and discard the remaining audio portion. Then we split each of these regions into two adjacent 5-second parts, the first to be used for training and the second reserved for validation in the TCN model; alternatively, we use the entire 10 s for training the Bayesian model with off-the-shelf parameters. Considering that each snippet only lasts 10 s, those data subsets add up to approximately 40 s, 1.5, 3, 6, 9, and 12 min of annotations. The rationale behind this strategy is that, given a set of recordings of such Latin American music traditions in real-world applications, it would be reasonable to ask a user to annotate just a few seconds to a few minutes of data; of course, the less data needed, the better.

Since we are using very few data points to train the models, performance is strongly affected by data sampling. To mitigate this, we repeat all of our experiments 10 times with different seeds for the random data split generation, which means that models are trained 10 times with each of the differently-sized subsets. Note that selecting the best strategies for data sampling is not discussed in this section, and left to be addressed in Section 11.3. Test data are left uncut, i.e., we use the full 30 s, to keep compatibility with common model evaluation practices in meter tracking.

11.2.3 TCN Model

We use in our experiments the TCN multi-task model presented in [275] and reviewed in Section 9.1.2, particularly the open-source implementation of DAVIES et al. [241]. In this section, we focus on meter tracking, and ignore the tempo estimation head of the network. First, the TCN estimates the beat and downbeat likelihoods. Then, we use two different DBN implementations from `madmom` (v0.17.dev0) [72], one for beats and the other for downbeats, to infer the final positions of beats and downbeats respectively. Inferring them separately rather than jointly led to better results.

11.2.4 Training Strategies

Training from Scratch (TCN-FS)

When datasets have high similarity in terms of instrumentation, rhythmic patterns, and tempo, we anticipate that training a model from scratch with just a few data points will work well for most of the similar data.

Following the explanation in Section 11.2.2, we train one model per data subset, and repeat this 10 times with randomly initialized weights and seeds. We also consider the case in which all annotations are available and include the analysis of model performance when training with the entire 30-second excerpts. In this situation, we split the 74 train excerpts into train and validation (75%/25%). For every strategy, we use a learning rate of 0.005, and reduce it by a factor of 0.2 if validation loss does not improve after 10 epochs. We train for a maximum of 100 epochs, with early stopping if the validation loss does not improve after 20 epochs.

Fine-Tuning (TCN-FT)

We also approach the problem of meter tracking in a culture-specific setting from a "transfer learning" perspective. Following [17, 283, 320], we adapt a meter tracking model that has been previously trained for a different musical context. The intuition here is that if the model is first trained on a large dataset, even if it was built around Western music, it can serve as a good starting point for a model that is to be tuned for a specific out-of-training music tradition. This is a realistic approach since most of the available annotated data and trained models are Western-based. For this purpose, we trained a baseline TCN model on the Ballroom, Beatles, GTZAN, and RWC datasets. Due to the nature of its training data, this baseline model has to cope with many different meters, genres and acoustic conditions, which makes it a good starting point. We fine-tune it by using the same training procedure described previously with the initial learning rate reduced to 0.001, a fifth of the value used in the FS approach, as suggested in [17].

Data Augmentation (TCN-FTA, TCN-FSA)

Data augmentation techniques are useful for artificially increasing the number of training data points, which can be of great benefit in cases of low or insufficient data such as ours. In order to evaluate the impact of data augmentation in our models, we adopted a simple strategy inspired by [17, 275] in the experiments conducted with the TCN model, i.e., computing the input STFTs with different frame rates (varying hop sizes) so as to even out the distribution of tempi in the training set. Instead of randomly sampling tempi from a normal distribution around the annotated tempo, we selected a set of frames rates ±2.5% and ±5% around its value. This allowed us to increase our sample size five-fold while maintaining the same amount of annotation effort. Models obtained with the data augmentation procedure are labeled TCN-FSA and TCN-FTA.

11.2.5 Baselines

We include two types of baselines. Firstly, the BayesBeat statistical model [76] is used as a reference to the adaptability and computational cost of the TCN. It has fewer parameters, thus training is faster. The second type of baselines is composed of three off-the-shelf models — a signal processing technique, and two neural networks trained on Western music —; they illustrate the need for tailor-made solutions/adaptations in our context. Details are presented below.

BayesBeat

BayesBeat is based on the bar-pointer model, and simultaneously estimates beats, downbeats, tempo, meter, and rhythmic patterns, by expressing them as hidden variables in an HMM. An observation feature based on the spectral flux is computed from the audio signal and the observation model uses GMMs that are fitted during training to the feature values of each bin in a one-bar grid, so that rhythmic patterns are learned. Several patterns can be modeled, though is usually assumed that one pattern remains constant throughout a given music signal.

This model has a few hyperparameters that the user should choose depending on the music. Those are the number of rhythmic patterns, the type of feature to use (e.g., using only low, or low and high frequencies), and the feature grouping (e.g., how to compute the rhythmic pattern clusters), the tempo range, and the number of whole note subdivisions. In [76], it is reported that using two separate frequency bands ($\gtrless$ 250 Hz) helps finding the correct metrical level and is beneficial for beat and downbeat tracking. Considering more frequency bands did not seem to improve the results [76]. According to [264], using one rhythmic pattern per rhythm class is usually enough to achieve a good performance and provides the best results in most cases. Following these recommendations, we learn only one rhythmic pattern in two frequency bands.

Off-the-Shelf Baselines

We use the joint beat and downbeat tracking model of BÖCK et al. [271] (BO2) as implemented in `madmom` [72]. We note this LSTM-based model was trained in ten datasets spanning Western genres, and Carnatic, Cretan and Turkish music excerpts. We also include the beat tracker from ELLIS [245] (ELL), which uses a dynamic programming approach. As a final baseline, we include the same TCN architecture of BÖCK and DAVIES [275] trained on 38 h of Western music material from the datasets described in Section 11.2.1. We name this baseline “TCN-BL”.

11.2.6 Evaluation Metrics

We use as our main metric the F-measure, along with the CMLt and AMLt metrics. For the computational cost of the models, we simply report the time they take to train by using in-build timing functions in the code.

11.2.7 Performance of Models

Figure 11.1 shows the F-measure results for the TCN models trained for *samba* and *candombe* with different amounts of data using each of the training strategies, as well as BayesBeat, computed as the bootstrapped results of ten experiments (95% confidence) with different random seeds for each combination of model and data amount.

A first striking observation is that, for both beats and downbeats, the performance curve for most models has a small positive slope after a budget of 3 min, which means it is indeed possible to nearly achieve best model performance (which would require training with the complete dataset, represented by “all” in the figure) just by training with few samples. This is particularly true for the estimation of beat, for which models rapidly reach F-measure scores above 80% with 1.5 min of data in both Candombe and BRID for all configurations. This is an interesting result, meaning that not much gain in performance is expected with an increase in the number of annotations for these datasets. Therefore, an end-user could annotate about a minute of data and possibly obtain decent performance figures. The same holds for downbeat in the best performing models for Candombe, but not in BRID. For the latter, there is a clear gain when adding more data, which has to do with the differences between *candombe* and *samba*, as discussed in the following.

Differences Between *Candombe* and *Samba*

Observing the results in Figure 11.1, we see that the models tend to require more data to achieve better performance on BRID than on Candombe. Our first intuition behind this result is that, as *samba* has a greater combination of timbres and pitches than *candombe*, the decision about which snippets to annotate (i.e., the sampling) may be more critical for BRID than for Candombe. For instance, it might require ensuring the representation of timbre.

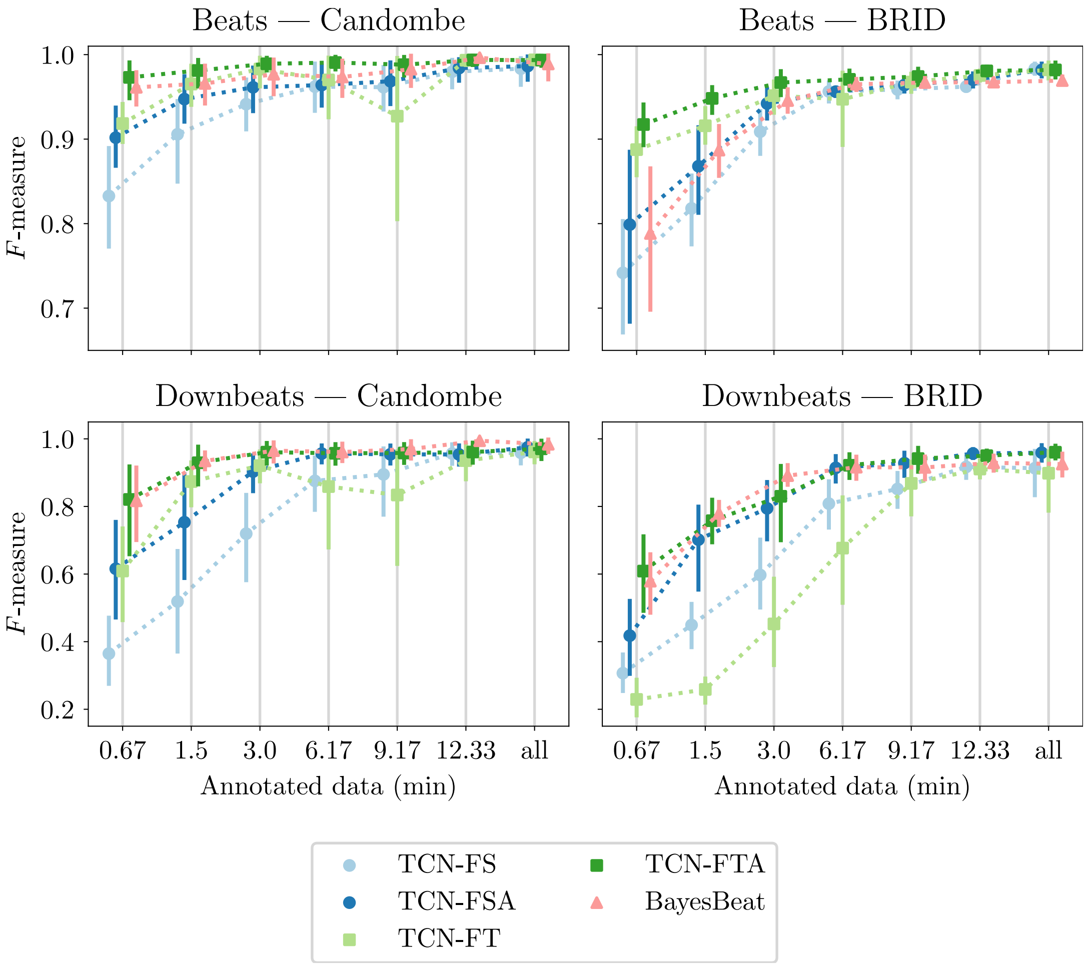


Figure 11.1: Performance of different model and training configurations. Label "all" indicates fully-annotated dataset.

Best Model Configuration

The best performing configuration for beat and downbeat tracking in both music traditions is the fine-tuned TCN model with data augmentation (FTA). Particularly, data augmentation produced significant improvement in performance for downbeat tracking in BRID. Interestingly, for the adaptive setting concerned in this section, the BayesBeat baseline is competitive with the TCN model, especially considering the computational cost (see Section 11.2.8).

Comparison with Off-the-Shelf Benchmarks

Table 11.1 shows the performance of the TCN and the BayesBeat baseline for different data subsets, namely the two smallest and the largest subsets, and the full dataset. It also shows the performance of the three off-the-shelf baselines presented in Section 11.2.5. In alignment with other works [63] and our previous findings (Chapter 10), the models trained with Western music (TCN-BL and BO2) perform very poorly in Candombe, and reach only about 66% beat F-measure in BRID,

Table 11.1: Mean F-measure (F) and continuity scores (CMLt = C, AMLt = A) in beat and downbeat tracking tasks across both genres.

	Candombe						BRID					
Model	Beat			Downbeat			Beat			Downbeat		
	C (%)	A (%)	F (%)	C (%)	A (%)	F (%)	C (%)	A (%)	F (%)	C (%)	A (%)	F (%)
BayesBeat (0.67)	95.0	95.0	96.1	82.2	92.0	81.7	70.5	74.2	78.8	57.4	72.9	57.9
BayesBeat (1.50)	95.3	95.4	96.6	93.8	94.8	93.5	82.3	85.2	88.7	76.8	83.9	77.9
BayesBeat (12.33)	99.6	99.6	99.6	99.8	99.8	99.4	93.5	96.0	96.7	92.5	94.9	92.9
BayesBeat (all)	98.6	98.6	98.9	98.8	98.8	98.4	94.0	96.0	96.9	92.0	95.3	92.5
TCN-FSA (0.67)	88.2	89.1	90.2	56.5	70.4	61.6	74.4	78.2	79.9	24.3	60.1	41.8
TCN-FSA (1.50)	94.3	95.1	94.7	71.9	81.5	75.3	82.0	86.7	86.8	58.9	77.9	70.2
TCN-FSA (12.33)	97.9	98.0	98.4	95.8	98.0	95.4	94.3	96.3	97.0	92.8	97.5	95.7
TCN-FSA (all)	98.4	98.4	98.6	97.4	98.4	97.5	96.0	98.6	98.2	95.0	96.6	95.9
TCN-FTA (0.67)	96.7	96.7	97.3	81.8	89.7	82.1	85.3	93.0	91.7	28.5	81.5	60.9
TCN-FTA (1.50)	98.0	98.0	98.1	94.1	97.9	92.9	89.7	95.3	94.8	52.2	89.2	75.8
TCN-FTA (12.33)	99.4	99.4	99.4	96.5	99.5	96.1	95.8	97.8	98.1	92.0	97.2	95.0
TCN-FTA (all)	99.4	99.4	99.4	97.4	99.5	97.1	96.3	98.4	98.2	96.1	97.7	96.0
TCN-BL	11.1	18.7	15.9	14.9	31.9	4.1	46.5	65.6	60.0	5.9	52.5	9.6
ELL	34.8	38.1	38.0	-	-	-	82.3	87.6	87.1	-	-	-
BO2	11.7	14.4	11.5	26.7	40.3	0.5	46.9	76.0	66.4	5.2	66.6	2.0

both significantly lower than the typical performance of the same models in Western music datasets. ELL scores considerably better in BRID, comparably to the results in the previous chapter, but is not as consistent in Candombe. This shows the necessity of adapting meter tracking models to these music genres, as even the models trained with the smallest subsets of data (0.67 and 1.5 min) outperform the baselines.

11.2.8 How Much Time do the Models Take to Train?

Our analysis is motivated by the adaptation of meter tracking models in real-world use cases. Therefore, it would be interesting if this adaptation could be done quickly, i.e., consistently with moderate computational demands. In this regard, we estimate the time each model configuration takes in training, and contrast it with the BayesBeat baseline. Figure 11.2 shows how the train duration varies with the size of the training set for BRID (very similar results were obtained for Candombe). To showcase the accessibility of our method for researchers and practitioners without access to high-performance computing resources, all experiments were strictly conducted using the CPU (Intel Xeon CPU E5-2403 @ 1.80GHz).

The TCN has a minimum training time of about 100 s for the smallest subset. Among the TCN configurations, the most expensive ones use data augmentation. This makes sense given that more data is used for training. As expected, the BayesBeat trains significantly faster than the TCN, taking on average 1.62 s to train with 0.67 min of data, and being in the order of 50 to 350 times faster than the TCN when data augmentation is not used. This big gap in computing time, together with the results of Figure 11.1 and Table 11.1, makes BayesBeat an overall good alternative for adapting meter tracking to these Latin American music. We observe that all configurations take about the same inference time, around 20 s for the full test set.

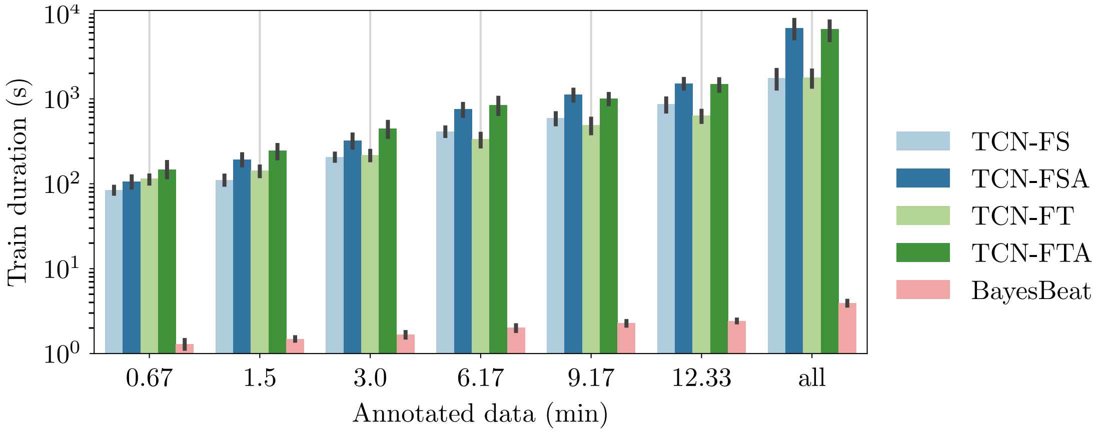


Figure 11.2: Training time for the different amounts of data.

11.2.9 When Can We Train With Small Data?

Our intuition is that the more variability in the data (in terms of meters, rhythmic patterns, and instrumentation), the harder it is for a model to learn with small data. This aligns with our experiments on the adaptability of these methods to *samba* and *candombe*. To have a more quantitative understanding of this, we derived a bar profile for each type of music. First, we extracted a feature map from each excerpt using the beat/downbeat annotations to time-quantize a locally normalized onset strength function [334] at the tatum scale — this was done with the `carat` toolbox [335] considering the tatum duration as one quarter of the time-span between successive beats. Then, for each dataset, we summarize these feature maps across time using the downbeat annotations, which results in a distribution of feature values per tatum in a bar (16 tatums in $\frac{4}{4}$ meter, 8 tatums in $\frac{2}{4}$). To allow an analysis of these profiles in different regions of the spectrum, we compute the onset strength in two frequency bands (from 20 to 200 Hz; and above 200 Hz). We present these distributions as violin plots in Figure 11.3 for the Candombe and BRID datasets. To enrich our discussion and prepare the following section, we have also added these bar-wrapped tatum-quantized onset strength distributions for tracks in $\frac{4}{4}$ meter of the Ballroom dataset.

We verify in this figure that, for some tatums, strength distributions are concentrated around 1 or 0, indicating a strong characteristic accent or lack thereof at that point of the bar, respectively. High variance, in its turn, means "fuzzyness" in the rhythmic pattern, which could justify the difficulty in learning that rhythm with a meter tracking model, specially under small data constraints.

Samba, which has eight sixteenth-note tatums per bar at a $\frac{2}{4}$ meter, is known for having a strong metrical accent at beat "2", which we may readily identify in the low-frequency channel for BRID at tatum 5. The first beat (tatum 1) of this profile also has a high median value but is less "deterministic" due to its high variance. In turn, the low-frequency profile of Candombe displays a high-variance downbeat, no accent on beat "2" (tatum 5), and strong accents on beats "3" and "4" (tatums 9 and 13), but a strong contrametric accent at tatum 4. These characteristics could help explain why the off-the-shelf beat tracking models, which expect beats to be accentuated, perform worse on Candombe. Looking back at the BRID profile, we see that tatums 2 and 3 show small standard deviations and correspond to "off" tatums; together with beat "2", they make three out of eight tatums that exhibit very small variance in the low channel. In the Candombe profile, besides tatum 4, tatums 2, 3, 7, 8, 9, 14, and 16 also present small variance. This abundance of "anchor" points could justify why adaptation in *candombe* comes with little data.

In Ballroom, we clearly see that beats are distinct for having high strength and

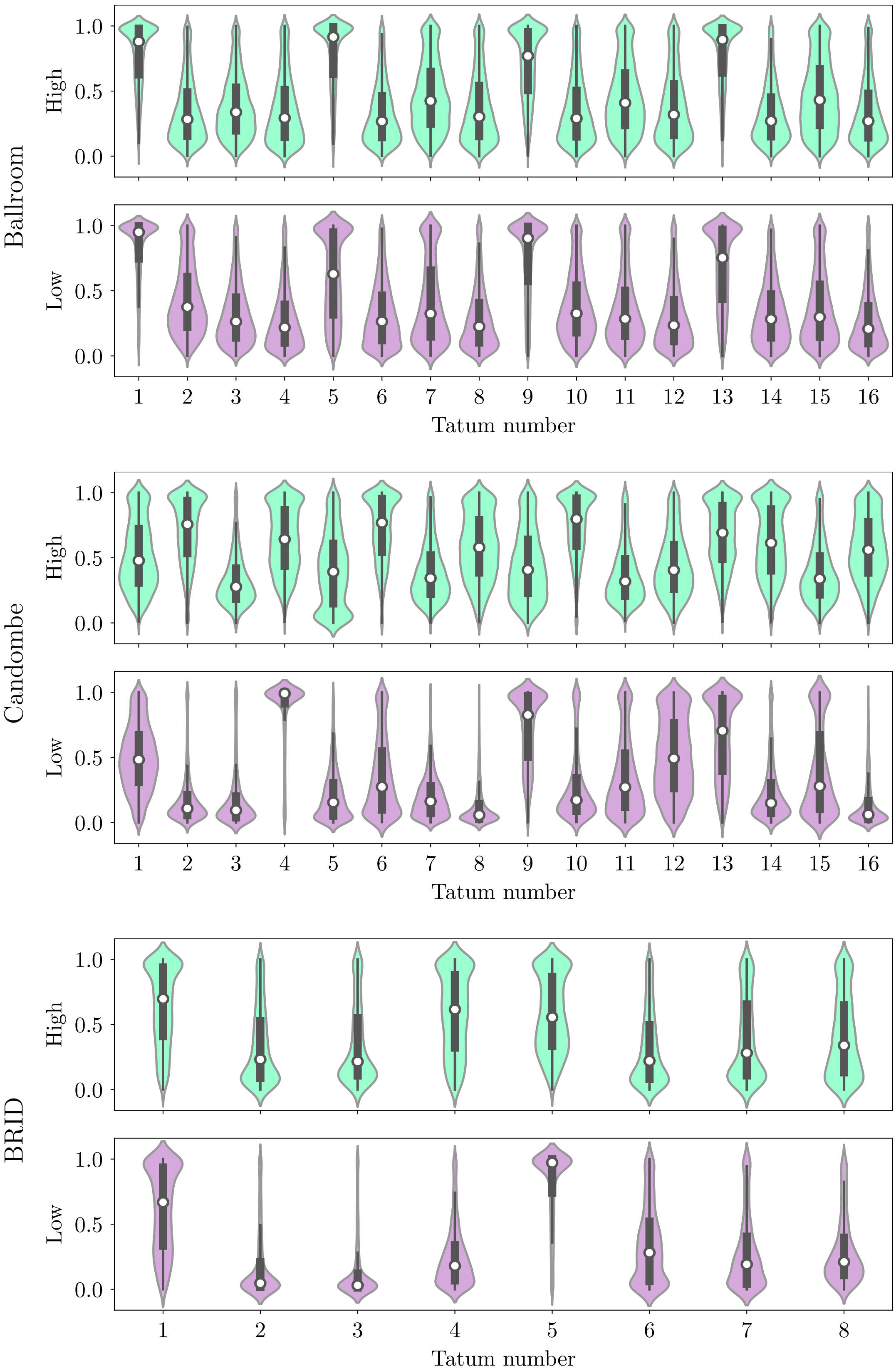


Figure 11.3: Tatum strength distribution per frequency band for Ballroom (just 4/4 tracks), Candombe, and BRID.

low variance in both channels, whereas the rest of the tatums show no clear trend. Its fewer reference points (only 4 out of 16) could however pose a challenge for learning models. Furthermore, beat patterns (the combination of the four tatums in-between beats, including the beat itself) are also indistinguishable from one another, which could aggravate this matter specifically for downbeat tracking. To test these observations, we trained a set of models from scratch for Ballroom using the same methodology used for BRID and Candombe. Results are depicted in Figure 11.4. The performance results correlate with the intuition that Ballroom is a more challenging dataset, particularly for beat tracking, given that it comprises multiple genres, and also that for learning beat and downbeat more data would be needed.

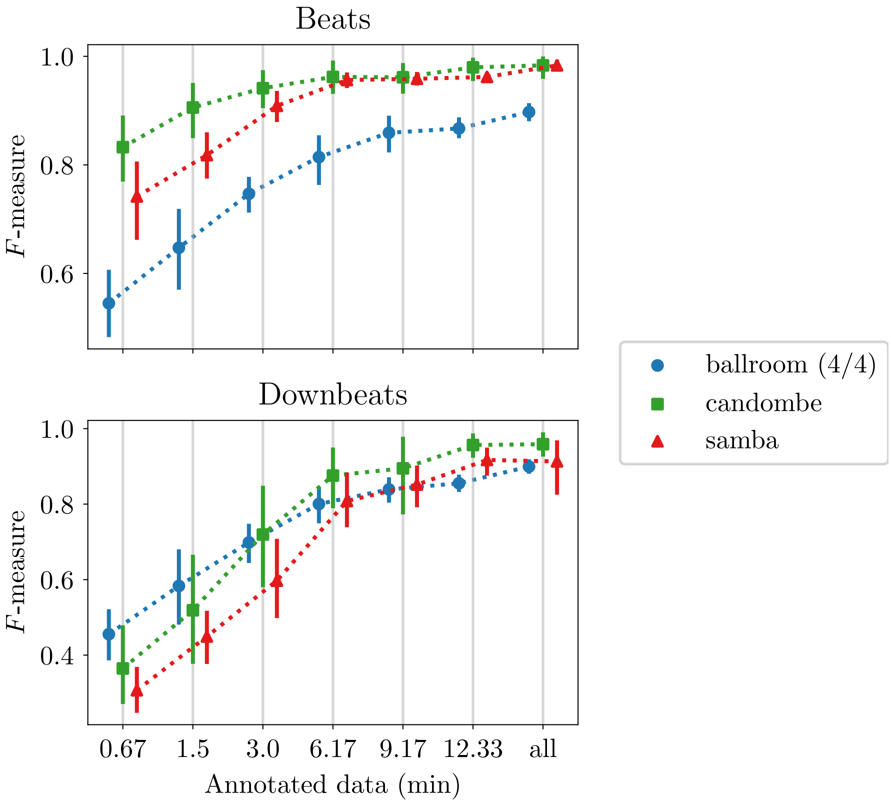


Figure 11.4: TCN-FS performance in Ballroom.

11.3 Selective Annotation of Few Data

After verifying the validity of our approach to training meter tracking models with limited data, specifically in music datasets with a high degree of self-similarity (homogeneous), we now address the issue of data selection.

We present in this section an offline data-driven framework that allows the selection of informative training data for state-of-the-art beat tracking models, under a constrained annotation budget and given this homogeneity condition. At the first step, we extract a rhythmically meaningful feature from each track of the dataset. The second step consists of selecting, with an appropriate sampling technique, the portion of the dataset that should be annotated. The pipeline resumes and the annotated samples are used to train the model so that beats can be estimated for the remaining unannotated tracks.

Our experiments investigate the importance of selection in low-data scenarios; explore the suitability of two rhythmic descriptors; and evaluate the performance of a TCN-based state-of-the-art beat tracking model trained with data selected according to four different sampling schemes as well as a random selection baseline. To validate our methodology, we perform the data selection on BRID and Candombe, and compare the results with those of the Ballroom dataset. We also investigate the different rhythmic properties of each dataset and dive deeper into the meaning of “homogeneity”.

11.3.1 Data Selection Methodology

Our data selection pipeline is a two-step process. First, by using a rhythmic descriptor, we represent each track i in a dataset of size N as a vector $\mathbf{x}_i$. At this stage, we use two feature representations — STM and OPH — exactly as described in Chapter 8. Notice that the former is a tempo-robust descriptor, while the latter is sensitive to tempo variations. We investigate both features following the conclusions of HOLZAPFEL et al. [224] that, depending on the tempo distribution of the dataset, it might be better to use tempo-robust or tempo-sensitive features.

At the second step, given a user-defined annotation budget, we perform sampling in the feature space using selection techniques based on representativeness and diversity. The examples selected by the algorithm are then annotated by the user, and form the training set for a beat tracking model. This is summarized in Figure 11.5.

We assume that, under annotation budget constraints, if we wish to achieve a good[1] beat tracking performance for a given dataset represented by a set of points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ in the rhythmic feature space, the most informative training samples can be retrieved by an appropriate model of the input distribution. Therefore, our objective is to select samples to be annotated and serve as training data for a state-of-the-art beat tracking model. The cardinality of the resulting labeled set $\mathcal{L}$ is the labeling budget M, and all remaining samples in the unlabeled set $\mathcal{U}$ serve as test set for the model. The selected data have to be informative in the sense that, by training the tracking model on the examples in $\mathcal{L}$, we should achieve good evaluation results over tracks in $\mathcal{U}$. We will abuse our notation and refer to points in $\mathcal{X}$, $\mathcal{L}$, and $\mathcal{U}$ as both the tracks and their corresponding features.

[1] In the rest of this chapter, “good” will mostly be used in place of “better than the performance of an equivalent system trained on randomly selected data”.

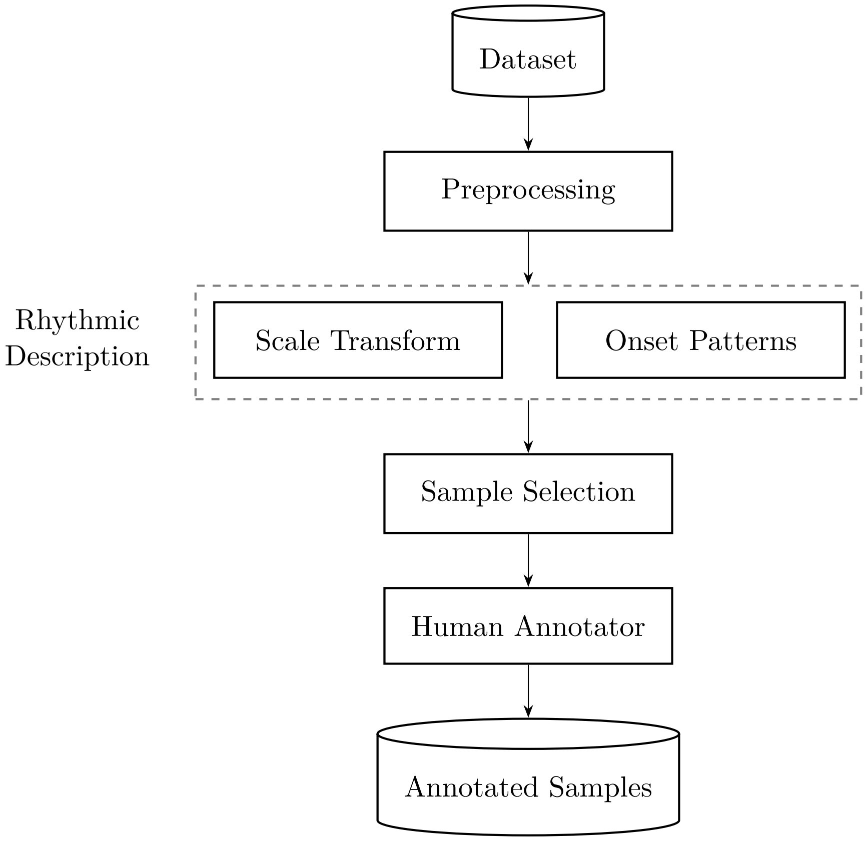


Figure 11.5: Construction of a set of annotated samples.

11.3.2 Selection Schemes

We borrow some selective annotation techniques from [331] and [325], which are presented below.

Fast Vote-k (VTK)

One of the selection techniques we use is the graph-based selective annotation method, proposed by SU et al. [331], which determines a set of simultaneously diverse and representative examples given the annotation budget. First, a directed graph $G = (V, E)$ is created where each feature vector in $\mathcal{X}$ is a vertex in V. Edges E are defined from each vertex to its k nearest neighboring vertices in the embedding space, according to the cosine similarity. We start with $\mathcal{L} = \emptyset$ and $\mathcal{U} = \mathcal{X}$. Then, at every iteration, unlabeled vertices $\mathbf{u} \in \mathcal{U}$ receive a score

$$\varsigma(\mathbf{u}) = \sum_{\mathbf{v} \in \{\mathbf{v} | (\mathbf{v}, \mathbf{u}) \in E, \mathbf{v} \in \mathcal{U}\}} w(\mathbf{v}), \tag{11.1}$$

where

$$w(\mathbf{v}) = \rho^{-|\{\boldsymbol{\ell} \in \mathcal{L} | (\mathbf{v}, \boldsymbol{\ell}) \in E\}|}, \tag{11.2}$$

with $\rho > 1$ and $|\cdot|$ the cardinality of the set. The score $\varsigma(\mathbf{u})$ depends on the vertices $\mathbf{v}$ from which $\mathbf{u}$ can be reached. Each $\mathbf{v}$ contributes with its weight, $w(\mathbf{v})$, which is small for $\mathbf{v}$ close to vertices already in $\mathcal{L}$. These two properties account for representativeness and diversity in the selected set, respectively. At every iteration, a vertex

$$\mathbf{u}^* = \arg\max_{\mathbf{u}\in\mathcal{U}} \varsigma(\mathbf{u}) \tag{11.3}$$

is moved from $\mathcal{U}$ to $\mathcal{L}$, until $|\mathcal{L}| = M$. At the first iteration, the algorithm selects the most reachable vertex. We experimented with a few values for $k \in \{3, 5, 7, 10, 15, 20, 25, 30\}$, but ended up choosing $k = 5$ as it provided good results across all budgets and datasets.

Diversity (DIV)

Another technique we use in this work focuses on maximizing the diversity of the labeled set. Following SU et al. [331], beginning with a random sample, at every iteration $t \leq M$ the furthest sample from those already in $\mathcal{L}$ is selected.

Maximum Facility Location (MFL)

We also employ a representativeness selection based on an algorithm by LIN and BILMES [336] adapted for the facility location problem [331]. This greedy algorithm optimizes the representativeness of selected samples by measuring the pairwise cosine similarity between embeddings. At every iteration $t \leq M$, it selects the most representative example $\mathbf{u}^*$ as

$$\mathbf{u}^* = \arg\max_{\mathbf{u}\in\mathcal{U}} \sum_{j=1}^{N} \max\{0, s_{\cos}(\mathbf{x}_j, \mathbf{u}) - \rho_j\}, \tag{11.4}$$

where ρ_j is the maximum similarity of $\mathbf{x}_j$ to the selected samples. At every step, ρ_j, which starts as $-1\forall j$, is updated to $\max\{\rho_j, s_{\cos}(\mathbf{x}_j, \mathbf{u}^*)\}$.

k-Medoids (MED)

We include a data selection scheme inspired by the work of SHUYANG et al. [325]. We first cluster data with a k-medoids algorithm. Since the medoids returned by this clustering algorithm are the center points that represent local distributions and, at the same time, reside in distinct places of the feature space, we set $k = M$ and directly use the medoids as the set of selected samples $\mathcal{L}$. As in the case of vote-k, this selection scheme aims to provide simultaneously diverse and representative examples for training.

Random (RND)

All selection schemes are compared to a random baseline: a subset of size M is randomly selected from the dataset to make up $\mathcal{L}$.

By leveraging representativeness and diversity, all of the presented selection schemes (except random sampling) are indirectly conditioned by the input data distribution. Since the musical properties (e.g., global tempo, rhythmic patterns, complexity, density) pertinent to our task and features vary differently according to genre and performer, we expect that each distinct dataset might benefit from a different sampling method. In the case of a highly homogeneous dataset, it is perhaps better to annotate a set of more diverse examples. For datasets with increasingly more heterogeneous data distributions, representativeness should be weighted more. Moreover, if the dataset is unimodal and highly homogeneous (e.g., composed of a single music genre and displaying little variance in its rhythmic properties), we would expect to observe little improvement in using smart selection schemes over training on randomly selected data. For less homogeneous (still unimodal) datasets, a proper smart selection should be able to systematically provide better training examples for beat tracking. Finally, in a dataset containing different genres with particular characteristics: (1) a single selection scheme might not be effective for all genres; (2) on average, random sampling will select more examples from the most populated genres, possibly overlooking the less populated ones. In contrast, when genres have the same number of tracks, due to (1) we should not expect large improvements in employing data selection methods.

11.3.3 Training Strategy

In all experiments, we train the TCN model from scratch with the labeled set $\mathcal{L}$ output by the data selection stage. We stand on the idea, verified in Section 11.2, that one can overfit a neural network model for a specific musical genre by training it with few samples, provided the dataset is sufficiently homogeneous in terms of instrumentation, rhythmic patterns, and tempo. Unless otherwise specified, we evaluate the results over the remaining data ($\mathcal{U}$). This matches our real-world application, where an end user would employ a small annotation effort (with a budget of M tracks) and train a model on the labeled data hoping to obtain good beat time estimates for the remaining unlabeled tracks. The annotation step by a human-user is emulated by retrieving the corresponding ground truth annotations.

Previously in Section 11.2 (following [17]), we extracted a single 10-second segment from each musical sample and split it into two disjoint 5-second regions, the first reserved for training and the second for validation. This allowed for more control when tuning the model's hyperparameters, despite sacrificing half the available information. We now assume a slightly different approach: we split each audio track from $\mathcal{L}$ in half and use the first and second halves as training and validation, respectively; test data are not cut. We use the same training parameters as seen in Section 11.2.4 for the FS training scheme.

11.3.4 Dataset Homogeneity

Preceding our experiments, we investigate tempo and rhythmic variability of tracks from each dataset.

Figure 11.6 presents the datasets' global (per-track average) tempo distributions smoothed by a Gaussian kernel density estimation technique. Candombe exhibits a slim distribution, averaging 132 bpm (8 bpm standard deviation), while BRID is approximately bimodal, whose peaks at 95 and 130 bpm can be respectively associated with *samba/partido-alto* and *samba de enredo* subgenres. Unsurprisingly, Ballroom's multi-genre characteristic is disclosed by multiple modes; individual distributions are described by [76].

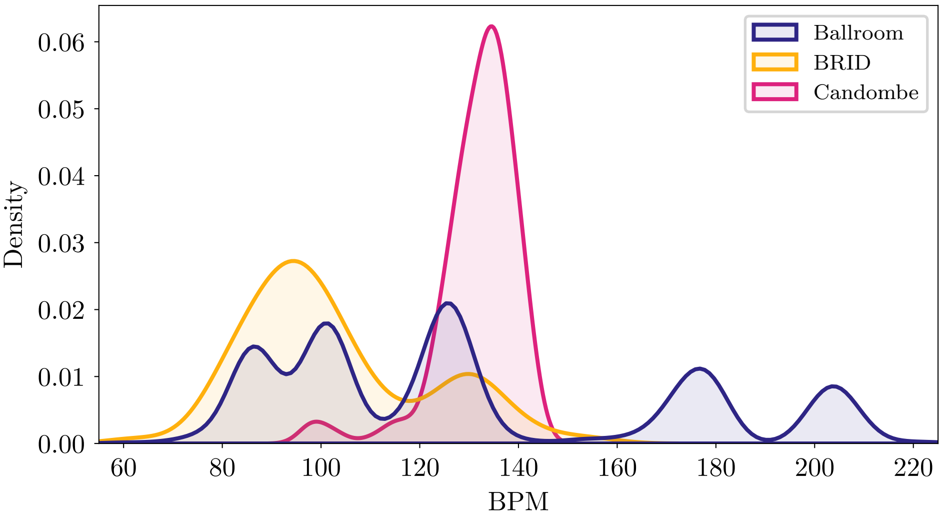


Figure 11.6: Global tempo distributions.

A representation of the rhythmic patterns across all datasets is displayed in Figure 11.7. The STM feature was obtained from each track following the procedure described in Section 8.3. Then, manifold learning with UMAP [227] was used to reduce the feature space dimension from 400 to 2 using the cosine distance as a metric. UMAP diagnoses that Ballroom patterns mostly lie in regions whose local dimension is estimated as high, which means they are less accurately represented in this embedding, and thus display greater rhythmic variation than can be represented in two dimensions. Candombe has a small set of outliers but is mostly represented in a compact structure, whereas BRID, despite having fewer examples, is more spread out in the embedding space. Interestingly, the subset of Ballroom located near BRID and Candombe is mostly composed of tracks labeled as samba, with few examples of jive.

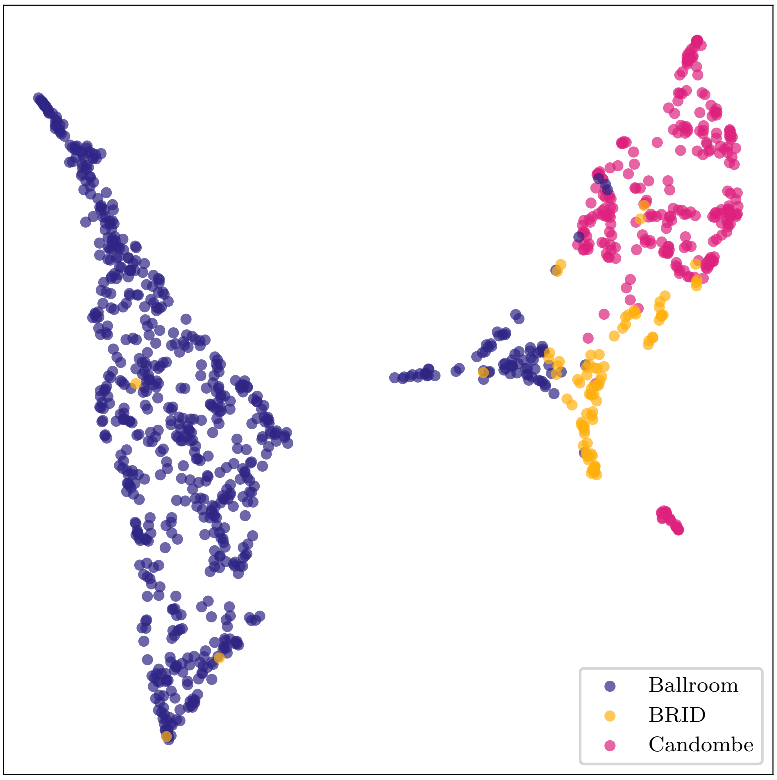


Figure 11.7: STM features embedded by UMAP (cosine metric, n-neighbors = 15, min-dist = 0.1).

11.3.5 State-of-the-Art Results Without Selection

To contextualize the outcomes of our experiments, we present in Table 11.2 the beat tracking performances on BRID and Candombe of models using the architecture of [275] under three different training schemes:

- "Pre-trained": results of the TCN-BL model from Section 11.2.5 — network trained on 38 h of Western music material from six datasets (including Ballroom), and tested on the entire BRID and Candombe datasets. Results were extracted from Table 11.1.
- "Fine-tuned": the "pre-trained" model that we fine-tuned for each dataset with 3 min of randomly selected data (tracks were split in half for training and validation), tested on the remaining data. We used 10 random selections,

Table 11.2: Performance figures of the state of the art (with random data selection): mean (standard deviation) in %.

	Beat F-measure (%)	
Model	BRID	Candombe
Pre-trained (TCN-BL)	60.0	15.9
Fine-tuned (3 min)	93.4 (3.4)	98.2 (1.1)
Trained from scratch (all)	98.9 (1.2)	99.8 (0.3)

30 training seeds, and the fine-tuning parameters as seen in Section 11.2.4 for the FT training scheme.

- "Trained from scratch": the TCN model initialized randomly and trained for each dataset on full 30-second tracks, using an eight-fold cross-validation scheme. One fold was used for testing, one for validation, and six for training. The training was repeated until all folds had been used for testing. The training parameters are the same as the main model and the FS training of Section 11.2.4.

11.3.6 Experiment 1: Does Sampling Matter?

In this first experiment, we assess how beat tracking performance is affected by random sampling of the training sets in low-data scenarios. Depending on the size of the dataset and the annotation budget, it might not be feasible to explore all possible training sets combinations. We choose to focus on the BRID dataset, which has the smallest number of tracks of all datasets, so that we can be able to survey a larger proportion of all random combinations. In this sense, we set the annotation budget to $M = 4$ tracks, which yields around 3 million possible combinations of four distinct elements out of 93 total tracks. Then, we select 1000 of these combinations with an in-house algorithm that forces all tracks in the dataset to be about equally represented overall. We use each unique combination of four tracks (about 40 s of annotations) to train/validate the TCN model, which we evaluate over the complementary test set of 89 files. We repeat each training 30 times with different randomly initialized weights and seeds.

Figure 11.8 shows the averages and standard deviations for the performances of all trained models in ascending order of mean beat F-measure. We note that mean beat F-measures range from 46.5% to 90.1% depending on the training set, with the 5th and 95th percentiles corresponding to 61.0% and 85.4%. A mean F-measure of 74.4% is achieved on average.

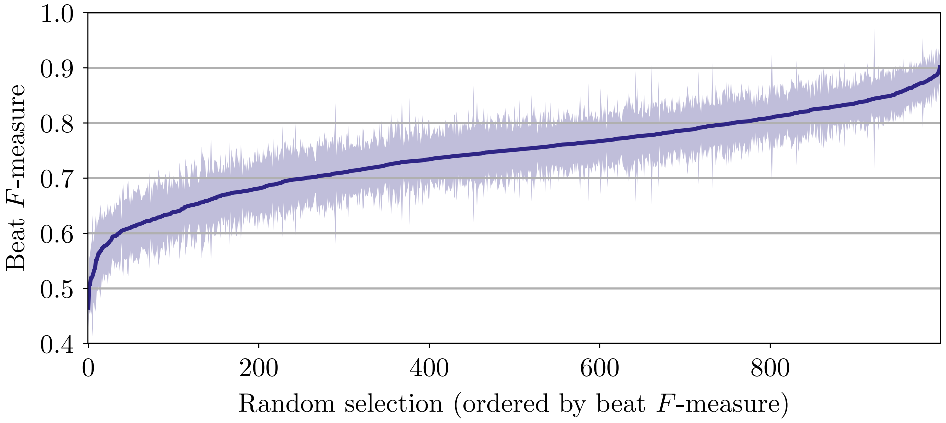


Figure 11.8: Random data selections on the BRID dataset ordered by mean F-measure, showing standard deviations (shaded area).

This experiment shows the importance of adequate data selection for the TCN model when dealing with low-data training scenarios. With an annotation budget of $M = 4$ tracks, we observe a considerable improvement of about 16 percent points over the average case and 44 percent points over the worst combination when estimating beats on the BRID dataset.

11.3.7 Experiment 2: Feature Structure

In this second experiment, we investigate the local structure of the feature space generated by each rhythm description feature (STM and OPH) and its capability of conveying meaning for the data selection scheme. For this purpose, considering the distribution of a dataset in the feature space, we analyze the performance of the TCN model when trained with points sampled from different regions around single test tracks. We hypothesize that regions closer to the test sample provide better training examples, thus yielding good beat tracking results. Again we set the annotation budget to $M = 4$, but this time we experiment with all datasets.

The regions are limited by concentric hyperspheres centered at each test sample, whose radii depend on the distribution of the dataset in the feature space. If Q_1, Q_2, and Q_3 are the first, second, and third quartiles of the pairwise feature distances of points in the whole dataset, respectively, we define $\mathcal{R}_1^j$, $\mathcal{R}_2^j$, and $\mathcal{R}_3^j$, the regions in increasing distance from a given test file, $\mathbf{x}_j$, as

$$\mathcal{R}_1^j := \left\{\, \mathbf{x}_i \,\middle|\, \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq Q_1 \,\right\} \tag{11.5}$$

$$\mathcal{R}_2^j := \left\{\, \mathbf{x}_i \,\middle|\, Q_1 < \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq Q_2 \,\right\} \tag{11.6}$$

$$\mathcal{R}_3^j := \left\{\, \mathbf{x}_i \,\middle|\, Q_2 < \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq Q_3 \,\right\}, \tag{11.7}$$

where $i \neq j$. We also define the set of remaining points, which lie outside the largest

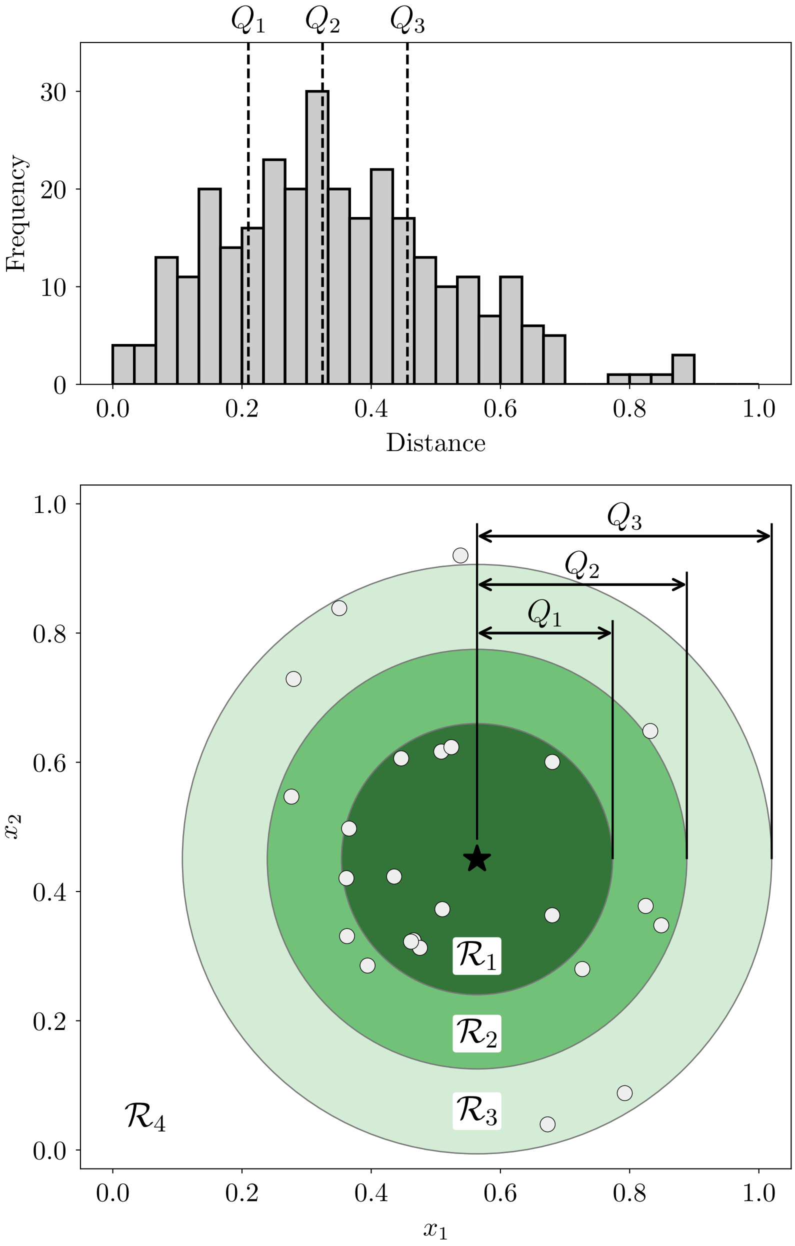


Figure 11.9: Example of pairwise feature distance frequencies and regions surrounding a single test sample (star) from a normal data distribution. The distance distribution (top) defines quartile regions in the feature domain (bottom).

hypersphere, as

$$\mathcal{R}_4^j \coloneqq \left\{ \mathbf{x}_i \,\middle|\, \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) > Q_3 \right\}. \tag{11.8}$$

Figure 11.9 exemplifies the computation of these regions from a normal data distribution in a two-dimensional space using the Euclidean distance. In practice, we use the cosine distance, i.e.,

$$\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) = 1 - s_{\cos}(\mathbf{x}_i, \mathbf{x}_j). \tag{11.9}$$

Once all regions are determined for a reference track, $\mathbf{x}_j$, we can compare how well models trained and validated on sets of randomly selected points from each region perform on the test set $\mathcal{U} = \{\mathbf{x}_j\}$. We repeat this process for all points in the dataset that contain at least $M = 4$ examples in each of its regions.[2] This way, we end up training four different models per reference sample. Once again, we repeat the training process 30 times, with different seeds, keeping the same training sets.

The results of this experiment are presented in Figure 11.10. We show the average F-measure gain $\Delta F(\mathcal{R}_i)$ across all models when using points from each region ($\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$) over using points from the farthest region ($\mathcal{R}_4$). Mathematically, if $F(\mathcal{R}_i^j)$ is the F-measure of a model trained on points from the region $\mathcal{R}_i^j$ around sample $\mathbf{x}_j$, $j \in \{1, \ldots, L\}$, then

$$\Delta F(\mathcal{R}_i) = \frac{1}{L} \sum_j \Delta F(\mathcal{R}_i^j), \tag{11.10}$$

where $\Delta F(\mathcal{R}_i^j) = F(\mathcal{R}_i^j) - F(\mathcal{R}_4^j)$, for $i \in \{1, 2, 3\}$. For all datasets, we observe that the best models are trained on points from the closest regions ($\mathcal{R}_1$), independently of the feature. We may also compare the gain from using $\mathcal{R}_1$ over $\mathcal{R}_2$, for example. Using a set of immediate neighbors leads to significant gains in BRID, as shown by the substantial difference between results in the two regions. In Ballroom this gain is much smaller, and in Candombe it is almost negligible. It is worth noting that for Candombe, the absolute beat F-measure values are 91.7% and 96.4% for STM and OPH, respectively, in $\mathcal{R}_2$. These compare to 62.6% and 65.3%, respectively, for the same region in BRID. This means that there is more room for improvement in BRID than in Candombe. Without forgetting that the definition of these regions depends on the data distribution of each dataset, we can say that data selection must be more important in the former than in the latter.

We have shown with this experiment that one can train beat tracking models that are better able to generalize to recordings in local neighborhoods defined in the

[2] If a data point does not meet this criterion, it is disregarded in this analysis.

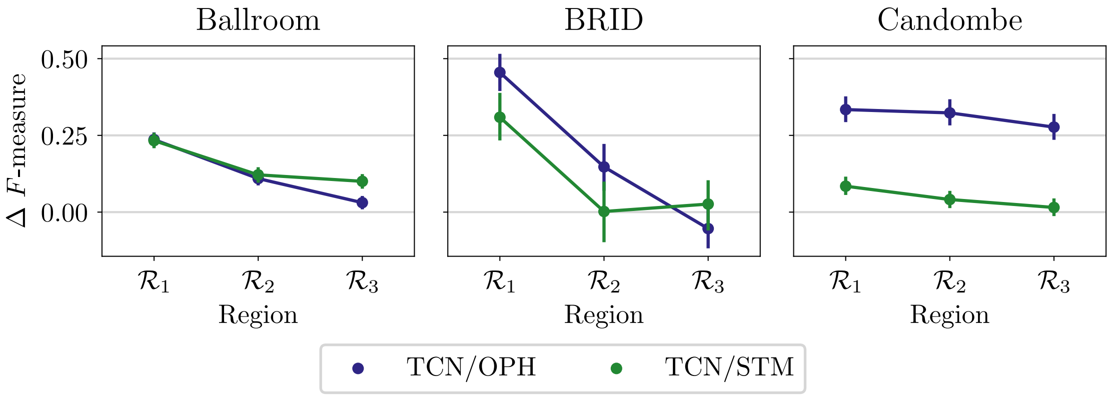


Figure 11.10: Average beat F-measure gains (95% confidence interval) w.r.t. sampling from $\mathcal{R}_4$.

space of the rhythmic features. This is a promising result that suggests that these features can be employed to retrieve informative samples.

11.3.8 Experiment 3: Sampling Strategies

In this experiment, we examine all setups of rhythm description and sample selection techniques across different annotation budgets; and evaluate the beat tracking performance of a TCN model trained on the selected data and tested on the remainder of each dataset. This experiment represents the use case of our proposal. As we mentioned before, the main difference to the real-world scenario is that we use ground-truth annotations instead of asking for a human to provide labels for $\mathcal{L}$. We wish to investigate how much a sampling strategy can improve tracking performance against random sampling. Naturally, this depends on the properties (e.g., tempo and rhythm pattern distributions) of each dataset as well as on the specified annotation budget.

For BRID and Candombe, we vary the annotation budget from 4 to 14 samples (in steps of 2), i.e., about 2–7 min of annotations. Since Ballroom has many more tracks (about 7.5 and 2.5 times more than BRID and Candombe, respectively) and genres, we use larger budgets for this dataset, $M = \{10, 16, 22, 28, 34, 40\}$ ($\sim$5–20 min), i.e., around 2.5 times more data. As in all experiments, training files are split in half for training and validation purposes. Regarding the selective sampling techniques, we observe that MFL and VTK are deterministic and as such always provide the same labeling sets. DIV and MED depend on a random initialization. However, we noticed that a considerable number of files (usually $M - 2$ or $M - 1$) were repeatedly selected over multiple executions of the DIV sampling process, which means that, especially for larger budgets, it is nearly deterministic. In the case of MED, smart initialization of cluster centers with a k-means++ algorithm [337] and multiple runs were used to obtain a more robust clustering. Once these sampling techniques have provided consistent results in terms of the selection, we use them to select one set of training examples for each setup, which is composed of dataset, labeling budget, and feature representation. Finally, for the random baselines (RND), 10 random selections of M files were carried through for each dataset–budget pair; these are used to indicate the expected performance of the TCN model. For each selected data, we repeat the training process 30 times with different seeds.

Beat tracking performance (means and standard deviations) is summarized in Table 11.3. Looking at these results, we notice that, in most cases, selective sampling techniques are consistently better than random sampling across all budgets, although the best feature–selection pair greatly depends on the dataset and training size. In particular, STM+MFL stands out as the best setup for Ballroom, closely followed by STM+MED and OPH+MED, showing gains of up to 5.2 percentage points ($M = 16$) over the random baseline. In extremely low-data scenarios, OPH+MED produces the best results for BRID, with an 18.3 points increase over random at the smallest budget, although STM+MED and OPH+MFL provide good results as well. In both Ballroom and BRID, diversity sampling gives worse results than random for STM (-4 percent points on average). DIV is also worse than RND with OPH in Ballroom (almost -5 points on average), and inconsistent in BRID when paired with the same feature representation. Finally, the RND baseline performance in Candombe is already very high (94.0–96.8%), which leaves little room for improvement in this case. However, except for the two smallest budgets, OPH+DIV provides moderate gains for this dataset (2.3 points average).

The general conclusion is that using sampling techniques can provide better training examples for the TCN model, since most feature–selection setups are shown to outperform the random baseline. This performance gain is typically larger the smaller the annotation budget. We also note that a smart data selection can reduce the standard deviation of the results, leading to more stable solutions than those obtained through random selection. Unsurprisingly, since there are many possible dataset configurations (e.g., highly homogeneous, highly heterogeneous), there is no optimal setup. We discuss a set of recommendations based on our results in the following.

11.3.9 Discussion

In this section, we have studied the influence of data selection on the effectiveness of beat tracking systems that are trained using a limited amount of data. We found out that selective sampling techniques, which take into account the data distribution, can significantly improve beat tracking performance compared to a random selection

Table 11.3: Performances of the different selective sampling setups: mean value (standard deviation) in %. In boldface, we highlight the best-performing selective sampling technique given M (budget) and feature, for each dataset; in gray, we highlight the best-performing setup in each dataset–budget pair. Sampling techniques: diversity (DIV), k-medoids (MED), maximum facility location (MFL), vote-k (VTK), random (RND).

		Beat F-measure (%)								
		Onset patterns histogram (OPH)				Scale transform magnitudes (STM)				RND
Dataset	M	DIV	MED	MFL	VTK	DIV	MED	MFL	VTK	
Ballroom	10	69.5(2.3)	**77.2(2.0)**	76.7(2.5)	74.8(3.0)	66.0(2.9)	**77.4(2.0)**	75.5(1.9)	75.2(1.8)	72.5(4.5)
	16	72.3(2.7)	**81.1(1.1)**	78.4(2.3)	77.9(3.6)	76.7(1.9)	80.4(1.1)	**82.1(1.1)**	78.8(0.9)	76.9(3.2)
	22	74.1(1.5)	**82.2(1.1)**	82.0(1.2)	79.7(1.0)	79.8(2.3)	84.1(0.7)	**85.4(0.6)**	81.3(1.4)	81.1(2.8)
	28	79.0(1.5)	**83.8(0.7)**	83.0(1.2)	81.0(0.9)	77.8(2.4)	84.7(0.8)	**85.9(0.5)**	83.2(0.8)	83.5(1.5)
	34	79.8(1.0)	**85.6(0.8)**	84.3(0.9)	83.0(1.2)	78.1(2.0)	85.7(0.8)	**85.8(0.6)**	85.3(0.8)	84.6(1.4)
	40	81.1(1.4)	**85.2(0.9)**	84.9(1.0)	83.5(0.9)	79.3(1.8)	84.8(1.0)	**85.2(1.3)**	**85.2(0.5)**	85.2(1.4)
BRID	4	83.9(4.4)	**91.0(2.2)**	88.7(3.8)	81.8(3.3)	66.7(8.2)	**86.0(2.7)**	76.3(9.5)	75.0(4.1)	72.7(8.4)
	6	75.9(5.3)	**90.9(2.8)**	89.2(4.2)	86.7(1.7)	72.5(4.9)	**88.2(5.7)**	82.9(3.4)	84.2(4.6)	76.3(8.3)
	8	81.4(5.0)	89.9(3.8)	89.6(3.0)	**90.6(2.1)**	87.4(3.1)	82.8(4.3)	89.4(2.4)	**91.2(1.9)**	78.2(8.4)
	10	84.3(4.6)	93.7(1.9)	**94.9(1.4)**	89.1(1.2)	79.6(3.7)	91.3(2.5)	89.2(2.6)	**94.3(1.7)**	82.7(8.7)
	12	90.5(1.7)	93.3(1.7)	**94.0(6.0)**	91.0(1.7)	80.7(4.7)	89.6(4.8)	90.7(2.6)	**94.1(1.5)**	85.5(6.9)
	14	87.9(2.3)	92.7(2.0)	**94.1(1.9)**	91.2(1.4)	80.7(3.3)	91.4(3.2)	91.5(2.2)	**95.8(1.1)**	89.3(4.7)
Candombe	4	81.2(7.4)	**91.6(2.5)**	82.8(3.7)	90.3(2.5)	89.5(2.8)	90.5(4.5)	**94.9(0.8)**	93.7(1.1)	94.0(3.7)
	6	83.7(13.7)	**95.2(2.6)**	91.7(1.7)	93.2(1.8)	90.3(2.4)	**96.4(0.6)**	95.1(0.7)	95.7(1.0)	95.0(1.8)
	8	**97.0(1.3)**	96.1(1.7)	92.5(1.9)	92.5(1.0)	94.6(2.6)	**96.0(0.7)**	95.2(0.8)	**96.0(0.7)**	95.2(1.5)
	10	**98.2(1.2)**	96.5(1.2)	94.4(1.5)	93.0(0.7)	**96.8(0.7)**	96.2(0.6)	96.3(0.5)	96.0(0.8)	95.9(1.7)
	12	**99.0(0.3)**	95.4(2.6)	96.8(1.0)	93.8(0.9)	**98.2(0.7)**	96.1(0.6)	96.3(0.6)	96.1(0.6)	96.5(1.5)
	14	**99.2(0.2)**	98.8(0.1)	97.1(1.0)	93.8(0.5)	**98.4(0.4)**	96.1(0.6)	96.2(0.5)	96.1(0.5)	96.8(1.5)

baseline, while also reducing its variance. This improvement was observed even when working with small training sets. The baseline results are consistent with those of the previous section, which used the same datasets but did not employ any specific selection scheme.

We have noticed that, in general, when the size of the training set is smaller, performance improvements tend to be more significant. This is mainly due to two reasons. Firstly, it becomes more challenging to improve performance when the results are already very good, which is usually the case with larger annotation budgets. Secondly, with less data, each selected sample contributes proportionally more to what the model sees during training, thus having a greater impact on beat estimation, e.g., changing a single sample in a set of four is more critical than changing it in a set of 20 samples.

Regarding feature representations, it is currently unclear which is preferable, as both OPH and STM allowed for good training samples to be selected. Initially, we had a suspicion that OPHs, which encode tempo information, would produce better results than STM in general, given that the TCN model is sensitive to the tempo distribution of the dataset [275]. However, it should be noted that, at the post-processing DBN stage, tempo is dissociated from the rhythmic pattern and separately encoded in the state variable. Additionally, the difference in dimensionality between the two features cannot be ignored. Further investigation is needed to determine which representation is more effective.

This study has also examined the results of sample selection on different datasets. Although our work primarily focuses on single-genre datasets from Afro-rooted traditions, we highlight the moderate performance gains observed in Ballroom, which is highly diverse with various genres, meters, and patterns (see Section 11.3.4). We then turn our attention to the two main datasets, Candombe and BRID. Candombe, which displays the smallest tempo range and little pattern variability, benefited less from tailored sets of training examples. However, we note that models trained on random selections already accurately track Candombe excerpts, which means there is less room for improvement. On the other hand, BRID, which is less homogeneous, seemed to profit the most from sample selection. It is yet to be determined how exactly these two characteristics — tempo and rhythm — affect the impact of selection in each dataset, and whether general rules could be established to inform when selective sampling is most beneficial. We underscore that OPH+DIV (which maximizes diversity) and STM+MFL (which maximizes representativeness) were the best-performing setups for Candombe (most homogeneous) and Ballroom (most heterogeneous), respectively. For BRID, on the other hand, MED and VTK — both making a compromise between diversity and representativeness — were the better sampling schemes.

Finally, it is worth comparing our results with the state-of-the-art procedures discussed in Section 11.3.5. First, we observe that our results consolidate the idea of adapting to challenging music already expressed by [17, 290]. We saw in Experiment 1, under a restrictive scenario (2 min of training data), that a large majority of the adapted models has greatly surpassed the “pre-trained” model (Table 11.2), which was trained on hours of data from standard datasets. This improvement was more evident when data selection was carefully planned (Table 11.3). Our selective sampling strategy has proven to be much more effective when compared to the random baseline, as we have managed to achieve results that are very close to the “trained from scratch” model (which we consider a “full-dataset” performance). For example, in BRID, with the same 2 min of data, there is only a 7.9 percent difference, while RND is behind by 26.2 points. It is also worth mentioning that our selective sampling approach is comparable to transfer-learning-based procedures. With a budget of $M = 6$ samples (3 min of annotations), we have obtained results that are on par with the “fine-tuned” models for Candombe and BRID, with only a slight difference of -1.8 and -2.5 percent points, respectively, considering the best feature–selection pairs, but with lower variability. We must note that our approach is not only affordable but should be considered more general, as pre-trained networks may have been trained on data that is not relevant to the object of study, which could compromise the stability of the results.

Chapter 12

Contributions to Microtiming Analysis

We have discussed in Chapter 9 that many music styles related to African or Afro-diasporic practices contain an element of rhythmic expression found in a very fine timescale. In particular, for Afro-Brazilian dance music, this fine structure is composed by deviations from the nominal grid at a subtactus level, i.e., the sixteenth-note level. By definition, prior knowledge of the underlying isochronous subdivisions is needed to capture the actual non-isochronous duration patterns. Some works that were reviewed in Section 9.2 quantize the positions in-between successive beats or downbeats with a fixed number of points to generate this reference [300]. Others measure the distances (normalized by the IBI) between each individual pulse and the preceding beat or downbeat [38, 201, 203, 301]. Others still investigate microtiming as patterns of duration, measuring the length of each individual note relative to the IBI [297].

In this chapter, we provide a few contributions to the analysis of microtiming patterns in *samba*. First, we conduct a reference investigation, using annotations, to verify prior musicological studies [296, 297]. Next, we present a novel model, inspired by the bar-pointer model, that allows simultaneously tracking beats and microtiming. This is an appropriate approach given the natural relation between their corresponding timescales (tactus and tatum). This latter study was originally presented in [21], but it is expanded here: besides the exact inference approach, we also present an approximate inference approach that exploit a sequential Monte Carlo method (particle filter) to reduce the computational cost and the constraints upon the inference. We take a non-traditional approach to alleviate certain limitations from the original proposal.

12.1 Characterization of Microtiming in *Tamborim Carreteiro*

In this section, we perform a preliminary investigation of microtiming patterns in selected tracks from BRID. To simplify our analysis, we consider a set of *tamborim* solos playing the typical *carreteiro* pattern (see Figure 2.19a) of *samba-enredo*. Our subset is composed of the seven tracks found in region 2 of Figure 8.7, namely: `[0131]`, `[0132]`, and `[0136]`, by musician 1 (`S1`); `[0216]` and `[0218]`, by musician 2 (`S2`); `[0304]` and `[0306]`, by musician 3 (`S3`). With all these renditions, we have a large number of examples of the same pattern, from the same instrument, where all tatums are articulated (i.e., there are always four fast pulses to every beat), and from which we can extract the microtiming information. In total, these tracks add up to 396 beats and 1584 onsets.

As in [14, 38, 201, 203, 301], we measure the position of each tatum with respect to the preceding beat. If we call Δ_i the distance between tatum i and its previous beat, we can define our normalized microtiming feature as the normalized tatum position

$$m_i = \frac{\Delta_i}{\text{IBI}}, \tag{12.1}$$

for $i = \{1, 2, 3, 4\}$, where IBI is the local inter-beat interval (the distance between the preceding and succeeding beats). Since, in our case, the position preceding beat is equivalent to that of the first tatum (see Section 3.1.3), it is evident that $\Delta_1 = 0$. Note however, that this is not a necessity, and could be different if the beat was estimated by some other method (e.g., the mean position of the first onset for multiple instruments, as in the case of [201, 301]). Figure 12.1 presents an example of this definition in an excerpt of track `[0306]`.

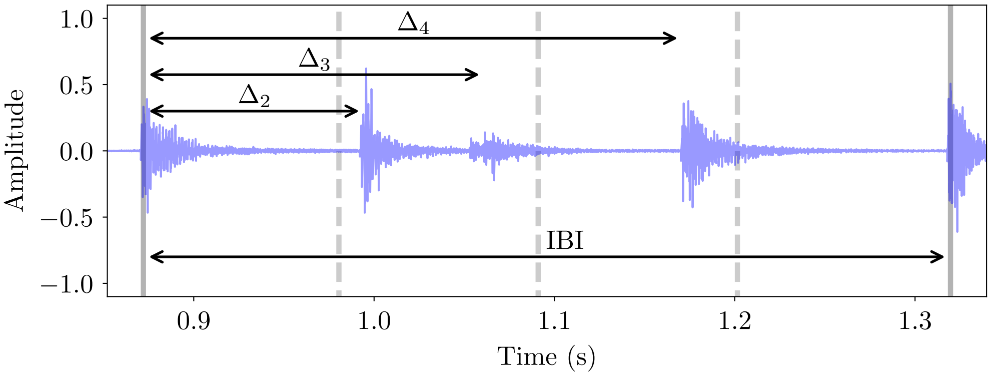


Figure 12.1: Example of microtiming deviations at the sixteenth-note level. We show beat annotations (solid, vertical) and the underlying isochronous grid (dashed). Deviations (Δ) for tatums 2, 3, and 4 are indicated, as well as the inter-beat interval.

We also point out that the transformation into the system used by GERISHER [297] (patterns of duration) is easily achievable by doing

$$\begin{cases} \delta_1 = \Delta_2 \\ \delta_2 = \Delta_3 - \Delta_2 \\ \delta_3 = \Delta_4 - \Delta_3 \\ \delta_4 = \mathrm{IBI} - \Delta_4 \end{cases} \tag{12.2}$$

and normalizing all durations by the IBI to obtain $\tilde{\delta}_i = \frac{\delta_i}{\mathrm{IBI}}$. For transforming from duration patterns back to relative durations, it suffices to compute

$$m_i = \sum_{j=1}^{i-1} \tilde{\delta}_j \tag{12.3}$$

for each m_i, $i \in [2, 4]$.

We can now explore the microtiming profiles of the *carreteiro* tracks. Figure 12.2 shows every beat-length pattern in the subset. We have stacked the patterns of the different tracks, separately for each performer. The average microtiming feature $\bar{m}_i$ is informed as a percentage of the IBI. We can readily verify that, in all cases, while there is not a major deviation in tatum 2 (it is played on average slightly behind time), tatums 3 and 4 lie mostly ahead of time by a significant margin. We note that tatum 4 appears almost precisely in a subdivision in three of the beat (triplet feel). Expressed in vector form ($[\bar{m}_1, \bar{m}_2, \bar{m}_3, \bar{m}_4]$), the average microtiming profile in the subset (across all musicians) is $\bar{\mathbf{m}} = [0.000, 0.265, 0.433, 0.671]$, with standard deviations of $[0.000, 0.011, 0.019, 0.012]$.

Our results confirm findings from [201, 297, 299–301] in different types of *samba*. In particular, in [299], at a tempo of 133 bpm, the average duration pattern was $\tilde{\boldsymbol{\delta}} = [0.27, 0.15, 0.25, 0.34]$, measured in proportion to the IBI, which is equivalent to $\bar{\mathbf{m}} = [0.00, 0.27, 0.42, 0.67]$. This is strikingly close to our results in Figure 12.2, which were obtained at a tempo of (130.0 ± 2.4) bpm.

12.2 Beat and Microtiming Tracking

In this section, we introduce our proposal of a fully-automatic system for simultaneously tracking beats and microtiming. Our model is built upon a CRF that uses beat and onset activations as observations, and combines them for investigating rhythmic expression at the sixteenth-note level. Since all variables in the model are discrete, we can perform exact inference using the Viterbi algorithm. This was originally reported in [21]. This is only feasible, with regard to the computational cost, due

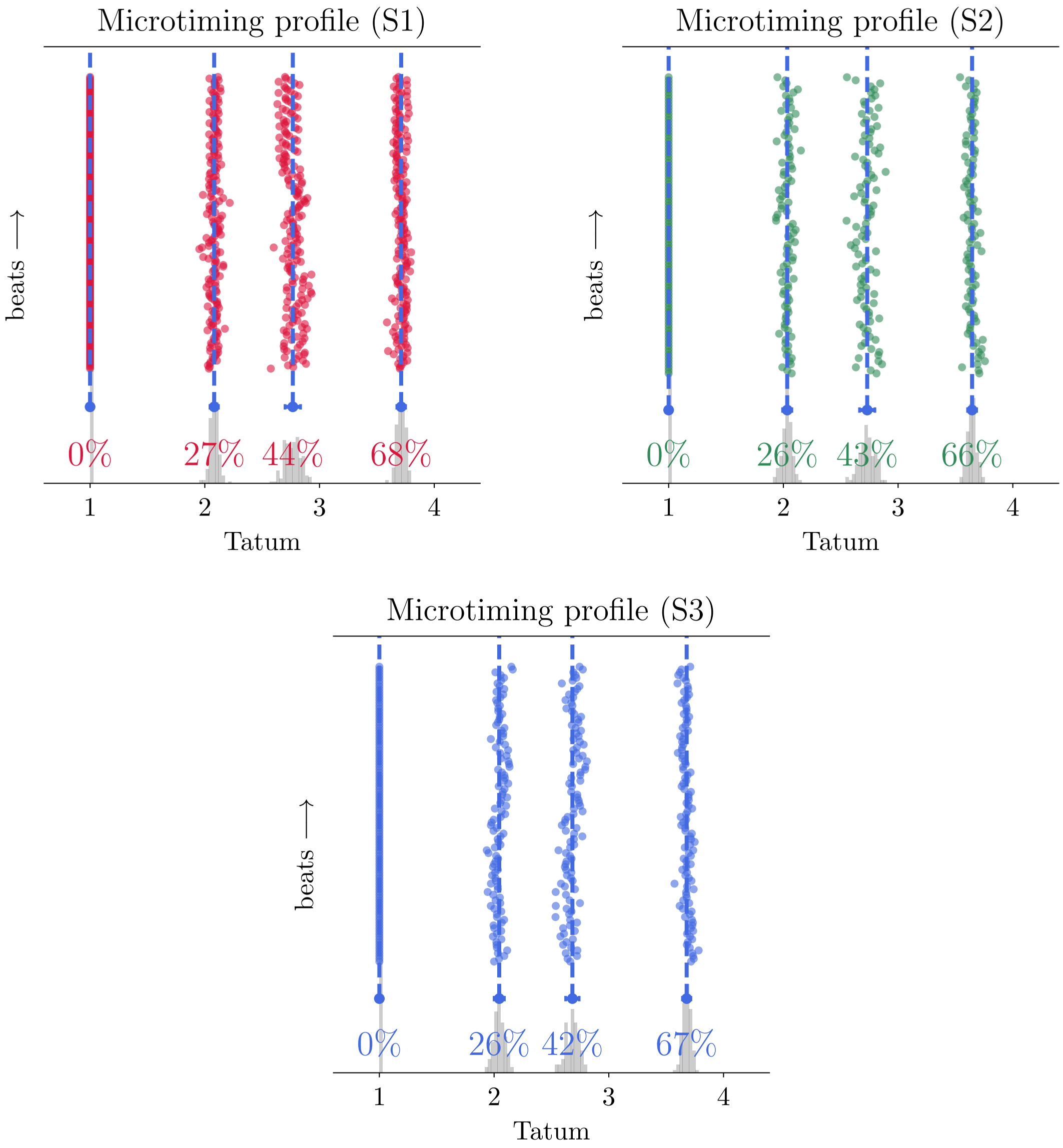


Figure 12.2: Microtiming profiles by musician.

to a coarse discretization of the microtiming variable. We also describe an approach for inference using particle filters (PF), which does not require discretization, but yields only approximate solutions to the tracking problem. Finally, we showcase our system and the two inference schemes with an experiment over the aforementioned subset of seven *tamborim carreteiro* tracks.

12.2.1 Model Structure

Our model consists of a linear-chain CRF [338, 339]. We write the conditional probability of a label sequence $\mathbf{x}_{1:K} = [\mathbf{x}_1, ..., \mathbf{x}_K]$ given an input sequence of observations $\mathbf{y}_{1:K} = [\mathbf{y}_1, ..., \mathbf{y}_K]$ of length K frames as

$$P(\mathbf{x}_{1:K}|\mathbf{y}_{1:K}) = \frac{1}{Z(\mathbf{y}_{1:K})} \prod_{k=1}^{K} \psi(\mathbf{x}_k, \mathbf{x}_{k-1})\, \phi(\mathbf{x}_k, \mathbf{y}_k), \tag{12.4}$$

where $\psi(\mathbf{x}_k, \mathbf{x}_{k-1})$ and $\phi(\mathbf{x}_k, \mathbf{y}_k)$ are the transition and observation potentials, respectively. These potentials work similarly to transition and observation probabilities in DBNs and HMMs, but they are not required to be proper probabilities, hence the need for a normalization factor, $Z(\mathbf{y}_{1:K})$.

Output Variables

The output labels $\mathbf{x}_{1:K}$ are composed of three variables,

$$\mathbf{x}_k := [f_k, l_k, \mathbf{m}_k], \tag{12.5}$$

where $f_k \in \{1, \ldots, l_k\}$ is a frame counter that describes the position inside the beat; $l_k \in l_{\min}, \ldots, l_{\max}$ is the length of the beat interval in frames (which is related to tempo); and $\mathbf{m}_k \in \{\mathbf{m}_1, ..., \mathbf{m}_M\}$ is the microtiming variable. The observations $\mathbf{y}_{1:K}$ are based on estimated beat and onset likelihoods, as detailed later. The problem of obtaining the beat positions and microtiming profiles can be formulated as finding the sequence of labels

$$\mathbf{x}^*_{1:K} = \underset{\mathbf{x}_1, ..., \mathbf{x}_K}{\arg\max}\, P(\mathbf{x}_{1:K}|\mathbf{y}_{1:K}). \tag{12.6}$$

As mentioned in the beginning of this chapter, we want to investigate microtiming in beat-length rhythmic patterns where all four sixteenth notes are articulated. Therefore, we define the microtiming variable $\mathbf{m}_k$ at frame k similarly to what is shown in Figure 12.1 as

$$\mathbf{m}_k := [m^k_2, m^k_3, m^k_4], \tag{12.7}$$

where $m^k_i = \frac{\Delta^k_i}{l_k} \in [\frac{i-1}{4} + L_i, \frac{i-1}{4} + U_i]$, and Δ^k_i is the distance in frames between an articulated sixteenth note and the beginning of the beat interval. Each m^k_i

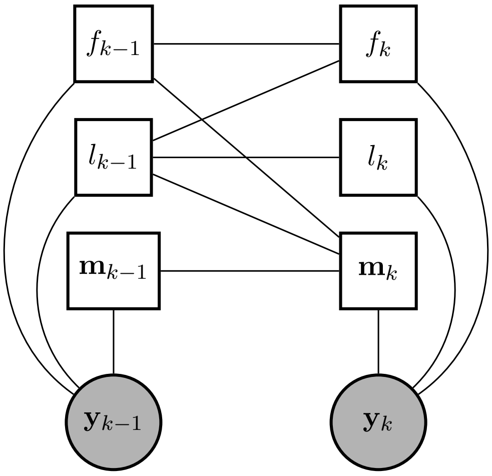


Figure 12.3: CRF graph of the model. Gray nodes are the observed variables, and white nodes are the labels.

models the position of the i-th sixteenth note (disregarding the beat itself, tatum $i = 1$) with respect to the beginning of the beat, and relative to the total beat length. For instance, if all tatums are located exactly on their nominal positions, we have $\mathbf{m} = [0.25, 0.50, 0.75]$. To account for different microtiming profiles, the value of m_i^k is estimated within an interval determined by lower and upper deviations bounds, L_i and U_i, measured in fractions of the beat-length interval. We observe that, with minor adjustments, our framework can be adapted to track other kinds of microtiming deviations and rhythmic patterns with a different number of tatums.

Our model is represented in Figure 12.3.

A Priori Knowledge

We incorporate to the system some a priori knowledge in the form of the following assumptions, which are further explained later:

1. The tempo is constant within a beat and rarely changes;
2. The microtiming profile changes smoothly and only on beat transitions;
3. The tempo is in the range 125 to 135 bpm, to ensure an appropriate temporal resolution.
4. The microtiming is bounded by deviations $\mathbf{L} = [-0.005, -0.105, -0.110]$ and $\mathbf{U} = [0.035, -0.025, -0.050]$.

Transition Potential

The transition potential is given in terms of f_k, l_k, and $\mathbf{m}_k$ by

$$\psi(\mathbf{x}_k, \mathbf{x}_{k-1}) = \psi_{f,l}(f_k, f_{k-1}, l_k, l_{k-1})\, \psi_m(f_{k-1}, l_{k-1}, \mathbf{m}_k, \mathbf{m}_{k-1}). \tag{12.8}$$

As in [260, 270], the frame counter f_k increases by one at each step, up to the maximum beat length considered, going back to one at the end of the beat. Our assumption (1) states that changes in the beat interval duration are unlikely (i.e., tempo changes are rare) and only allowed at the end of the beat. We constrain these changes to be smooth, giving inertia to tempo transitions. These rules can be expressed by

$$\psi_{f,l}(f_k, f_{k-1}, l_k, l_{k-1}) = \begin{cases} 1, & \text{if } f_k = (f_{k-1} \mod l_{k-1}){+}1, \\ & \quad f_{k-1} \neq l_{k-1}; \\ 1 - p_f, & \text{if } l_k = l_{k-1}, \\ & \quad f_k = 1, \\ & \quad f_{k-1} = l_{k-1}; \\ \frac{p_f}{2}, & \text{if } l_k = l_{k-1} \pm 1, \\ & \quad f_k = 1; \\ 0, & \text{otherwise.} \end{cases} \tag{12.9}$$

The microtiming descriptor m_k changes smoothly (with resolution r) and only at the end of the beat, that is,

$$\psi_{m}(f_{k-1}, l_{k-1}, \mathbf{m}_k, \mathbf{m}_{k-1}) = \begin{cases} 1, & \text{if } \mathbf{m}_k = \mathbf{m}_{k-1}, \\ & \quad f_{k-1} \neq l_{k-1}; \\ 1 - p_m, & \text{if } \mathbf{m}_k = \mathbf{m}_{k-1}, \\ & \quad f_{k-1} = l_{k-1}; \\ \frac{p_m}{2}, & \text{if } m_i^k = m_i^{k-1} \pm r, \forall i, \\ & \quad f_{k-1} = l_{k-1}; \\ 0, & \text{otherwise.} \end{cases} \tag{12.10}$$

In the transition potential, p_f and p_m represent the probability of changing the length of the beat interval and the probability of changing the microtiming profile, respectively. Following previous works, we have set $p_f = 10^{-3}$ to indicate the unlikeliness of tempo transitions. We have experimented with a few values for p_m.

Since m_i^k is given as a fraction of the IBI, the resolution of our microtiming scale is also related to that variable. In fact, it is given by the relation between the feature rate f_{r} and the tempo τ (in bpm):

$$r = \frac{\tau}{60 \times f_{\mathrm{r}}}. \tag{12.11}$$

Considering our literature review, we assume that a resolution of 2% of the IBI is sufficient for representing microtiming deviations [201, 300]. To keep the computational complexity low but at the same time guarantee a resolution $r = 0.02$, we

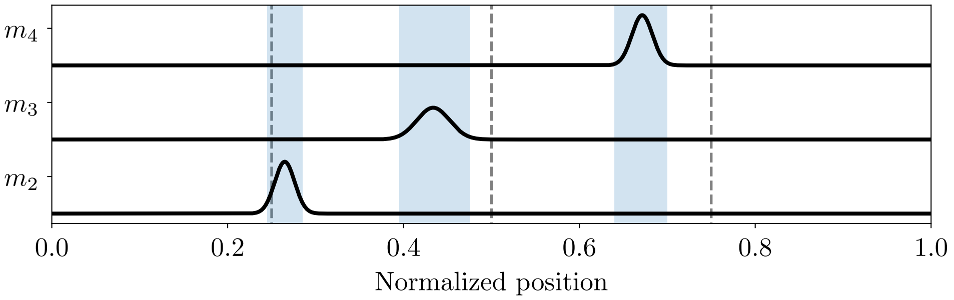


Figure 12.4: Microtiming distributions in the subset and search regions (blue) for the model.

sample the observation features at a rate of 110 Hz and limit tempo to the range of 125 to 135 bpm, which is approximately around two standard deviations of our subset tempo distribution mean. The bounding regions of the microtiming variable (assumption (4)) are such that we can encompass the subset within about two standard deviations of the mean (see Figure 12.4). These choices are valid for the music under study, and could be adapted to cope with different music genres.

Observation Potential

We exploit as onset likelihood the spectral flux from `librosa` [305], which is based on a filtered time-difference of the log-power mel spectrogram. For beat, we obtain the activation using the previously discussed TCN architecture. For this purpose, we train the model from scratch with the *tamborim carreteiro* subset: we use each track as test file once, and extract a beat activation for it from a network trained on the remaining six tracks (each divided in half for training and validation purposes). Training parameters can be seen in Chapter 11. The main difference here is that we compute these features at a higher rate of 110 bpm.

The observation potential depends on the beat and onset likelihoods (b_k and o_k, respectively), the frame counter f_k, the local beat length l_k, and the microtiming variable $\mathbf{m}_k$. We write this potential as

$$\phi(f_k, l_k, \mathbf{m}_k, \mathbf{y}_k) = \begin{cases} b_k, & \text{if } f_k = 1; \\ o_k - b_k, & \text{if } \frac{f_k}{l_k} \in \mathbf{m}_k; \\ 1 - o_k, & \text{otherwise.} \end{cases} \tag{12.12}$$

12.2.2 Inference Methods

Exact Inference

We perform the inference in the model presented in Section 12.2.1 using a generalized Viterbi decoding to obtain the MAP sequence of labels $\mathbf{x}^*_{1:K}$. However, for the model to be feasible in terms of its computational complexity, we had to coarsely discretize the microtiming space and limit the tempo range. Even under these strict conditions, the total size of the state space is $S = 15\,000$ — there are 250 frame-counter–tempo states, three states for m_2, five for m_3 and four for m_4. We note that S could rapidly increase if we were to add more tatums to track (more m_i), improve the microtiming resolution, or expand the tempo range.

Particle Filters

We exploit the Markov-like form of our CRF model and employ a simple approximate inference scheme based on particle filters to avoid the computational overhead of exact inference. Even though potentials are not proper conditional probabilities, we still use them for sampling next candidates for labels, normalizing weights and parameters accordingly.

In the particle filter, the posterior probability is approximated by a sum of N weighted particles in the state space. Following [196], we rewrite Equation (12.4) as

$$P(\mathbf{x}_{1:K}|\mathbf{y}_{1:K}) \approx \sum_{j=1}^{N} w_K^{(j)}\, \delta(\mathbf{x}_{1:K} - \mathbf{x}_{1:K}^{(j)}), \tag{12.13}$$

where each j-th particle $\mathbf{x}_{1:K}^{(j)}$, for $j = \{1, \ldots, N\}$, is associated with a weight $w_K^{(j)}$ at frame K, and $\delta(\mathbf{x})$ is the Dirac delta defined in $\mathbb{R}^D$,

$$\delta(\mathbf{x}) = \delta(x_1)\, \delta(x_2) \cdots \delta(x_D) \tag{12.14}$$

Through a sequential importance sampling (SIS) [340] scheme, at each time step, we update the labels of each particle according to the transition potential $\psi(\mathbf{x}_k, \mathbf{x}_{k-1})$, which is here redesigned in continuous form. We sample a new tempo for frame k from a Gaussian distribution centered at the previous tempo (at $k-1$) with standard deviation given by parameter σ_l. Note that we use tempo as a variable, instead of the beat length, to simplify the sampling process and make updates of $\pm\Delta\tau$ bpm equally likely. Either way, if this standard deviation σ_l is small, i.e., the corresponding distribution is thin, we can sustain assumption (1) on the inertia of tempo changes. Microtiming is sampled in a similar fashion, but with a distribution

$\mathcal{N}(m_i^{(j)}, \sigma_{m_i}^2)$, for each dimension m_i, where

$$\sigma_{m_i} = (U_i - L_i)\sigma_{\mathbf{m}}, \tag{12.15}$$

and $\sigma_{\mathbf{m}}$ is another adjustable parameter. We define the standard deviation of each microtiming dimension as proportional to the searchable region defined by the upper and lower bounds (L_i, U_i), so that we can have more detail in smaller regions, and less detail in larger regions. We have empirically chosen the values $\sigma_l = 0.02$ and $\sigma_{\mathbf{m}} = 0.05$.

We iteratively update the weights $w_k^{(j)}$ using the observations, b_k and o_k, as expressed by the observation potential ϕ:

$$w_k^{(j)} \propto w_k^{(j)}\, \phi(\mathbf{x}_k^{(j)}, \mathbf{y}_k), \tag{12.16}$$

where weights are normalized such that

$$\sum_{j=1}^{N} w_k^{(j)} = 1, \quad \forall k. \tag{12.17}$$

After all particle trajectories $\{\mathbf{x}_{1:K}^{(j)}\}$ are determined, we select as MAP the trajectory $\mathbf{x}_{1:K}^{(j)}$ with the highest weight $w_K^{(j)}$ [196].

PFs are subject to the degeneracy problem [340], where the variance in the particle importance increases with time and the weight of most particles approaches zero. In the ideal case, we would have a perfect approximation to the posterior probability and low-variance weights. We follow [196] and approach this problem with two simple resampling schemes: systematic resampling (SISR) [341] and auxiliary particle filter (APF) [342]. To put it briefly, the main idea behind SISR is to replace particles of low importance by particles with high importance. This is done with a resampling procedure that selects particles in proportion to their weights. We follow the approach in [196] and only perform this resampling when the effective sample size

$$N_{\text{ESS}} = \left(\sum_{i=1}^{N} \left(w_k^{(i)} \right)^2 \right)^{-1} \tag{12.18}$$

is below a threshold of ρN. If the probability mass function defined by the normalized weights is close to uniform, N_{ESS} is high; if instead, the PMF is concentrated on few weights, the effective sample size is small [343]. We have set $\rho = 0.1$ for APF and $\rho = 0.02$ for SISR, following [196]. However, this resampling procedure creates another problem: the impoverishment of particle diversity, i.e., when the resampled particles only represent a limited region of the state space. The APF method attempts to remedy this by compressing the weights of each particle before resampling (and later decompressing them) to even out the chances of a particle being selected. Here, we have used a compression of the form

$$g(w) = w^{\beta}, \tag{12.19}$$

with $\beta = \frac{1}{4}$ (fourth root).

12.2.3 Evaluation Metrics

We evaluate both inference methods (CRF and PF) with respect to accuracy and computational cost.

For measuring beat tracking accuracy, we employ the F-measure with the default tolerance of ±70 ms (see again Section 10.1).

Microtiming is also evaluated with the F-measure, as this metric is also used in onset detection (cf. Chapter 7). To do so, we first decode an estimated microtiming sequence $\mathbf{m}_{1:K}$ with the knowledge of the estimated beats to obtain the estimated onset positions. Then, for each tatum $i \in \{2, 3, 4\}$ and its corresponding set of decoded onsets, we compute the F-measures, $\{F_i\}$. The final microtiming tracking score for a piece is given by the average

$$F_{\mathbf{m}} = \frac{1}{3} \sum_{i=2}^{4} F_i. \tag{12.20}$$

We evaluate this F-measure with a set of tolerance values (from ±5 to ±25 ms) around estimated onsets, instead of using a single one (e.g., ±50 ms), as it is commonly done in onset detection (cf. Chapter 7). This allows us to have a better understanding of the model's accuracy under different conditions.

Lastly, we evaluate the computational cost of the inference step by measuring its runtime per track. We disregard the feature extraction step, which is the same for all models.

We present the average results across the subset that was described in Section 12.1. In the case of PFs, which are stochastic in nature, we run the inference step ten times for each file reporting averages and standard deviations.

12.2.4 Performance of Models

We have tried different values for $p_{\mathbf{m}} \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$, the probability of change in microtiming profile, in the CRF model. For PF approaches, we have varied the number of particles $N \in \{2000, 4000, 6000, 8000, 10\,000\}$. Table 12.1 gives an overview of the performance of each model, with regard to its accuracy in tracking beats and microtiming, and its computational cost. In the same table, we

Table 12.1: Performance figures of the exact and approximate inference models: mean (standard deviation) in %. In gray, we highlight the best F-measures across all configurations for each inference model; in boldface, we highlight the best F-measure overall. Results for the analysis of the *tamborim carreteiro* subset. Microtiming scores were obtained using a tolerance of ±25 ms.

Inference	Model	$p_{\mathbf{m}}$	N	Beat F (%)	$F_{\mathbf{m}}$ (%)	Runtime (min)
Exact	CRF	0	-	**94.3**	**90.2**	103.7(6.0)
		10^{-3}	-	**94.3**	89.7	103.9(5.5)
		10^{-2}	-	**94.3**	89.0	104.1(5.3)
		10^{-1}	-	**94.3**	88.6	104.2(6.7)
Approx.	SISR	-	2000	66.6(21.8)	58.2(24.9)	0.2(0.1)
		-	4000	71.3(21.2)	62.1(24.8)	0.3(0.1)
		-	6000	73.6(19.2)	66.3(23.0)	0.5(0.1)
		-	8000	76.6(19.2)	69.3(22.4)	0.9(0.3)
		-	10 000	75.6(17.4)	68.4(20.8)	0.9(0.2)
Approx.	APF	-	2000	90.3(13.5)	88.0(15.0)	0.2(0.0)
		-	4000	91.8(9.3)	89.7(10.7)	0.4(0.1)
		-	6000	88.8(17.0)	86.6(18.2)	0.6(0.1)
		-	8000	90.8(12.8)	88.6(13.8)	0.8(0.1)
		-	10 000	89.1(18.1)	87.0(19.4)	1.0(0.2)
CRF solo beat estimation F-measure: 96.8%						

also report a reference beat F-measure value, which was obtained with a simplified CRF model. This model just tracks beats with target variables f_k and l_k, using the same potential $\psi_{f,l}$ and an observation potential given by

$$\phi(f_k, l_k, \mathbf{y}_k) = \begin{cases} b_k, & \text{if } f_k = 1; \\ 1 - b_k, & \text{otherwise.} \end{cases} \tag{12.21}$$

First, we evaluate the accuracy of each model in the task of beat tracking. Our reference CRF sets the bar at an F-measure of 96.8%, and is closely followed by the CRF models with microtiming targets (all tied at 94.3%). The APF system provides a high score (with the best average result of 91.8% for $N = 4000$ particles), but the SISR system is unable to properly capture beats at the same level (best average result of 76.6%). These results clearly attest the effect of impoverishment in the quality of the particles. There is also a lot of variability in SISR estimates, when compared with APF in terms of the standard distributions.

We now turn our attention to microtiming estimation; its F-measure scores (computed as expressed in Section 12.2.3) are reported in Table 12.1 for a tolerance window of ±25 ms. We can see from the CRF results that a constant microtiming profile ($p_{\mathbf{m}} = 0$) is sufficient for tracking rhythmic expression in our subset — these

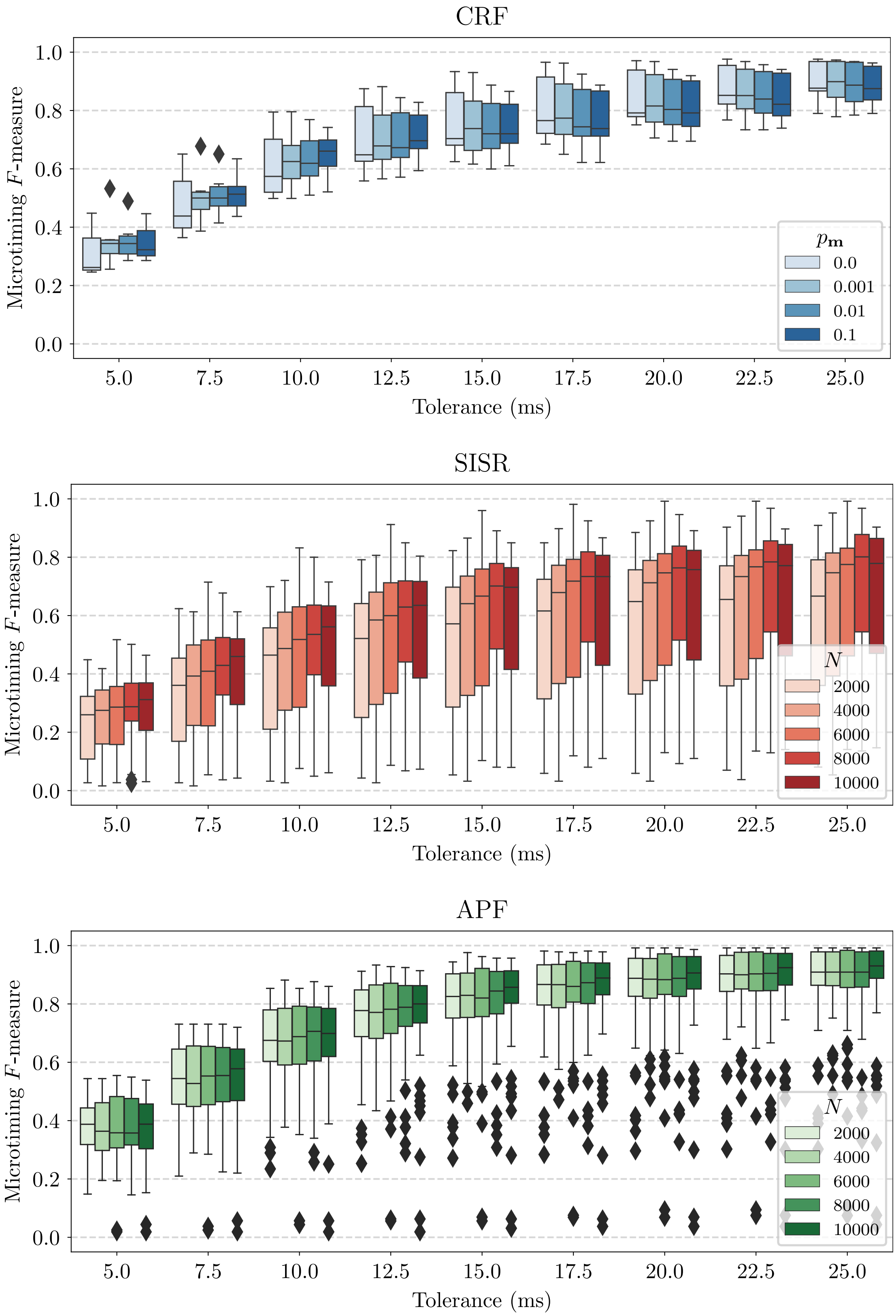


Figure 12.5: F-measures for microtiming tracking with various tolerance windows (±tolerance value around the onset annotation). Diamonds indicate outliers.

are the best results overall. In SISR, the worse beat estimation results in similarly degraded microtiming F-measures. Since we have also restricted the tempo range for particle filter systems, these values of about 73% do not correspond to octave errors; they are phase problems. Most importantly, we observe that APF's results are very comparable with those from the CRF system. Figure 12.5 lets us dive deeper into the matter of the estimates' accuracy by exploring a range of tolerance values, from ±5 ms (1% of the IBI at 130 bpm) to ±25 ms (5% of the IBI at the same tempo). First, we can identify that there is a difference among the CRF models with respect to the probability $p_{\mathbf{m}}$: for narrower tolerance windows, the median performance of the constant microtiming tracker is slightly lower than that of the others. Moreover, we readily verify that, starting at ±22.5 ms, again as the tolerance window gets narrower, median microtiming F-measures for APF surpass those of the CRF. This means that the approximate inference performs better than the exact inference one. This may seem strange at first, but it can be explained by the coarse quantization of the state space for the microtiming target in the CRF implementation, which reduces its ability to consistently and accurately track onset positions. The particle filter approach does not require quantizing the state space in such manner. For APF, we can still achieve reasonable results above 80% in F-measure for tolerance windows of ±12.5 ms and up. Of course, this becomes limited by the fact that our feature rate is 110 Hz. Figure 12.6 presents the microtiming ground truth and estimation with APF for track `[0131]`; we also display a smoothed version of the ground truth that better aligns with our assumption of a slowly changing microtiming profile, achieved by filtering out a portion of the "motor noise" [344] (i.e., variability in time-keeping due to physical constraints).

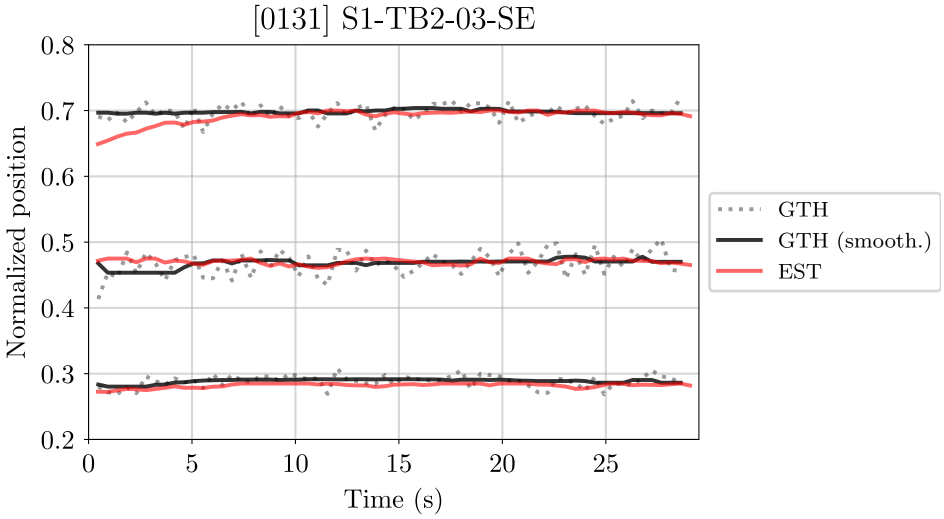


Figure 12.6: Microtiming ground truth and estimation with APF for track `[0131]` (Beat F-measure = 99.2% with tolerance of ±70 ms, $F_{\mathbf{m}}$ = 99.4% with tolerance of ±25 ms). We present the ground truth (GTH) and the estimation (EST), as well as a version of the ground truth smoothed with a median window of length 21 beats.

We end this investigation with a discussion about the computational cost of the different methods. As we can verify in Table 12.1, both particle filter systems are inexpensive, with the best APF solution running the entire inference in under 30 s (files in our subset have an average duration of 28.5 s). We can also perceive that the time cost of particle filter-based approaches is nearly linear with the number of particles. Finally, we notice the great difference between the PF approaches and the CRF, whose computational costs are two orders of magnitude greater. This, together with the competitive performance in both beat and microtiming tracking tasks, makes the APF approach a very good alternative to exact inference methods.

Chapter 13

Conclusions

This thesis has presented a series of investigations and contributions in the field of automatic music transcription, specifically concentrating on the problems of drum sound classification and rhythmic description. These techniques were showcased with two datasets of *samba* music, which were also organized within the context of this work. This concluding chapter reflects on our achievements in this work and discusses prospects for future contributions.

13.1 Summary and Conclusions

We structure the summary of our work and our main conclusions in the light of the main objectives of this thesis and the outputs of each model and experiment.

Data Curation and Annotation

One of the main contributions of this thesis was the curation and annotation of two moderately sized datasets of *samba* music, BRID and SAMBASET. These two datasets present complementary perspectives, from stripped-down solo performances to commercial-quality recordings of live *samba-enredo* performances — the closest one can get to the *Avenida* in a controlled environment. These two sides of the dataset have allowed the investigation of nuances in the performances with low-level features as well as larger-scale descriptions of the musical phenomena. The collected metadata can provide insights when paired with computational analyses, and the production of accurate annotations for onsets, beats, and downbeats is invaluable. We have also displayed a few ideas to improve this annotation pipeline, with a semi-automatic procedure for beat and downbeat annotation that leverages prior musicological information about meter in *samba*, among other things.

Classification of Note Articulations

In one of the parts of this research, we have collaborated with a more expressive view of percussive performances concerning timbre. Usually, in automatic drum transcription, the quality of the sound events produced by a single instrument is regarded as mostly immutable, or at least, the more complex rudiments are discarded in the search for a simplified detection and recognition model. However, we can find many examples in African and Asian cultures, for example, of the drummer manifesting great expression through different modes of excitation. These articulations are also important in Afro-Brazilian music practices and are well-correlated with the sensation of groove elicited by the performance.

We have investigated the problem of drum sound classification using a subset of BRID containing *tantã* and *repique* solo recordings. More than a simple binary instrument classification, we set out to determine the features that were useful for distinguishing each instrumental playing technique, which is very nuanced. We have presented the entire pipeline, from onset detection and note segmentation to the feature extraction and the classification itself, providing some insights. For instance, we have showed that, after a careful grid search of the peak picking parameters, our modified RCD is among the top performing ODFs both in terms of the F-measure and of the MAE, i.e., it produces the closest estimates to the annotated beat positions. Due to the percussive nature of our signals, other ODFs have also been deemed suitable (e.g., HFC, E, SF). However, we have shown that onsets estimated with RCD were closer to our annotations and that RCD had better recall overall. This has justified a slight increase in the necessary computation power.

We have extracted features from four domains — temporal, spectral, cepstral, and modulation — and used them in two types of classification experiments. The first type approached the classification of specific articulations of each instrument. We have shown that our proposed CQT-cascade modulation spectrum yielded the best performance, along with the regular temporal features. We have also performed a classification of archetypal strokes, identifying the commonalities in function between the articulations of *tantã* and *repique*. Again, these two sets of features displayed the best results, which improved when they were aggregated. A note of caution is due in the interpretation of these results and the discriminative power of each feature set since these temporal and modulation-based features have very different dimensions. Either way, both were equally robust in face of the challenges of our subset, in particular, the overlap between notes.

Enabling Improved Tracking of Beats in Small Data Scenarios

Our efforts in the annotation of beats and downbeats, as well as our preliminary investigation of the tracking ability of state-of-the-art models in the developed datasets, have led to the proposal of an entire annotation methodology.

First, we adapted a TCN-based meter tracking model using small quantities of data to work in datasets of *samba* and of another Latin American music tradition, *candombe*, assuming a certain level of homogeneity in the music genres. We have shown that, at least under this homogeneity condition, it is indeed possible to train a model with a few minutes of annotated data and training cycles, and obtain almost a “full-dataset” performance. We have also shown that, in this particular task, fine-tuning a base model, trained on a larger dataset, and using data augmentation in the process can largely improve the overall F-measure of the model.

Then, considering an end user’s perspective, we proposed an effective methodology for selecting training samples in this small data scenario. Our framework combined tempo-sensitive or tempo-robust rhythmic features with selective sampling techniques that exploit the internal distribution of the data. The system output was a selection of meaningful examples, which were subject to a user-informed annotation budget. In real-world applications, the user is then given this selection and should produce corresponding annotations. Finally, beat positions for the remaining tracks in the dataset were estimated with the TCN model.

Our experiments with this framework have highlighted the importance of carefully selecting the training data for the TCN model since our results demonstrated a marked improvement when compared to outcomes obtained through random data selection. The experiments have also indicated that there are complex non-linear interactions between the sizes of training and testing sets, the rhythmic properties of the dataset at hand, the features for rhythmic representation, and the different strategies for sampling. Nonetheless, we have confirmed our intuition that a more appropriate data selection has to leverage diversity and representativeness. Even though the datasets used in this study are very percussive, we believe that the same framework should also work for music with little to no percussive content.

Integrated Beat and Microtiming Tracking

Our final contribution in this work was in the characterization of microtiming profiles. The *ostinato* patterns played in the *tamborim* in *samba-enredo* were a good motivation for the definition of a model capable of simultaneously tracking beats and these genre-defining small-scale deviations. After having presented a statistical analysis of our *tamborim* subset, we presented our graphical model for tackling the extraction of beat-length microtiming patterns. To the best of our knowledge, this is the first method for simultaneously tracking beats and microtiming with this kind of model.

We have used our model with two inference techniques. First, by discretizing the microtiming feature, we performed exact inference with a CRF. Then, approximate inference approaches were exploited — two particle filter techniques with a resampling scheme to avoid degeneration. We have shown that, by searching for optimal trajectories in a continuous state space, the APF method can more accurately track microtiming than the exact inference approach.

It is important to note that our method for microtiming analysis is generalizable and adaptable to other genres, with different pattern lengths and number of tatums. In particular, it can be directly modified to deal with the estimation of swing ratio in jazz recordings, a problem that has been extensively reported in the literature.

13.2 Future Work

While the primary focus was on *samba* music, one of our most important goals was to ensure that the developed tools and methodologies were as general as possible. This means that they should be adaptable and applicable to other underrepresented music genres, thereby promoting diversity and inclusivity in the MIR field and expanding our comprehension of the human perception of musical phenomena. We are aware, however, that the viewpoint of our work is very limited and that there are several possibilities for further research in the topics we have presented.

For instance, our approach to the classification of articulations is very cumbersome, requiring multiple steps for the detection of notes, the segmentation of the performance, etc. Moreover, we did not consider modeling the grammar-like aspects of the sequence of articulations, which should probably provide the same level of improvement of the language modeling approach proposed by GILLET and RICHARD [176]. Of course, the entire pipeline, from detection to the investigation of temporal context could be absorbed by the more powerful deep learning-based techniques available to us in the last few years [281].

Our methodology for beat tracking with few data has promising consequences in real-world applications, as it opens the possibility of adapting such models to other music genres with modest labeling efforts. Furthermore, we believe that a similar pipeline could be utilized for efficient data selection in other supervised learning problems in MIR. This includes mood and genre classification as well as other retrieval tasks. However, it would also be important to validate the current methodology on other challenging music datasets for beat tracking (e.g., SMC [289], Hainsworth [9]). One simpler extension could be to investigate the effect of the selection pipeline on the related task of downbeat tracking. Another limitation of this work was that our analysis was conducted by looking just at two musical parameters: global tempo and rhythmic patterns. The complex interactions observed in the results could start to be disambiguated if we were to investigate the dynamics of these parameters (i.e., tempo and pattern changes within each track) as well as other important musical aspects (e.g., timbre, pattern complexity/density, among others). Finally, although we provided a solution for the selection of informative data for training, we have not investigated how to predict the necessary annotation budget for a certain expected tracking performance. It would be very interesting to understand what the model considers "challenging" — this goes along with current trends in AI explainability [345].

In our analysis of microtiming profiles, we envision many fronts for future contributions. It would be greatly beneficial to further investigate particle filter techniques and see if our results for tracking microtiming can be improved. We would also like to loosen the restrictions for tempo. This would be very important in music genres were there is great tempo variation during the performance (e.g., in *candombe* [346], in *maracatu* [320]). Moreover, this would enable an investigation of the dependency of microtiming pattern and tempo, as in [299]. This system could be coupled with a source separation approach and used in the investigation of entrainment among musicians in ensemble recordings [347].

Continued efforts are needed before we are able to capture expressiveness in percussive performances with fully-automatically computational methods. We hope that the pathways here envisioned can contribute to this field of music transcription, specially in providing some momentum for the development and annotation of other culturally diverse music datasets and for the pursuit of more humanistic and multicultural approaches in MIR methodologies.

Bibliography

[1] COLEMAN, E. *To be a drum.* Morton Grove, USA, Albert Whitman & Company, 1998.

[2] BLADES, J. *Percussion Instruments and Their History.* 4 ed. Westport, USA, The Bold Strummer, 1992.

[3] RANDEL, D. M. *The Harvard Dictionary of Music.* 4 ed. Cambridge, USA, Belknap Press, 2003.

[4] JONES, A. M. *Studies in African Music*, v. I. London, UK, Oxford University Press, 1959.

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

[6] SERRA, X. "A multicultural approach in music information research". In: *Proc. 12th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 151–156, Miami, USA, Oct. 2011.

[7] 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 Process.*, v. 90, n. 4, pp. 1032–1048, Apr. 2010.

[8] HUANG, R., HOLZAPFEL, A., STURM, B., et al. "Beyond Diverse Datasets: Responsible MIR, Interdisciplinarity, and the Fractured Worlds of Music", *Trans. Int. Soc. Music Inform. Retr.*, v. 6, n. 1, pp. 43–59, 2023.

[9] HAINSWORTH, S. W., MAACLEOD, M. D. "Particle Filtering Applied to Musical Tempo Tracking", *EURASIP J. Appl. Signal Process.*, pp. 2385–2395, Nov. 2004.

[10] GOUYON, F., DIXON, S. "A Review of Automatic Rhythm Description Systems", *Comput. Music J.*, v. 29, n. 1, pp. 34–54, Mar. 2005.

[11] TEMPERLEY, D. *The Cognition of Basic Musical Structures.* Cambridge, USA, Mit Press, 2001.

[12] LERDAHL, F., JACKENDOFF, R. *A Generative Theory of Tonal Music.* Cambridge, USA, MIT Press, 1983.

[13] LONDON, J. *Hearing in Time: Psychological Aspects of Musical Meter.* New York, USA, Oxford University Press, 2004.

[14] POLAK, R., LONDON, J., JACOBY, N. "Both Isochronous and Non-Isochronous Metrical Subdivision Afford Precise and Stable Ensemble Entrainment: A Corpus Study of Malian Jembe Drumming", *Frontiers in Neuroscience*, v. 10, n. 285, Jun. 2016.

[15] COCHARRO, D., BERNARDES, G., BERNARDO, G., et al. "A Review of Musical Rhythm Representation and (Dis)similarity in Symbolic and Audio Domains". In: Correia Castilho, L., Dias, R., Pinho, J. F. (Eds.), *Perspectives on Music, Sound and Musicology. Current Research in Systematic Musicology*, Springer, pp. 189–208, Cham, Switzerland, 2021.

[16] KLAPURI, A. "Introduction to Music Transcription". In: Klapuri, A., Davy, M. (Eds.), *Signal Processing Methods for Music Transcription*, Springer, pp. 3–20, New York, USA, 2006.

[17] PINTO, A. S., BÖCK, S., CARDOSO, J. S., et al. "User-Driven Fine-Tuning for Beat Tracking", *Electronics*, v. 10, n. 13, Jun. 2021.

[18] PROCKUP, M., SCHMIDT, E. M., SCOTT, J., et al. "Toward Understanding Expressive Percussion Through Content Based Analysis". In: *Proc. 14th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 143–148, Curitiba, Brazil, Nov. 2013.

[19] 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: *Proc. 2018 Latin American Congr. Audio Eng. Soc. (AES-LAC)*, pp. 53–60, Montevideo, Uruguay, Sep. 2018.

[20] MAIA, L. S., FUENTES, M., BISCAINHO, L. W. P., et al. "SAMBASET: A Dataset of Historical Samba de Enredo Recordings for Computational Music Analysis". In: *Proc. 20th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 628–635, Delft, The Netherlands, Nov. 2019.

[21] FUENTES, M., MAIA, L. S., ROCAMORA, M., et al. "Tracking Beats and Microtiming in Afro-Latin American Music Using Conditional Random

Fields and Deep Learning". In: *Proc. 20th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 251–258, Delft, The Netherlands, Nov. 2019.

[22] MAIA, L. S., ROCAMORA, M., BISCAINHO, L. W. P., et al. "Adapting Meter Tracking Models to Latin American Music". In: *Proc. 23rd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 361–368, Bengaluru, India, Dec. 2022.

[23] MAIA, L. S., ROCAMORA, M., BISCAINHO, L. W. P., et al. "Selective Annotation of Few Data for Beat Tracking of Latin American Music Using Rhythmic Features", *Trans. Int. Soc. Music Inform. Retr.*, v. 7, n. 1, pp. 99–112, Mar. 2024.

[24] ARAUJO JUNIOR, S. M. *Acoustic Labor in the Timing of Everyday Life: A Critical Contribution to the History of Samba in Rio de Janeiro.* PhD Thesis, University of Illinois, Urbana, USA, 1992.

[25] RUGENDAS, J. M. *Voyage pittoresque dans le Brésil.* Paris, France, Engelmann & Cie., 1827.

[26] VAINSENCHER, S. A. "Castro Alves". Pesquisa Escolar Online, Fundação Joaquim Nabuco, Recife, Brazil, Mar. 2018. Available at: `https://pesquisaescolar.fundaj.gov.br/en/artigo/castro-alves/`. Accessed: September 4, 2020.

[27] MOURA, R. *Tia Ciata e a Pequena África no Rio de Janeiro.* Coleção Biblioteca Carioca, v. 32. 2 ed. Rio de Janeiro, Brazil, Secretaria Municipal de Cultura, Departamento Geral de Documentação e Informação Cultural, 1995.

[28] IPHAN. *Matrizes do Samba no Rio de Janeiro: partido-alto, samba de terreiro, samba-enredo.* Brasília, Brazil, Instituto do Patrimônio Histórico e Artístico Nacional, 2014. Dossiê Iphan, 10.

[29] DAMATTA, R. *Carnavais, Malandros e Heróis: Para uma Sociologia do Dilema Brasileiro.* 6 ed. Rio de Janeiro, Brazil, Rocco, 1997.

[30] LOPES, N. *Enciclopédia Brasileira da Diáspora Africana.* 4 ed. São Paulo, Brazil, Selo Negro, 2014.

[31] SANDRONI, C. *Feitiço Decente: Transformações do samba no Rio de Janeiro (1917-1933).* 2 ed. Rio de Janeiro, Brazil, Jorge Zahar Ed./Ed. UFRJ, 2008.

[32] MARIANI, M. E. “African Influences in Brazilian Dance”. In: Asante, K. W. (Ed.), (1996). *African Dance: An Artistic, Historical, and Philosophical Inquiry*, Africa World Press, pp. 79–97, Trenton, USA, 2002.

[33] AMARAL, E. *Alguns Aspectos da MPB.* 2 ed. Rio de Janeiro, Brazil, Esteio Editora, 2010.

[34] CABRAL, S. *Escolas de Samba do Rio de Janeiro.* Rio de Janeiro, Brazil, Lazuli, 2011.

[35] NETO, L. *Uma História do Samba*, v. I. São Paulo, Brazil, Companhia das Letras, 2006.

[36] FRYER, P. *Rhythms of Resistance: African Musical Heritage in Brazil.* London, UK, Pluto Press, 2000.

[37] FAUSTO, B. *História Concisa do Brasil.* 2 ed. São Paulo, Brazil, Editora da Universidade de São Paulo, 2006.

[38] ROCAMORA, M. *Computational Methods for Percussion Music Analysis: the Afro-Uruguayan Candombe Drumming as a Case Study.* PhD Thesis, Universidad de la República, Montevideo, Uruguay, 2018.

[39] TINHORÃO, J. R. *Os Sons que Vêm da Rua.* 2 ed. São Paulo, Brazil, Editora 34, 2000.

[40] MASCHEK, E. “Planta da cidade do Rio de Janeiro e de uma parte dos Subúrbios”. Laemmert e Cia., 1885. Map, litograph, 81 x 100cm. Scale 1:10000. Available at: `http://objdigital.bn.br/objdigital2/acervo_digital/div_cartografia/cart219156/cart219156.jpg`. Accessed: September 5, 2020.

[41] SOARES, C. E. L. *Valongo, Cais dos Escravos: Memória da Diáspora e Modernização Portuária na Cidade do Rio de Janeiro.* Postdoctoral final report, National Museum, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, 2013.

[42] EFEGÊ, J. *Maxixe — A Dança Excomungada.* 1 ed. Rio de Janeiro, Brazil, Conquista, 1974.

[43] FERREZ, M. “Negras”. 1870-1899. Photograph, 15 x 21.5 cm. Available at: `http://objdigital.bn.br/acervo_digital/div_iconografia/icon846210.jpg`. Accessed: September 23, 2020.

[44] MOURA, R. M. *No Princípio, Era a Roda: Um estudo sobre samba, partido-alto e outros pagodes*. Rio de Janeiro, Brazil, Rocco, 2004.

[45] MALTA, A. "Praça 11 de Junho - Rio". 1922. Photograph, 17.3 x 23.3 cm. Available at: `http://objdigital.bn.br/objdigital2/acervo_digital/div_iconografia/icon1411486/icon1411486.jpg`. Accessed: September 17, 2020.

[46] DEBRET, J. B. *Voyage pittoresque et historique au Brésil*, v. 2. Paris, France, Firmin Didot Frères, 1835.

[47] GONÇALVES, R. S. "Cronistas, folcloristas e os ranchos carnavalescos: perspectivas sobre a cultura popular", *Estudos Históricos*, v. 2, n. 32, pp. 89–105, Jul. 2003.

[48] FERNANDES, N. N. *Escolas de Samba: Sujeitos Celebrantes e Objetos Celebrados*. 3 ed. Rio de Janeiro, Brazil, Editora 34, 2001.

[49] CARNEIRO, E. *Carta do Samba*. Rio de Janeiro, Brazil, Centro Nacional de Folclore e Cultura Popular, 2012.

[50] CANÇADO, T. M. L. "O 'fator atrasado' na música brasileira: evolução, características e interpretação", *Per Musi*, v. 2, pp. 5–14, Jul. 2000.

[51] PAZ, E. A. *500 Canções Brasileiras*. 2 ed. Brasília, Brazil, MusiMed, 2010.

[52] GONZAGA, F. "Gaúcho: O Corta-Jaca de Cá e Lá; Tango Brasileiro". Acervo Digital Chiquinha Gonzaga, [S.l.], 2011. Available at: `http://www.chiquinhagonzaga.com/acervo/partituras/gaucho_ca-e-la_piano.pdf`. Accessed: September 15, 2020. Digital.

[53] NAZARETH, E. "Odeon: Tango Brasileiro". Musica Brasilis & Instituto Moreira Salles, [S.l.], 2012. Available at: `http://ernestonazareth150anos.com.br/files/uploads/work_elements/work_136/odeon_piano.pdf`. Accessed: September 24, 2020. Digital.

[54] ABREU, Z. "Tico-Tico no Fubá: Chôro Sapéca". Irmãos Vitale S/A, São Paulo, Brazil, 1941. Available at: `https://imslp.org/wiki/Special:ReverseLookup/211430`. Accessed: September 15, 2020. Print.

[55] SANTOS, E. "Pelo Telephone: Samba Carnavalesco". [s.n.], Rio de Janeiro, Brazil, 1916. Available at: `http://objdigital.bn.br/acervo_digital/div_musica/mas1147601.pdf`. Accessed: September 15, 2020. Manuscript.

[56] NKETIA, J. H. K. *The Music of Africa.* New York, USA, W. W. Norton, 1974.

[57] VIZEU, C. M. O. *O Samba-Enredo Carioca e Suas Transformações nas Décadas de 70 e 80: Uma análise musical.* Master Dissertation, University of Campinas, Campinas, Brazil, 2004.

[58] PRADO, Y. "Padrões Musicais do Samba-Enredo na Era do Sambódromo", *Música em Perspectiva*, v. 8, n. 1, pp. 155–195, Jun. 2015.

[59] AMORIM, L. C. *As Baterias das Escolas de Samba Cariocas do Grupo Especial: Trabalho Acústico e os Impactos do Concurso Sobre o seu Caráter Humanizador.* Master Dissertation, Federal University of the State of Rio de Janeiro, Rio de Janeiro, Brazil, 2014.

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

[61] MCFEE, B., HUMPHREY, E. J., BELLO, J. P. "A Software Framework for Musical Data Augmentation". In: *Proc. 16th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 248–254, Málaga, Spain, Oct. 2015.

[62] CARTWRIGHT, M., BELLO, J. P. "Increasing Drum Transcription Vocabulary Using Data Synthesis". In: *Proc. 21st Int. Conf. Digit. Audio Effects (DAFx)*, pp. 72–79, Aveiro, Portugal, Sep. 2018.

[63] NUNES, L. O., ROCAMORA, M., JURE, L., et al. "Beat and downbeat tracking based on rhythmic patterns applied to the Uruguayan candombe drumming". In: *Proc. 16th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 264–270, Málaga, Spain, Oct. 2015.

[64] SILLA JR., C. N., KOERICH, A. L., KAESTNER, C. A. A. "The Latin Music Database". In: *Proc. 9th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 451–456, Philadelphia, USA, Sep. 2008.

[65] SOUSA, J. M., PEREIRA, E. T., VELOSO, L. R. "A robust music genre classification approach for global and regional music datasets evaluation". In: *Proc. 2016 IEEE Int. Conf. Digit. Signal Process. (DSP)*, pp. 109–113, Beijing, China, Oct. 2016.

[66] GILLET, O., RICHARD, G. "ENST-Drums: An Extensive Audio-Visual Database for Drum Signals Processing". In: *Proc. 7th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 156–159, Victoria, Canada, Oct. 2006.

[67] SOUTHALL, C., WU, C.-W., LERCH, A., et al. “MDB Drums: An Annotated Subset of MedleyDB for Automatic Drum Transcription”. In: *18th Int. Soc. Music Inform. Retr. Conf. (ISMIR) Late Breaking and Demo Papers*, Suzhou, China, Oct. 2017.

[68] BITTNER, R., SALAMON, J., TIERNEY, M., et al. “MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research”. In: *Proc. 15th Conf. Int. Soc. Music Inform. Retr. (ISMIR)*, pp. 155–160, Taipei, Taiwan, Oct. 2014.

[69] VOGL, R., WIDMER, G., KNEES, P. “Towards Multi-Instrument Drum Transcription”. In: *Proc. 21st Int. Conf. Digit. Audio Effects (DAFx)*, pp. 57–64, Aveiro, Portugal, Sep. 2018.

[70] 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.

[71] DIXON, S. “Onset Detection Revisited”. In: *Proc. 9th Int. Conf. Digit. Audio Effects (DAFx)*, pp. 133–137, Montreal, Canada, Sep. 2006.

[72] BÖCK, S., KORZENIOWSKI, F., SCHLÜTER, J., et al. “madmom: a New Python Audio and Music Signal Processing Library”. In: *Proc. 24th ACM Int. Multimedia Conf. (MM)*, pp. 1174–1178, Amsterdam, Netherlands, Oct. 2016.

[73] DA COSTA, M. V. M. *Novel Time-Frequency Representations for Music Information Retrieval*. PhD Thesis, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, 2020.

[74] ROCAMORA, M., JURE, L., MARENCO, B., et al. “An audio-visual database of Candombe performances for computational musicological studies”. In: *Memorias del II Congreso Int. de Ciencia y Tecnología Musical (CICTeM)*, pp. 17–24, Buenos Aires, Argentina, Sep. 2015.

[75] GOUYON, F., KLAPURI, A., DIXON, S., et al. “An Experimental Comparison of Audio Tempo Induction Algorithms”, *IEEE Trans. Audio, Speech, Lang. Process.*, v. 14, n. 5, pp. 1832–1844, Sep. 2006.

[76] KREBS, F., BÖCK, S., WIDMER, G. “Rhythmic Pattern Modeling for Beat and Downbeat Tracking in Musical Audio”. In: *Proc. 14th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 227–232, Curitiba, Brazil, Nov. 2013.

[77] HAYKIN, S., VEEN, B. V. *Signals and Systems.* New York, USA, John Wiley and Sons, 1999.

[78] ALLEN, J. B., RABINER, L. R. "A Unified Approach to Short-Time Fourier Analysis and Synthesis", *Proc. IEEE*, v. 65, n. 11, pp. 1558–1564, Nov. 1977.

[79] BROWN, J. C. "Calculation of a Constant Q Spectral Transform", *J. Acoust. Soc. Amer.*, v. 89, n. 1, pp. 425–434, Jan. 1991.

[80] SCHÖRKHUBER, C., KLAPURI, A. "Constant-Q Transform Toolbox for Music Processing". In: *Proc. 7th Sound Music Comput. Conf. (SMC)*, Barcelona, Spain, Jul. 2010.

[81] BROWN, J. C., PUCKETTE, M. S. "An Efficient Algorithm for the Calculation of a Constant Q Transform", *J. Acoust. Soc. Amer.*, v. 92, n. 5, pp. 2698–2701, Nov. 1992.

[82] BURRUS, C. S., GOPINATH, R. A., GUO, H. *Introduction to Wavelets and Wavelet Transforms: A Primer.* Upper Saddle River, USA, Prentice Hall, 1998.

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

[84] CLARK, P., ATLAS, L. "A Sum-of-Products Model for Effective Coherent Modulation Filtering". In: *Proc. 2009 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 4485–4488, Taipei, Taiwan, Apr. 2009.

[85] LOUGHLIN, P. J., TACER, B. "On the Amplitude- and Frequency-Modulation Decomposition of Signals", *J. Acoust. Soc. America*, v. 100, n. 3, pp. 1594–1601, Sep. 1996.

[86] COHEN, L., LOUGHLIN, P., VAKMAN, D. "On an Ambiguity in the Definition of the Amplitude and Phase of a Signal", *Signal Process.*, v. 79, n. 3, pp. 301–307, Dec. 1999.

[87] ATLAS, L., SHAMMA, S. A. "Joint Acoustic and Modulation Frequency", *EURASIP J. Appl. Signal Process.*, pp. 668–675, Jun. 2003.

[88] COHEN, L. "The scale representation", *IEEE Trans. Signal Process.*, v. 41, n. 12, pp. 3275–3292, Dec. 1993.

[89] DE SENA, A., ROCCHESSO, D. "A Fast Mellin and Scale Transform", *EURASIP J. Adv. Signal Process.*, v. 2007, Dec. 2007. doi: 10.1155/2007/89170.

[90] WILLIAMS, W. J., ZALUBAS, E. J. "Helicopter Transmission Fault Detection via Time-Frequency, Scale and Spectral Methods", *Mech. Syst. Signal Process.*, v. 14, n. 4, pp. 545–559, 2000. doi: 10.1006/mssp.2000.1296.

[91] ALÍAS, F., SOCORÓ, J. C., SEVILLANO, X. "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds", *Appl. Sciences*, v. 6, n. 5, May 2016.

[92] PEETERS, G. *A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project.* Technical report, IRCAM, Paris, France, 2004.

[93] SCHLOSS, W. A. *On the Automatic Transcription of Percussive Music – From Acoustic Signal to High-Level Analysis.* PhD Thesis, Stanford University, Stanford, USA, 1985.

[94] ZÖLZER, U. *Digital Audio Signal Processing.* New York, USA, John Wiley and Sons, 1997.

[95] ESSID, S. *Classification Automatique des Signaux Audio-Fréquences : Reconnaissance des Instruments de Musique.* PhD Thesis, Université Pierre et Marie Curie, Paris, France, 2005.

[96] CAETANO, M. F., RODET, X. "Improved Estimation of the Amplitude Envelope of Time Domain Signals Using True Envelope Cepstral Smoothing". In: *Proc. 2011 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 11–21, Prague, Czech Republic, May 2011.

[97] GOUYON, F., PACHET, F., DELERUE, O. "On the Use of Zero-Crossing Rate for an Application of Classification of Percussive Sounds". In: *Proc. COST G-6 Conf. Digit. Audio Effects (DAFx)*, Verona, Italy, Dec. 2000.

[98] BELLO, J. P., DAUDET, L., ABDALLAH, S., et al. "A Tutorial on Onset Detection in Music Signals", *IEEE Trans. Speech Audio Process.*, v. 13, n. 5, pp. 1035–1047, Sep. 2005.

[99] VAN STEELANT, D., TANGHE, K., DEGROEVE, S., et al. "Classification of Percussive Sounds Using Support Vector Machines". In: *Proc. Annu. Machine Learn. Conf. Belgium Netherlands (BeNeLearn)*, pp. 146–153, Brussels, Belgium, Jan. 2004.

[100] FRASER, A., FUJINAGA, I. "Toward Real-Time Recognition of Acoustic Musical Instruments". In: *Proc. 1999 Int. Comput. Music Conf. (ICMC)*, pp. 175–177, Beijing, China, Oct. 1999.

[101] HERRERA, P., YETERIAN, A., GOUYON, F. "Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques". In: *Proc. 2nd Int. Conf. Music Artif. Intell. (ICMAI)*, pp. 69–80, Edinburgh, Scotland, Sep. 2002.

[102] WU, C.-W., LERCH, A. "On Drum Playing Technique Detection in Polyphonic Mixtures". In: *Proc. 17th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 218–224, New York, USA, Aug. 2016.

[103] CHORDIA, P. "Segmentation and Recognition of Tabla Strokes". In: *Proc. 6th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 107–114, London, UK, Sep. 2005.

[104] GOUYON, F., HERRERA, P. "Exploration of Techniques for Automatic Labeling of Audio Drum Tracks Instruments". In: *Proc. Workshop Current Directions Comput. Music (MOSART)*, Barcelona, Spain, Nov. 2001.

[105] SCHUBERT, E., WOLFE, J., TARNOPOLSKY, A. "Spectral Centroid and Timbre in Complex, Multiple Instrumental Textures". In: *Proc. 8th Int. Conf. Music Perception Cogn. (ICMPC)*, pp. 654–657, Evanston, USA, Aug. 2004.

[106] BOGDANOV, D., WACK, N., GÓMEZ, E., et al. "ESSENTIA: an Audio Analysis Library for Music Information Retrieval". In: *Proc. 14th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 493–498, Curitiba, Brazil, Nov. 2013.

[107] AKKERMANS, V., SERRÀ, J., HERRERA, P. "Shape-Based Spectral Contrast Descriptor". In: *Proc. 6th Sound Music Comput. Conf. (SMC)*, pp. 143–148, Porto, Portugal, Jul. 2009.

[108] JIANG, D.-N., LU, L., ZHANG, H.-J., et al. "Music Type Classification by Spectral Contrast Feature". In: *Proc. 2016 IEEE Int. Conf. Multimedia Expo (ICME)*, pp. 113–116, Lausanne, Switzerland, Aug. 2002.

[109] MOREAU, A., FLEXER, A. "Drum Transcription in Polyphonic Music Using Non-Negative Matrix Factorisation". In: *Proc. 8th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 353–354, Vienna, Austria, Sep. 2007.

[110] CHILDERS, D. G., SKINNER, D. P., KEMERAIT, R. C. "The Cepstrum: A Guide to Processing", *Proc. IEEE*, v. 65, n. 10, pp. 1428–1443, Oct. 1977.

[111] BROWN, J. C. "Computer Identification of Musical Instruments Using Pattern Recognition with Cepstral Coefficients as Features", *J. Acoust. Soc. Amer.*, v. 105, n. 3, pp. 1933–1941, Mar. 1999.

[112] TODISCO, M., DELGADO, H., EVANS, N. "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification", *Comput. Speech & Lang. (CSL)*, v. 45, pp. 516–535, Sep. 2017.

[113] O'SHAUGHNESSY, D. *Speech Communication: Human and Machine.* Reading, USA, Addison-Wesley, 1987.

[114] RABINER, L. R., JUANG, B.-H. *Fundamentals of Speech Recognition.* Englewood Cliffs, USA, PTR Prentice Hall, 1993.

[115] GANCHEV, T., FAKOTAKIS, N., KOKKINAKIS, G. "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task". In: *Proc. 10th Int. Conf. Speech Comput. (SPECOM)*, pp. 191–194, Patras, Greece, Oct. 2005.

[116] REN, E., LOELIGER, H.-A. "Exact Discrete-Time Realizations of the Gammatone Filter". In: *Proc. 2019 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 316–320, Brighton, UK, May 2019.

[117] PATTERSON, R. D., ROBINSON, K., HOLDSWORTH, J., et al. "Complex Sounds and Auditory Images". In: Cazals, Y., Demany, L., Horner, K. (Eds.), *Auditory Physiology and Perception: Proc. 9th International Symposium on Hearing*, Pergamon Press, pp. 429–446, Oxford, UK, 1992.

[118] SHAO, Y., JIN, Z., WANG, D., et al. "An Auditory-Based Feature for Robust Speech Recognition". In: *Proc. 2009 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 4625–4628, Taipei, Taiwan, Apr. 2009.

[119] ZHAO, X., WANG, D. "Analyzing Noise Robustness of MFCC and GFCC Features in Speaker Identification". In: *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 7204–7208, Vancouver, Canada, May 2013.

[120] FASTL, H., ZWICKER, E. *Psychoacoustics: Facts and Models.* 3 ed. Berlin, Germany, Springer, 2007.

[121] TRAUNMÜLLER, H. "Analytical Expressions for the Tonotopic Sensory Scale", *J. Acoust. Soc. America*, v. 89, n. 1, pp. 97–100, Jul. 1990.

[122] HERMANSKY, H., HANSON, B. A., WAKITA, H. "Low-Dimensional Representation of Vowels Based on All-Pole Modeling in the Psychophysical Domain", *Speech Commun.*, v. 4, n. 1-3, pp. 181–187, Aug. 1985.

[123] BOSI, M., GOLDBERG, R. E. *Introduction to Digital Audio Coding and Standards.* New York, USA, Springer, 2003.

[124] HERRERA, P., DEHAMEL, A., GOUYON, F. "Automatic Labeling of Unpitched Percussion Sounds". In: *Proc. 114th Audio Eng. Soc. Conv. (AES)*, pp. 69–80, Amsterdam, The Netherlands, Mar. 2003.

[125] BRENT, W. "Perceptually Based Pitch Scales in Cepstral Techniques for Percussive Timbre Identification". In: *Proc. 2009 Int. Comput. Music Conf. (ICMC)*, Montreal Canada, Aug. 2009.

[126] BRENT, W. "Cepstral Analysis Tools for Percussive Timbre Identification". In: *Proc. 3rd Int. Pure Data Conv. (PdCon)*, pp. 121–124, São Paulo, Brazil, Jul. 2009.

[127] BRENT, W. *Physical and Perceptual Aspects of Percussive Timbre.* PhD Thesis, University of California, San Diego, USA, 2010.

[128] ANDÉN, J., MALLAT, S. "Deep Scattering Spectrum", *IEEE Trans. Signal Process.*, v. 62, n. 16, pp. 4414–4428, Aug. 2014.

[129] KINGSBURY, B. E. D., MORGAN, N., GREENBERG, S. "Robust Speech Recognition Using the Modulation Spectrogram", *Speech Commun.*, v. 25, n. 1, pp. 117–132, Jun. 1998.

[130] KINNUNEN, T. "Joint Acoustic-Modulation Frequency for Speaker Recognition". In: *Proc. 2006 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 665–668, Toulouse, France, May 2006.

[131] KINNUNEN, T., LEE, K., LI, H. "Dimension Reduction of the Modulation Spectrogram for Speaker Verification". In: *Proc. 2008 Speaker Lang. Recognit. Workshop (Odyssey)*, Stellenbosch, South Africa, Jan. 2008.

[132] FALK, T. H., CHAN, W. "Modulation Spectral Features for Robust Far-Field Speaker Identification", *IEEE Trans. Audio, Speech, Lang. Process.*, v. 18, n. 1, pp. 90–100, Jan. 2010.

[133] AHMADI, S., AHADI, S. M., CRANE, B., et al. "Sparse Coding of the Modulation Spectrum for Noise-Robust Automatic Speech Recognition", *EURASIP J. Audio, Speech, Music Process.*, Oct. 2014.

[134] SARRIA-PAJA, M., FALK, T. H. “Whispered Speech Detection in Noise Using Auditory-Inspired Modulation Spectrum Features”, *IEEE Signal Process. Lett.*, v. 20, n. 8, pp. 783–786, Aug. 2013.

[135] WU, S., FALK, T. H., CHAN, W. “Automatic Speech Emotion Recognition Using Modulation Spectral Features”, *Speech Commun.*, v. 53, n. 5, pp. 768–785, May 2011.

[136] ZHU, Z., MIYAUCHI, R., ARAKI, Y., et al. “Feasibility of Vocal Emotion Conversion on Modulation Spectrogram for Simulated Cochlear Implants”. In: *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1884–1888, Kos, Greece, Aug. 2017.

[137] MCKINNEY, M. F., BREEBAART, J. “Features for Audio and Music Classification”. In: *Proc. 4th Int. Conf. Music Inform. Retr. (ISMIR)*, Baltimore, USA, Oct. 2003.

[138] LEE, C., SHIH, J., YU, K., et al. “Automatic Music Genre Classification Using Modulation Spectral Contrast Feature”. In: *Proc. 2007 IEEE Int. Conf. Multimedia Expo (ICME)*, pp. 204–207, Beijing, China, Jul. 2007.

[139] LEE, C., SHIH, J., YU, K., et al. “Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features”, *IEEE Trans. Multimedia*, v. 11, n. 4, pp. 670–682, Jun. 2009.

[140] CONT, A., DUBNOV, S., WESSEL, D. “Realtime Mutiple-Pitch and Multiple-Instrument Recognition for Music Signals using Sparse Non-Negative Constraints”. In: *Proc. 10th Int. Conf. Digit. Audio Effects (DAFx)*, pp. 85–92, Bordeaux, France, Sep. 2007.

[141] SHI, Y., ZHU, X., KIM, H., et al. “A Tempo Feature via Modulation Spectrum Analysis and its Application to Music Emotion Classification”. In: *Proc. 2006 IEEE Int. Conf. Multimedia Expo (ICME)*, pp. 1085–1088, Toronto, Canada, Jul. 2006.

[142] SCHIMMEL, S. M., ATLAS, L. E., NIE, K. “Feasibility of Single Channel Speaker Separation Based on Modulation Frequency Analysis”. In: *Proc. 2007 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 605–608, Honolulu, USA, Apr. 2007.

[143] BARKER, T., VIRTANEN, T. “Non-Negative Tensor Factorization of Modulation Spectrograms for Monaural Sound Source Separation”. In: *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, pp. 827–831, Lyon, France, Aug. 2013.

[144] SUKITTANON, S., ATLAS, L. E., PITTON, J. W. "Modulation-Scale Analysis for Content Identification", *IEEE Trans. Signal Process.*, v. 52, n. 10, pp. 3023–3035, Oct. 2004.

[145] VINTON, M. S., ATLAS, L. E. "Scalable and Progressive Audio Codec". In: *Proc. 2001 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 3277–3280, Salt Lake City, USA, May 2001.

[146] THOMPSON, J. K., ATLAS, L. E. "A Non-Uniform Modulation Transform for Audio Coding with Increased Time Resolution". In: *Proc. 2003 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 397–400, Hong Kong, China, Apr. 2003.

[147] GREENBERG, S., KINGSBURY, B. E. D. "The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech". In: *Proc. 1997 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 1647–1650, Munich, Germany, Apr. 1997.

[148] ELLIS, D., ZENG, X., MCDERMOTT, J. "Classifying Soundtracks with Audio Texture Features". In: *Proc. 2011 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 5880–5883, Prague, Czech Republic, May 2011.

[149] SCHIMMEL, S., ATLAS, L. "Coherent Envelope Detection for Modulation Filtering of Speech". In: *Proc. 2005 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 221–224, Philadelphia, USA, Mar. 2005.

[150] NAGATHIL, A., GERKMANN, T., MARTIN, R. "Musical Genre Classification Based on a Highly-Resolved Cepstral Modulation Spectrum". In: *Proc. 18th Eur. Signal Process. Conf. (EUSIPCO)*, pp. 462–466, Aalborg, Denmark, Aug. 2010.

[151] REN, J., WU, M., JANG, J. R. "Automatic Music Mood Classification Based on Timbre and Modulation Features", *IEEE Trans. Affective Comput.*, v. 6, n. 3, pp. 236–246, Jul. 2015.

[152] CHI, T., RU, P., SHAMMA, S. A. "Multiresolution Spectrotemporal Analysis of Complex Sounds", *J. Acoust. Soc. Amer.*, v. 118, n. 2, pp. 887–906, Aug. 2005.

[153] SCHÖRKHUBER, C., KLAPURI, A., HOLIGHAUS, N., et al. "A Matlab Toolbox for Efficient Perfect Reconstruction Time-Frequency Transforms with Log-Frequency Resolution". In: *Proc. 53rd Audio Eng. Soc. Int. Conf. Semantic Audio (AES)*, London, UK, Jan. 2014.

[154] BAUGÉ, C., LAGRANGE, M., ANDÉN, J., et al. "Representing Environmental Sounds Using the Separable Scattering Transform". In: *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 8667–8671, Vancouver, Canada, Oct. 2013.

[155] ANDÉN, J., LOSTANLEN, V., MALLAT, S. "Joint Time-Frequency Scattering", *IEEE Trans. Signal Process.*, v. 67, n. 14, pp. 3704–3718, Jul. 2019.

[156] CROS VILA, L. *Musical Instrument Recognition Using the Scattering Transform.* Master Dissertation, KTH Royal Institute of Technology, Stockholm, Sweden, 2020.

[157] HAN, H., LOSTANLEN, V. "WAV2SHAPE: Hearing the Shape of a Drum Machine". In: *Proc. Forum Acusticum*, pp. 647–654, Lyon, France, Dec. 2020.

[158] WANG, C., LOSTANLEN, V., BENETOS, E., et al. "Playing Technique Recognition by Joint Time–Frequency Scattering". In: *Proc. 2020 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 881–885, Barcelona, Spain, May 2020.

[159] LOSTANLEN, V., EL-HAJJ, C., ROSSIGNOL, M., et al. "Time-Frequency Scattering Accurately Models Auditory Similarities Between Instrumental Playing Techniques", *EURASIP J. Audio, Speech, Music Process.*, Jan. 2021.

[160] PAULUS, J. *Signal Processing Methods for Drum Transcription and Music Structure Analysis.* PhD Thesis, Tampere University of Technology, Tampere, Finland, 2009.

[161] RAMAN, C. V. "The Indian Musical Drums", *Proc. Indian Acad. Sci.*, v. 1, pp. 179–188, Sep. 1934.

[162] FLETCHER, N. H., ROSSING, T. D. *The Physics of Musical Instruments.* 2 ed. New York, USA, Springer, 1998.

[163] DITTMAR, C. *Source Separation and Restoration of Drum Sounds in Music Recordings.* PhD Thesis, University of Erlangen-Nuremberg, Erlangen-Nuremberg, Germany, 2018.

[164] OLSON, H. F. *Music, Physics and Engineering.* 2 ed. New York, USA, Dover Publications, 1967.

[165] BELL, R. *PAL: the Percussive Audio Lexicon.* PhD Thesis, Swinburne University of Technology, Melbourne, Australia, 2015.

[166] TINDALE, A. *Classification of Snare Drum Sounds Using Neural Networks.* Master Dissertation, McGill University, Montreal, Canada, 2004.

[167] GILLET, O. *Transcription des Signaux Percussifs. Application à l'Analyse de Scènes Musicales Audiovisuelles.* PhD Thesis, Télécom ParisTech, Paris, France, 2007.

[168] BEROIS, M. H. *Detecting and Describing Percussive Events in Polyphonic Music.* Master Dissertation, Universitat Pompeu Fabra, Barcelona, Spain, 2008.

[169] TINDALE, A. *Advancing the Art of Electronic Percussion.* PhD Thesis, University of Victoria, Victoria, Canada, 2009.

[170] WU, C.-W. *Addressing the Data Challenge in Automatic Drum Transcription with Labeled and Unsabeled Data.* PhD Thesis, Georgia Institute of Technology, Atlanta, USA, 2018.

[171] BILMES, J. A. *Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Expressive Timing in Percussive Rhythm.* Master Dissertation, Massachusets Institute of Technology, Cambridge, USA, 1993.

[172] TINDALE, A., KAPUR, A., TZANETAKIS, G., et al. "Retrieval of Percussion Gestures Using Timbre Classification Techniques". In: *Proc. 5th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 541–544, Barcelona, Spain, Oct. 2004.

[173] SOUZA, V. M. A., BATISTA, G. E. A. P. A., SOUZA-FILHO, N. E. "Automatic Classification of Drum Sounds with Indefinite Pitch". In: *Proc. 2015 Int. Joint Conf. Neural Networks (IJCNN)*, Killarney, Ireland, Jul. 2015.

[174] CHESHIRE, M., STABLES, R., HOCKMAN, J. "Investigating Timbral Differences of Varied Velocity Snare Drum Strikes". In: *Proc. 148th Audio Eng. Soc. Conv. (AES)*, Online, Jun. 2020.

[175] ROY, P., PACHET, F., KRAKOWSKI, S. "Analytical Features for the Classification of Percussive Sounds: the Case of the Pandeiro". In: *Proc. 10th Int. Conf. Digit. Audio Effects (DAFx)*, pp. 213–220, Bordeaux, France, Sep. 2007.

[176] GILLET, O., RICHARD, G. “Automatic Labelling of Tabla Signals”. In: *Proc. 4th Int. Conf. Music Inform. Retr. (ISMIR)*, Baltimore, USA, Oct. 2003.

[177] ROHIT, M. A., BHATTACHARJEE, A., RAO, P. “Four-Way Classification of Tabla Strokes with Models Adapted from Automatic Drum Transcription”. In: *Proc. 22nd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 19–26, Online, Nov. 2021.

[178] ANANTAPADMANABHAN, A., BELLUR, A., MURTHY, H. A. “Modal Analysis and Transcription of Strokes of the Mridangam Using Non-Negative Matrix Factorization”. In: *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 181–185, Vancouver, Canada, Oct. 2013.

[179] ANANTAPADMANABHAN, A., BELLO, J. P., KRISHNAN, R., et al. “Tonic-Independent Stroke Transcription of the Mridangam”. In: *Proc. 53rd Audio Eng. Soc. Int. Conf. Semantic Audio (AES)*, London, UK, Jan. 2014.

[180] RABINER, L. R. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, *Proc. IEEE*, v. 77, n. 2, pp. 257–286, Feb. 1989.

[181] KURIAKOSE, J., KUMAR, J. C., SARALA, P., et al. “Akshara Transcription of Mrudangam Strokes in Carnatic Music”. In: *Proc. 21st Nat. Conf. Commun. (NCC)*, Mumbai, India, Feb. 2015.

[182] BOLÃO, O. *Batuque is a Privilege: Percussion in the Music of Rio de Janeiro for Musicians, Arrangers and Composers.* São Paulo, Brazil, Irmãos Vitale, 2010.

[183] BÖCK, S., KREBS, F., SCHEDL, M. “Evaluating the Online Capabilities of Onset Detection Methods”. In: *Proc. 13th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 49–54, Porto, Portugal, Oct. 2012.

[184] DANIELSEN, A., NYMOEN, K., ANDERSON, E., et al. “Where is the beat in that note? Effects of attack, duration, and frequency on the perceived timing of musical and quasi-musical sounds”, *J. Exp. Psychol.: Human Perception and Performance*, v. 45, n. 3, pp. 402–418, Mar. 2019.

[185] LONDON, J., NYMOEN, K., LANGERØD, M. T., et al. “A Comparison of Methods for Investigating the Perceptual Center of Musical Sounds”, *Attention, Perception, & Psychophysics*, v. 81, n. 6, pp. 2088–2101, Jun. 2019.

[186] ANDREUX, M., ANGLES, T., EXARCHAKIS, G., et al. “Kymatio: Scattering Transforms in Python”, *J. Machine Learn. Res.*, v. 21, n. 60, pp. 1–6, Jan. 2020.

[187] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., et al. “Scikit-learn: Machine Learning in Python”, *J. Mach. Learn. Res.*, v. 12, pp. 2825–2830, Oct. 2011.

[188] PEETERS, G. “Spectral and Temporal Periodicity Representations of Rhythm for the Automatic Classification of Music Audio Signal”, *IEEE Trans. Audio, Speech, Lang. Process.*, v. 19, n. 5, pp. 1242–1252, Jul. 2011.

[189] TODD, N. P. M. “Multi-Scale Analysis of Expressive Signals: Recovery of Structure and Motion”. In: *Proc. Stockholm Music Acoust. Conf. (SMAC)*, pp. 146–149, Stockholm, Sweden, Jul. 1993.

[190] SMITH, L. M. *A Multiresolution Time-Frequency Analysis and Interpretation of Musical Rhythm.* PhD Thesis, University of Western Australia, Perth, Australia, 2000.

[191] TOUSSAINT, G. “A Comparison of Rhythmic Dissimilarity Measures”, *Forma*, v. 21, n. 2, pp. 129–149, 2006.

[192] TOUSSAINT, G. T. “A Comparison of Rhythmic Similarity Measures”. In: *Proc. 5th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 10–14, Barcelona, Spain, Oct. 2004.

[193] TOUSSAINT, G. T. *The Geometry of Musical Rhythm. What Makes a ‘Good’ Rhythm Good?* 2 ed. New York, USA, CRC Press, 2019.

[194] ROADS, C. *The Computer Music Tutorial.* Cambridge, USA, MIT Press, 1996.

[195] MISGELD, O., GULZ, T., MINIOTAITĖ, J., et al. “A Case Study of Deep Enculturation and Sensorimotor Synchronization to Real Music”. In: *Proc. 22nd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 460–467, Online, Nov. 2021. doi: 10.5281/zenodo.5624537.

[196] KREBS, F., HOLZAPFEL, A., CEMGIL, A. T., et al. “Inferring Metrical Structure in Music Using Particle Filters”, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, v. 23, n. 5, pp. 817–827, May 2015.

[197] DITTMAR, C., PFLEIDERER, M., BALKE, S., et al. “A swingogram representation for tracking micro-rhythmic variation in jazz performances”, *J. New Music Res.*, v. 47, n. 2, pp. 97–113, 2018.

[198] DITTMAR, C., PFLEIDERER, M., MÜLLER, M. “Automated Estimation of Ride Cymbal Swing Ratios in Jazz Recordings”. In: *Proc. 16th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 271–277, Málaga, Spain, Oct. 2015.

[199] IYER, V. “Embodied Mind, Situated Cognition, and Expressive Microtiming in African-American Music”, *Music Perception*, v. 19, n. 3, pp. 387–414, Mar. 2002.

[200] WAADELAND, C. H. “‘It Don’t Mean a Thing If It Ain’t Got That Swing’ - Simulating Expressive Timing by Modulated Movements”, *J. New Music Res.*, v. 30, n. 1, pp. 23–37, Mar. 2001.

[201] NAVEDA, L., GOUYON, F., GUEDES, C., et al. “Microtiming Patterns and Interactions with Musical Properties in Samba Music”, *J. New Music Res.*, v. 40, n. 3, pp. 225–238, Oct. 2011.

[202] WRIGHT, M., BERDAHL, E. “Towards Machine Learning of Expressive Microtiming in Brazilian Drumming”. In: *Proc. 2006 Int. Comput. Music Conf. (ICMC)*, pp. 572–575, New Orleans, USA, Nov. 2006.

[203] JURE, L., ROCAMORA, M. “Microtiming in the Rhythmic Structure of Candombe Drumming Patterns”. In: *Proc. 4th Int. Conf. Anal. Approaches World Music (AAWM)*, New York, USA, Jun. 2016.

[204] FOOTE, J., UCHIHASHI, S. “The Beat Spectrum: A New Approach to Rhythm Analysis”. In: *Proc. 2001 IEEE Int. Conf. Multimedia Expo (ICME)*, pp. 881–884, Tokyo, Japan, Aug. 2001. doi: 10.1109/ICME.2001.1237863.

[205] TZANETAKIS, G., COOK, P. “Musical Genre Classification of Audio Signals”, *IEEE Trans. Speech Audio Process.*, v. 10, n. 5, pp. 293–302, Jul. 2002. doi: 10.1109/TSA.2002.800560.

[206] DIXON, S., GOUYON, F., WIDMER, G. “Towards characterisation of music via rhythmic patterns”. In: *Proc. 5th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 509–517, Barcelona, Spain, Oct. 2004.

[207] GOUYON, F., DIXON, S., PAMPALK, E., et al. "Evaluating Rhythmic Descriptors for Musical Genre Classification". In: *Proc. 25th Audio Eng. Soc. Int. Conf. (AES)*, pp. 196–204, London, UK, Jun. 2004.

[208] PAULUS, J., KLAPURI, A. "Measuring the Similarity of Rhythmic Patterns". In: *Proc. 3rd Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 150–156, Paris, France, Oct. 2002.

[209] PEETERS, G. "Rhythm Classification Using Spectral Rhythm Patterns". In: *Proc. 6th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 644–647, London, UK, Sep. 2005. doi: 10.5281/zenodo.1417495.

[210] HOLZAPFEL, A., STYLIANOU, Y. "Rhythmic Similarity of Music Based on Dynamic Periodicity Warping". In: *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 2217–2220, Las Vegas, USA, Mar. 2008. doi: 10.1109/ICASSP.2008.4518085.

[211] HOLZAPFEL, A., STYLIANOU, Y. "A Scale Transform Based Method for Rhythmic Similarity of Music". In: *Proc. 2009 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 317–320, Taipei, Taiwan, Apr. 2009. doi: 10.1109/ICASSP.2009.4959584.

[212] HOLZAPFEL, A., STYLIANOU, Y. "Scale Transform in Rhythmic Similarity of Music", *IEEE Trans. Audio, Speech, Lang. Process.*, v. 19, n. 1, pp. 176–185, 2011. doi: 10.1109/TASL.2010.2045782.

[213] MARCHAND, U., PEETERS, G. "The Modulation Scale Spectrum and its Application to Rhythm-Content Description". In: *Proc. 17th Int. Conf. Digit. Audio Effects (DAFx)*, pp. 167–172, Erlangen, Germany, Sep. 2014.

[214] MARCHAND, U., PEETERS, G. "Scale and Shift Invariant Time/Frequency Representation Using Auditory Statistics: Application to Rhythm Description". In: *Proc. 2016 IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, pp. 1–6, Vietri Sul Mare, Italy, Sep. 2016. doi: 10.1109/MLSP.2016.7738904.

[215] GRUHNE, M., DITTMAR, C. "Improving Rhythmic Pattern Features Based on Logarithmic Preprocessing". In: *Proc. 126th Audio Eng. Soc. Conv. (AES)*, Munich, Germany, May 2009.

[216] JENSEN, J. H., CHRISTENSEN, M. G., JENSEN, S. H. "A Tempo-Insensitive Representation of Rhythmic Patterns". In: *Proc. 17th Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1509–1512, Glasgow, Scotland, Aug. 2009.

[217] PAMPALK, E., RAUBER, A., MERKL, D. “Content-based Organization and Visualization of Music Archives”. In: *Proc. 10th ACM Int. Multimedia Conf. (MM)*, pp. 570–579, Juan-les-Pins, France, Dec. 2002. doi: 10.1145/641007.641121.

[218] PAMPALK, E. *Computational Models of Music Similarity and their Application in Music Information Retrieval.* PhD Thesis, Vienna University of Technology, Vienna, Austria, 2006.

[219] LIDY, T., RAUBER, A. “Evaluation of Feature Extractors and Psycho-Acoustic Transformations for Music Genre Classification”. In: *Proc. 6th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 34–41, London, United Kingdom, Sep. 2005. doi: 10.5281/zenodo.1416856.

[220] POHLE, T., SCHNITZER, D., SCHEDL, M., et al. “On Rhythm and General Music Similarity”. In: *Proc. 10th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 525–530, Kobe, Japan, Oct. 2009. doi: 10.5281/zenodo.1418229.

[221] ABRASSART, M., DORAS, G. “And What If Two Musical Versions Don’t Share Melody, Harmony, Rhythm, or Lyrics”. In: *Proc. 23rd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 677–684, Bengaluru, India, Dec. 2022.

[222] FOROUGHMAND, H., PEETERS, G. “Deep-Rhythm for Global Tempo Estimation in Music”. In: *Proc. 20th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 636–643, Delft, The Netherlands, Nov. 2019. doi: 10.5281/zenodo.3527890.

[223] BITTNER, R. M., MCFEE, B., SALAMON, J., et al. “Deep Salience Representations for F0 Estimation in Polyphonic Music”. In: *Proc. 18th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 63–70, Suzhou, China, Oct. 2017. doi: 10.5281/zenodo.1417937.

[224] HOLZAPFEL, A., FLEXER, A., WIDMER, G. “Improving Tempo-Sensitive and Tempo-Robust Descriptors for Rhythmic Similarity”. In: *Proc. 8th Sound Music Comput. Conf. (SMC)*, pp. 247–252, Padua, Italy, Jul. 2011.

[225] PANTELI, M., DIXON, S. “On the Evaluation of Rhythmic and Melodic Descriptors for Music Similarity”. In: *Proc. 17th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 468–474, New York, USA, Aug. 2016. doi: 10.5281/zenodo.1417555.

[226] SEYERLEHNER, K., WIDMER, G., POHLE, T. “Fusing Block-Level Features for Music Similarity Estimation”. In: *Proc. 13th Int. Conf. Digit. Audio Effects (DAFx)*, pp. 225–232, Graz, Austria, Sep. 2010.

[227] MCINNES, L., HEALY, J., SAUL, N., et al. “UMAP: Uniform Manifold Approximation and Projection”, *J. Open Source Softw.*, v. 3, n. 29, pp. 861, 2018. doi: 10.21105/joss.00861.

[228] MCINNES, L., HEALY, J., MELVILLE, J. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. 2020. arXiv preprint, arXiv:1802.03426v3.

[229] HAINSWORTH, S. “Beat Tracking and Musical Metre Analysis”. In: Klapuri, A., Davy, M. (Eds.), *Signal Processing Methods for Music Transcription*, Springer, pp. 101–129, New York, USA, 2006.

[230] JIANG, J., CHIN, D., ZHANG, Y., et al. “Learning Hierarchical Metrical Structure Beyond Measures”. In: *Proc. 23rd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 201–209, Bengaluru, India, Dec. 2022.

[231] DURAND, S., BELLO, J. P., DAVID, B., et al. “Downbeat tracking with multiple features and deep neural networks”. In: *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 409–413, Brisbane, Australia, Apr. 2015.

[232] DURAND, S., BELLO, J. P., DAVID, B., et al. “Feature Adapted Convolutional Neural Networks for Downbeat Tracking”. In: *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 296–300, Shanghai, China, Mar. 2016.

[233] DURAND, S., ESSID, S. “Downbeat Detection with Conditional Random Fields and Deep Learned Features”. In: *Proc. 17th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 386–392, New York, USA, Aug. 2016.

[234] DURAND, S., BELLO, J. P., DAVID, B., et al. “Robust Downbeat Tracking Using an Ensemble of Convolutional Networks”, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, v. 25, n. 1, pp. 76–89, Jan. 2017.

[235] MARCHAND, U., PEETERS, G. “Swing Ratio Estimation”. In: *Proc. 18th Int. Conf. Digit. Audio Effects (DAFx)*, pp. 423–428, Trondheim, Norway, Dec. 2015.

[236] HENNIG, H., FLEISCHMANN, R., FREDEBOHM, A., et al. “The Nature and Perception of Fluctuations in Human Musical Rhythms”, *PLOS ONE*, v. 6, n. 10, Oct. 2011. doi: 10.1371/journal.pone.0026457.

[237] STABLES, R., ATHWAL, C., CADE, R. “Drum Pattern Humanisation using a Recursive Bayesian Framework”. In: *Proc. 133rd Audio Eng. Soc. Conv. (AES)*, San Francisco, USA, Oct. 2012.

[238] STABLES, R., ENDO, S., WING, A. “Multi-player Microtiming Humanisation using a Multivariate Markov Model”. In: *Proc. 17th Int. Conf. Digit. Audio Effects (DAFx)*, pp. 109–114, Erlangen, Germany, Sep. 2014.

[239] FUENTES, M., MCFEE, B., CRAYENCOUR, H. C., et al. “Analysis of Common Design Choices in Deep Learning Systems for Downbeat Tracking”. In: *Proc. 19th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 106–112, Paris, France, Sep. 2018.

[240] BÖCK, S., KREBS, F., WIDMER, G. “A Multi-model Approach to Beat Tracking Considering Heterogeneous Music Styles”. In: *Proc. 15th Conf. Int. Society Music Inform. Retr. (ISMIR)*, pp. 603–608, Taipei, Taiwan, Oct. 2014.

[241] DAVIES, M. E. P., BÖCK, S., FUENTES, M. “Tempo, Beat and Downbeat Estimation”. Nov. 2021. Available at: <`https://tempobeatdownbeat.github.io/tutorial/intro.html`>. Accessed: January 31, 2023.

[242] JIA, B., LV, J., LIU, D. “Deep Learning-Based Automatic Downbeat Tracking: A Brief Review”, *Multimedia Systems*, v. 25, pp. 617–638, Mar. 2019.

[243] BÖCK, S., SCHEDL, M. “Enhanced Beat Tracking with Context-Aware Neural Networks”. In: *Proc. 14th Int. Conf. Digit. Audio Effects (DAFx)*, pp. 135–139, Paris, France, Sep. 2011.

[244] KLAPURI, A. P., ERONEN, A. J., ASTOLA, J. T. “Analysis of the Meter of Acoustic Musical Signals”, *IEEE Trans. Audio, Speech, Lang. Process.*, v. 14, n. 1, pp. 342–355, Jan. 2006.

[245] ELLIS, D. P. W. “Beat Tracking by Dynamic Programming”, *J. New Music Res.*, v. 36, n. 1, pp. 51–60, Mar. 2007.

[246] DAVIES, M. E. P., PLUMBLEY, M. “Context-Dependent Beat Tracking of Musical Audio”, *IEEE Trans. Audio, Speech, Lang. Process.*, v. 15, n. 3, pp. 1009–1020, Mar. 2007.

[247] HOLZAPFEL, A., STYLIANOU, Y. "Beat Tracking Using Group Delay Based Onset Detection". In: *Proc. 9th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 653–658, Philadelphia, EUA, Sep. 2008.

[248] GOTO, M., MURAOKA, Y. "Real-Time Beat Tracking for Drumless Audio Signals: Chord Change Detection for Musical Decisions", *Speech Commun.*, v. 27, n. 3-4, pp. 311–335, Apr. 1999.

[249] DIXON, S. "Automatic Extraction of Tempo and Beat from Expressive Performances", *J. New Music Res.*, v. 30, n. 1, pp. 39–58, Mar. 2001.

[250] DIXON, S. "Evaluation of the Audio Beat Tracking System BeatRoot", *J. New Music Res.*, v. 36, n. 1, pp. 39–50, Mar. 2007.

[251] GOUYON, F., HERRERA, P., CANO, P. "Pulse-Dependent Analysis of Percussive Music". In: *Proc. 22nd Audio Eng. Soc. Conf. Virtual, Synthetic, Entertainment Audio (AES)*, Espoo, Finland, Jun. 2002.

[252] GROSCHE, P., MÜLLER, M. "Extracting Predominant Local Pulse Information from Musical Recordings", *IEEE Trans. Audio, Speech, Lang. Process.*, v. 19, n. 6, pp. 1688–1701, Aug. 2011.

[253] ELLIS, D., ARROYO, J. "Eigenrhythms: Drum Pattern Basis Sets for Classification and Generation". In: *Proc. 5th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 554–560, Barcelona, Spain, Oct. 2004.

[254] DAVIES, M. E. P., PLUMBLEY, M. D. "A Spectral Difference Approach to Downbeat Extraction in Musical Audio". In: *Proc. 14th Eur. Signal Process. Conf. (EUSIPCO)*, Florence, Italy, Sep. 2006.

[255] DURAND, S., DAVID, B., RICHARD, G. "Enhancing Downbeat Detection When Facing Different Music Styles". In: *Proc. 2014 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 3132–3136, Florence, Italy, May 2014.

[256] JEHAN, T. "Downbeat Prediction by Listening and Learning". In: *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, pp. 267–270, New Paltz, USA, Oct. 2005.

[257] HAINSWORTH, S., MACLEOD, M. "Beat Tracking with Particle Filtering Algorithms". In: *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, pp. 91–94, New Paltz, USA, Oct. 2003.

[258] SETHARES, W. A., MORRIS, R. D., SETHARES, J. C. "Beat Tracking of Musical Performances Using Low-Level Audio Features", *IEEE Trans. Speech Audio Process.*, v. 13, n. 2, pp. 275–285, Mar. 2005.

[259] WHITELEY, N., CEMGIL, A. T., GODSILL, S. "Bayesian Modelling of Temporal Structure in Musical Audio". In: *Proc. 7th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 29–34, Victoria, Canada, Oct. 2006.

[260] FILLON, T., JODER, C., DURAND, S., et al. "A Conditional Random Field System for Beat Tracking". In: *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 424–428, Brisbane, Australia, Apr. 2015.

[261] LAROCHE, J. "Estimating Tempo, Swing and Beat Locations in Audio Recordings". In: *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, pp. 135–138, New Paltz, USA, Oct. 2001.

[262] PEETERS, G., PAPADOPOULOS, H. "Simultaneous Beat and Downbeat-Tracking Using a Probabilistic Framework: Theory and Large-Scale Evaluation", *IEEE Trans. Audio, Speech, Lang. Process.*, v. 19, n. 6, pp. 1754–1769, Aug. 2011.

[263] SRINIVASAMURTHY, A., HOLZAPFEL, A., CEMGIL, A. T., et al. "Particle Filters for Efficient Meter Tracking with Dynamic Bayesian Networks". In: *Proc. 16th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 197–203, Málaga, Spain, Oct. 2015.

[264] HOLZAPFEL, A., KREBS, F., SRINIVASAMURTHY, A. "Tracking the 'Odd': Meter Inference in a Culturally Diverse Music Corpus". In: *Proc. 15th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 425–430, Taipei, Taiwan, Oct. 2014.

[265] KREBS, F., BÖCK, S., WIDMER, G. "An Efficient State-Space Model for Joint Tempo and Meter Tracking". In: *Proc. 16th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 72–78, Málaga, Spain, Oct. 2015.

[266] SRINIVASAMURTHY, A., HOLZAPFEL, A., CEMGIL, A. T., et al. "A Generalized Bayesian Model for Tracking Long Metrical Cycles in Acoustic Music Signals". In: *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 76–80, Shanghai, China, Mar. 2016.

[267] HEYDARI, M., DUAN, Z. "Don't Look Back: An Online Beat Tracking Method Using RNN and Enhanced Particle Filtering". In: *Proc. 2021*

IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 236–240, Toronto, Canada, Jun. 2021.

[268] HEYDARI, M., CWITKOWITZ, F., DUAN, Z. “Beatnet: CRNN and Particle Filtering for Online Joint Beat, Downbeat and Meter Tracking”. In: *Proc. 22nd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 270–277, Online, Nov. 2021.

[269] HEYDARI, M., MCCALLUM, M., EHMANN, A., et al. “A Novel 1D State Space for Efficient Music Rhythmic Analysis”. In: *Proc. 2022 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 421–425, Singapore, Singapore, May 2022.

[270] KREBS, F., BÖCK, S., DORFER, M., et al. “Downbeat Tracking Using Beat Synchronous Features with Recurrent Neural Networks”. In: *Proc. 17th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 129–135, New York, USA, Aug. 2016.

[271] BÖCK, S., KREBS, F., WIDMER, G. “Joint Beat and Downbeat Tracking with Recurrent Neural Networks”. In: *Proc. 17th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 255–261, New York, USA, Aug. 2016.

[272] HOLZAPFEL, A., GRILL, T. “Bayesian Meter Tracking on Learned Signal Representations”. In: *Proc. 17th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 262–268, New York, USA, Aug. 2016.

[273] BÖCK, S., DAVIES, M. E. P., KNEES, P. “Multi-Task Learning of Tempo and Beat: Learning One to Improve the Other”. In: *Proc. 20th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 486–493, Delft, The Netherlands, Nov. 2019.

[274] DAVIES, M. E. P., BÖCK, S. “Temporal Convolutional Networks for Musical Audio Beat Tracking”. In: *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, A Coruña, Spain, Sep. 2019.

[275] BÖCK, S., DAVIES, M. E. P. “Deconstruct, Analyse, Reconstruct: How to Improve Tempo, Beat, and Downbeat Estimation”. In: *Proc. 21st Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 574–582, Montréal, Canada, Oct. 2020.

[276] CHEN, T.-P., SU, L. “Toward Postprocessing-Free Neural Networks for Joint Beat and Downbeat Estimation”. In: *Proc. 23rd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 27–35, Bengaluru, India, Dec. 2022.

[277] DI GIORGI, B., MAUCH, M., LEVY, M. "Downbeat Tracking with Tempo-Invariant Convolutional Neural Networks". In: *Proc. 21st Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 216–222, Montréal, Canada, Oct. 2020.

[278] HUNG, Y.-N., WANG, J.-C., SONG, X., et al. "Modeling Beats and Downbeats with a Time-Frequency Transformer". In: *Proc. 2022 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 401–405, Singapore, Singapore, May 2022.

[279] ZHAO, J., XIA, G., WANG, Y. "Beat Transformer: Demixed Beat and Downbeat Tracking with Dilated Self-Attention". In: *Proc. 23rd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 169–177, Bengaluru, India, Dec. 2022.

[280] KORZENIOWSKI, F., BÖCK, S., WIDMER, G. "Probabilistic Extraction of Beat Positions from a Beat Activation Function". In: *Proc. 15th Conf. Int. Soc. Music Inform. Retr. (ISMIR)*, pp. 513–518, Taipei, Taiwan, Oct. 2014.

[281] VAN DEN OORD, A., DIELEMAN, S., ZEN, H., et al. "WaveNet: A Generative Model for Raw Audio". 2016. arXiv preprint, arXiv:1609.03499.

[282] BAI, S., KOLTER, J. Z., KOLTUN, V. "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling". 2018. arXiv preprint, arXiv:1803.01271.

[283] FIOCCHI, D., BUCCOLI, M., ZANONI, M., et al. "Beat Tracking using Recurrent Neural Network: a Transfer Learning Approach". In: *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1929–1933, Rome, Italy, Sep. 2018.

[284] PINTO, A. S., DAVIES, M. E. P. "Tapping Along to the Difficult Ones: Leveraging User-Input for Beat Tracking in Highly Expressive Musical Content". In: Kronland-Martinet, R., Ystad, S., Aramaki, M. (Eds.), *Perception, Representations, Image, Sound, Music. CMMR 2019*, v. 12631, *Lecture Notes in Computer Science*, Springer, pp. 75–90, Cham, Switzerland, 2021. doi: 10.1007/978-3-030-70210-6_5.

[285] DAVIES, M. E. P., DEGARA, N., PLUMBLEY, M. D. *Evaluation Metrics for Musical Audio Beat Tracking Algorithms.* Technical Report C4DM-TR-09-06, Centre for Digital Music, Queen Mary University of London, London, UK, 2009.

[286] HAINSWORTH, S. W. *Techniques for the Automated Analysis of Musical Audio.* PhD Thesis, Department of Engineering, Cambridge University, 2003.

[287] HOCKMAN, J. A., DAVIES, M. E. P., FUJINAGA, I. "One in the Jungle: Downbeat Detection in Hardcore, Jungle, and Drum and Bass". In: *Proc. 13th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, p. 169–174, Porto, Portugal, Oct. 2012.

[288] GOUYON, F. *A Computational Approach to Rhythm Description — Audio Features for the Computation of Rhythm Periodicity Functions and their use in Tempo Induction and Music Content Processing.* PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2005.

[289] HOLZAPFEL, A., DAVIES, M. E. P., ZAPATA, J. R., et al. "Selective sampling for beat tracking evaluation", *IEEE Trans. Audio, Speech, Lang. Process.*, v. 20, n. 9, pp. 2539–2548, Nov. 2012.

[290] YAMAMOTO, K. "Human-in-the-Loop Adaptation for Interactive Musical Beat Tracking". In: *Proc. 22nd Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 794–801, Online, Nov. 2021.

[291] MARCHAND, U., FRESNEL, Q., PEETERS, G. "GTZAN-rhythm: Extending the GTZAN test-set with beat, downbeat and swing annotations". In: *16th Int. Soc. Music Inform. Retr. Conf. (ISMIR) Late Breaking and Demo Papers*, Málaga, Spain, Oct. 2015.

[292] GOTO, M., HASHIGUCHI, H., NISHIMURA, T., et al. "RWC Music Database: Popular, Classical, and Jazz Music Databases". In: *Proc. 3rd Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 287–288, Paris, France, Oct. 2002.

[293] GOTO, M. "AIST Annotation for the RWC Music Database". In: *Proc. 7th Int. Conf. Music Inform. Retr. (ISMIR)*, pp. 359–360, Victoria, Canada, Oct. 2006.

[294] FUENTES, M. *Multi-Scale Computational Rhythm Analysis: A Framework for Sections, Downbeats, Beats, and Microtiming.* PhD Thesis, Université Paris-Saclay, Paris, France, 2019.

[295] ANKU, W. "Circles and time: A theory of structural organization of rhythm in African music", *Music Theory Online*, v. 6, n. 1, pp. 1–8, 2000.

[296] GRAEFF, N. “Fundamentos rítmicos africanos para a pesquisa da música Afro-Brasileira: o exemplo do Samba de Roda”, *Música e Cultura*, v. 9, n. 1, pp. 66–87, Oct. 2014.

[297] GERISHER, C. “O Suíngue Baiano: Rhythmic Feeling and Microrhythmic Phenomena in Brazilian Percussion”, *Ethnomusicology*, v. 50, n. 1, pp. 99–119, Oct. 2006.

[298] PINTO, T. O. “Som e Música. Questões de uma Antropologia Sonora”, *Revista de Antropologia*, v. 44, n. 1, pp. 221–286, May 2001. doi: 10.1590/S0034-77012001000100007.

[299] HAUGEN, M. R., DANIELSEN, A. “Effect of Tempo on Relative Note Durations in a Performed Samba Groove”, *J. New Music Res.*, v. 49, n. 4, pp. 349–361, 2020.

[300] GOUYON, F. “Microtiming in ‘Samba de Roda’ — Preliminary Experiments with Polyphonic Audio”. In: *Proc. 11th Brazilian Symp. Comput. Music (SBCM)*, pp. 197–203, São Paulo, Brazil, Sep. 2007.

[301] NAVEDA, L., GOUYON, F., GUEDES, C., et al. “Multidimensional microtiming in samba music”. In: *11th Brazilian Symp. Comput. Music (SBCM)*, pp. 1–12, Recife, Brazil, Sep. 2009.

[302] CANNAM, C., LANDONE, C., SANDLER, M. “Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files”. In: *Proc. ACM Multimedia 2010 Int. Conf. (MM)*, pp. 1467–1468, Firenze, Italy, Oct. 2010.

[303] RAFFEL, C., MCFEE, B., HUMPHREY, E. J., et al. “mir_eval: A Transparent Implementation of Common MIR Metrics”. In: *Proc. 15th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 367–372, Taipei, Taiwan, Oct. 2014.

[304] DAVIES, M. E. P., DEGARA, N., PLUMBLEY, M. D. “Measuring the Performance of Beat Tracking Algorithms Using a Beat Error Histogram”, *IEEE Signal Process. Lett.*, v. 18, n. 3, pp. 157–160, Mar. 2011.

[305] MCFEE, B., RAFFEL, C., LIANG, D., et al. “librosa: Audio and Music Signal Analysis in Python”. In: *Proc. 14th Python Science Conf. (SciPy)*, pp. 18–24, Austin, USA, Jul. 2015.

[306] HOLZAPFEL, A., DAVIES, M. E. P., ZAPATA, J. R., et al. “On the Automatic Identification of Difficult Examples for Beat Tracking: Towards

Building New Evaluation Datasets". In: *Proc. 2012 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 89–92, Kyoto, Japan, Mar. 2012.

[307] ZAPATA, J. R., HOLZAPFEL, A., DAVIES, M. E. P., et al. "Assigning a Confidence Threshold on Automatic Beat Annotation in Large Datasets". In: *Proc. 13th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 157–162, Porto, Portugal, Oct. 2012.

[308] SEUNG, H. S., OPPER, M., SOMPOLINSKY, H. "Query By Committee". In: *Proc. 5th Annu. Workshop Comput. Learn. Theory (COLT)*, pp. 287–294, Pittsburgh, USA, Jul. 1992.

[309] CANNAM, C., SANDLER, M., JEWELL, M. O., et al. "Linked Data and You: Bringing music research software into the Semantic Web", *J. New Music Res.*, v. 39, n. 4, pp. 313–325, 2010.

[310] BROSSIER, P. M. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD Thesis, Department of Electronic Engineering, Queen Mary University of London, 2006.

[311] DEGARA, N., RÚA, E. A., PENA, A., et al. "Reliability-Informed Beat Tracking of Musical Signals", *IEEE Trans. Audio, Speech, Lang. Process.*, v. 20, n. 1, pp. 290–301, Jan. 2012.

[312] OLIVEIRA, J. L., GOUYON, F., MARTINS, L. G., et al. "IBT: A Real-Time Tempo and Beat Tracking System". In: *Proc. 11th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 291–296, Utrecht, The Netherlands, Aug. 2010.

[313] BÖCK, S., KREBS, F., WIDMER, G. "Accurate Tempo Estimation based on Recurrent Neural Networks and Resonating Comb Filters". In: *Proc. 16th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 625–631, Málaga, Spain, Oct. 2015.

[314] ZAPATA, J. R., DAVIES, M. E. P., GÓMEZ, E. "Multi-Feature Beat Tracking", *IEEE/ACM Trans. Audio, Speech, and Language Process.*, v. 22, n. 4, pp. 816–825, Apr. 2014.

[315] MOEHN, F. J. "'The Disc Is Not the Avenue': Schismogenetic Mimesis in Samba Recording". In: Green, P. D., Porcello, T. (Eds.), *Wired for Sound: Engineering and Technologies in Sonic Cultures*, Wesleyan University Press, chap. 3, pp. 47–83, Middletown, USA, 2005.

[316] CUNHA, F. L. "Samba Locations: An Analysis on the Carioca Samba, Identities, and Intangible Heritage (Rio de Janeiro, Brazil)". In: Cunha, F. L., Santos, M., Rabassa, J. (Eds.), *Latin American Heritage*, The Latin America Studies Book Series, Springer International Publishing, chap. 1, pp. 3–20, Cham, Switzerland, 2018.

[317] SARRIA M., G. M., DIAZ, J., ARCE-LOPERA, C. "Analyzing and Extending the Salsa Music Dataset". In: *Proc. XXII Symp. Image, Signal Process., Artif. Vision (STSIVA)*, pp. 1–5, Bucaramanga, Colombia, Apr. 2019.

[318] CANO, E., MORA-ÁNGEL, F., LÓPEZ GIL, G. A., et al. "Sesquialtera in the Colombian Bambuco: Perception and Estimation of Beat and Meter". In: *Proc. 21st Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 409–415, Montreal, Canada, Oct. 2020.

[319] CANO, E., MORA-ÁNGEL, F., LÓPEZ GIL, G. A., et al. "Sesquialtera in the Colombian Bambuco: Perception and Estimation of Beat and Meter – Extended version", *Trans. Int. Soc. Music Inform. Retr.*, v. 4, n. 1, pp. 248–262, Dec. 2021. doi: 10.5334/tismir.118.

[320] FONSECA, J., FUENTES, M., BONINI BARALDI, F., et al. "On the Use of Automatic Onset Detection for the Analysis of Maracatu de Baque Solto". In: Castilho, L. C., Dias, R., Pinho, J. F. (Eds.), *Perspectives on Music, Sound and Musicology: Research, Education and Practice*, Springer, pp. 209–225,Cham, Switzerland, 2021.

[321] KOCH, G., ZEMEL, R., SALAKHUTDINOV, R. "Siamese Neural Networks for One-Shot Image Recognition". In: *Proc. 32nd Int. Conf. Mach. Learn. Deep Learn. Workshop*, Lille, France, Jul. 2015.

[322] VINYALS, O., BLUNDELL, C., LILLICRAP, T., et al. "Matching Networks for One Shot Learning". In: *Adv. Neural Inform. Process. Syst.*, v. 29, pp. 3630—3638, 2016.

[323] SNELL, J., SWERSKY, K., ZEMEL, R. "Prototypical Networks for Few-Shot Learning". In: *Adv. Neural Inform. Process. Syst.*, v. 30, pp. 4077–4087, 2017.

[324] SETTLES, B. *Active Learning Literature Survey.* Computer Sciences Technical Report 1648, University of Wisconsin–Madison, Madison, USA, 2009. Available at: <`https://burrsettles.com/pub/settles.activelearning.pdf`>.

[325] SHUYANG, Z., HEITTOLA, T., VIRTANEN, T. “Active Learning for Sound Event Classification by Clustering Unlabeled Data”. In: *Proc. 2017 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 751–755, New Orleans, USA, Mar. 2017.

[326] WANG, Y., CARTWRIGHT, M., BELLO, J. P. “Active Few-Shot Learning for Sound Event Detection”. In: *Proc. 23rd Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, pp. 1551–1555, Incheon, Korea, Sep. 2022.

[327] KIM, B., PARDO, B. “A Human-in-the-Loop System for Sound Event Detection and Annotation”, *ACM Trans. Interact. Intell. Syst.*, v. 8, n. 2, pp. 1–23, 2018. doi: 10.1145/3214366.

[328] WANG, Y., SALAMON, J., CARTWRIGHT, M., et al. “Few-shot drum transcription in polyphonic music”. In: *Proc. 21st Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 117–124, Montreal, Canada, Oct. 2020.

[329] WANG, Y., STOLLER, D., BITTNER, R. M., et al. “Few-Shot Musical Source Separation”. In: *Proc. 2022 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 121–125, Singapore, Singapore, May 2022.

[330] SARASÚA, Á., LAURIER, C., HERRERA, P. “Support Vector Machine Active Learning for Music Mood Tagging”. In: *Proc. 9th Int. Symp. Comput. Music Model. Retr. (CMMR)*, pp. 518–525, London, UK, Jun. 2012.

[331] SU, H., KASAI, J., WU, C. H., et al. “Selective Annotation Makes Language Models Better Few-Shot Learners”. In: *Proc. 11th Int. Conf. Learn. Representations (ICLR)*, Kigali, Rwanda, May 2023.

[332] GOTO, M. “Development of the RWC Music Database”. In: *Proc. 18th Int. Congr. Acoust. (ICA)*, pp. I–553–556, Kyoto, Japan, Apr. 2004.

[333] BITTNER, R. M., FUENTES, M., RUBINSTEIN, D., et al. “mirdata: Software for Reproducible Usage of Datasets”. In: *Proc. 20th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, pp. 99–106, Delft, The Netherlands, Nov. 2019.

[334] ROCAMORA, M., JURE, L., BISCAINHO, L. W. P. “Tools for detection and classification of piano drum patterns from Candombe recordings”. In: *Proc. 9th Conf. Interdisciplinary Musicology (CIM)*, pp. 382–387, Berlin, Germany, Dec. 2014.

[335] ROCAMORA, M., JURE, L. “carat: Computer-Aided Rhythmic Analysis Toolbox.” In: *Proc. 1st Anal. Approaches World Music Spec. Topics Symp. (AAWM)*, Birmingham, UK, Jul. 2019.

[336] LIN, H., BILMES, J. A. “How to Select a Good Training-Data Subset for Transcription: Submodular Active Selection for Sequences”. In: *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, pp. 2859–2862, Brighton, United Kingdom, Sep. 2009.

[337] ARTHUR, D., VASSILVITSKII, S. “k-means++: The Advantages of Careful Seeding”. In: *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, p. 1027–1035, New Orleans, USA, Jan. 2006.

[338] LAFFERTY, J., MCCALLUM, A., PEREIRA, F. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: *Proc. 18th Int. Conf. Machine Learn. (ICML)*, pp. 282–289, Williamstown, USA, Jun. 2001.

[339] SUTTON, C., MCCALLUM, A. “An Introduction to Conditional Random Fields for Relational Learning”. In: Getoor, L., Taskar, B. (Eds.), *Introduction to Statistical Relational Learning*, MIT Press, chap. 4, pp. 93–128, Cambridge, USA, 2006.

[340] DOUCET, A., JOHANSEN, A. M. “A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later”. In: Crisan, D., Rozovskii, B. (Eds.), *The Oxford Handbook of Nonlinear Filtering*, Oxford University Press, chap. 8.2, pp. 656–704, New York, USA, 2011.

[341] KITAGAWA, G. “Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models”, *J. Comput. Graph. Statist.*, v. 5, n. 1, pp. 1–25, Mar. 1996.

[342] JOHANSEN, A. M., DOUCET, A. “A Note on Auxiliary Particle Filters”, *Statist. Probab. Lett.*, v. 78, n. 12, pp. 1498–1504, Sep. 2008.

[343] MARTINO, L., ELVIRA, V., LOUZADA, F. “Effective Sample Size for Importance Sampling Based on Discrepancy Measures”, *Signal Process.*, v. 131, pp. 386–401, Feb. 2017. doi: 10.1016/j.sigpro.2016.08.025.

[344] DAVIES, M., MADISON, G., SILVA, P., et al. “The Effect of Microtiming Deviations on the Perception of Groove in Short Rhythms”, *Music Perception*, v. 30, n. 5, pp. 497–510, Jun. 2013.

[345] LONGO, L., BRCIC, M., CABITZA, F., et al. "Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions", *Inform. Fusion*, v. 106, 2024.

[346] JURE, L., ROCAMORA, M. "Subir la Llamada: Negotiating Tempo and Dynamics in Uruguayan Candombe Drumming". In: *Proc. 8th Int. Workshop Folk Music Anal. (FMA)*, Thessaloniki, Greece, Jun. 2018.

[347] CLAYTON, M., TARSITANI, S., JANKOWSKY, R., et al. "The Interpersonal Entrainment in Music Performance Data Collection", *Empirical Musicology Rev.*, v. 16, n. 1, pp. 65–84, Dec. 2021.

[348] KARTOMI, M. J. *On Concepts and Classifications of Musical Instruments.* Chicago, USA, University of Chicago Press, 1990.

[349] ROWELL, L. E. *Music and Musical Thought in Early India.* Chicago, USA, University of Chicago Press, 1992.

[350] HORNBOSTEL, E. M., SACHS, C. "Classification of Musical Instruments: Translated from the Original German by Antony Baines and Klaus P. Wachsmann", *Galpin Soc. J.*, v. 14, pp. 3–29, Mar. 1961.

[351] MAHILLON, V. *Catalogue Descriptif et Analytique du Musée Instrumental du Conservatoire Royal de Bruxelles: Précédé d'un Essai de Classification Méthodique de Tous les Instruments Anciens et Modernes.* Ghent, Belgium, C. Annoot-Braeckman, 1880.

[352] SACHS, C. *Reallexikon der Musikinstrumente.* Berlin, Germany, Julius Bard, 1913.

[353] LEE, D. "Hornbostel-Sachs Classification of Musical Instruments", *Knowledge Organization*, v. 47, n. 1, pp. 72–91, 2020.

Appendix A

BRID Tables

Tables A.1, A.2, A.3, and A.4 display the full content of the BRID dataset, i.e., the contents of acoustic mixture and solo (musicians #1, #2, and #3) tracks, respectively. We present here the `[GID#]`, filename style and duration (in seconds) for all tracks in the dataset. Instruments in each track are shown with their particular variation, i.e., "Pandeiro 2" is the 10" leather-head *pandeiro*.[1] Please refer to Table 3.1 for more detail on each variation.

[1] "Surdo 1", "Surdo 2", and "Surdo 3" are not to be confused with *surdos de primeira*, *de segunda*, and *de terceira* respectively. For solo tracks in the *samba-enredo* style, "Surdo 2" and "Surdo 3" were tuned differently and performed *de primeira*, *de segunda*, and *de terceira* patterns.

Table A.1: Overview of the acoustic mixture tracks in BRID.

GID#	Filename	#inst.	Musician #2	Musician #4	Musician #5	Musician #6	Style	Duration
0001	M4-01-SA	4	Pandeiro 2	Tantã 3	Surdo 2	Tamborim 2	Samba	0:00:36
0002	M4-02-SA	4	Pandeiro 3	Tantã 3	Surdo 2	Tamborim 2	Samba	0:00:34
0003	M4-03-PA	4	Pandeiro 2	Tantã 3	Tantã 4	Tamborim 2	Partido-alto	0:00:41
0004	M4-04-PA	4	Pandeiro 3	Tantã 3	Tantã 4	Tamborim 2	Partido-alto	0:00:37
0005	M4-05-PA	4	Pandeiro 3	Chocalho 3	Tantã 4	Tamborim 2	Partido-alto	0:00:42
0006	M4-06-SA	4	Pandeiro 5	Tantã 3	Surdo 2	Chocalho 3	Samba	0:00:45
0007	M4-07-SA	4	Repique 1	Pandeiro 2	Surdo 2	Reco-reco 2	Samba	0:00:31
0008	M4-08-SA	4	Repique 3	Pandeiro 2	Surdo 2	Reco-reco 2	Samba	0:00:29
0009	M4-09-SA	4	Tantã 3	Pandeiro 2	Agogô 1	Surdo 2	Samba	0:00:28
0010	M4-10-SE	4	Caixa 2	Surdo 2	Pandeiro 3	Reco-reco 2	Samba-enredo	0:00:28
0011	M4-11-SE	4	Caixa 2	Tamborim 3	Agogô 1	Tantã 4	Samba-enredo	0:00:28
0012	M4-12-SE	4	Cuíca 3	Caixa 2	Repique 2	Surdo 2	Samba-enredo	0:00:28
0013	M4-13-SE	4	Cuíca 3	Caixa 2	Tamborim 3	Surdo 2	Samba-enredo	0:00:27
0014	M4-14-SE	4	Repique 2	Cuíca 3	Tamborim 3	Surdo 2	Samba-enredo	0:00:19
0015	M4-15-SA	4	Cuíca 3	Agogô 1	Tantã 4	Chocalho 2	Samba	0:00:34
0016	M4-16-PA	4	Pandeiro 2	Tantã 3	Pandeiro 3	Surdo 2	Partido-alto	0:00:32
0017	M4-17-PA	4	Repique 1	Pandeiro 2	Pandeiro 3	Tantã 4	Partido-alto	0:00:23
0018	M4-18-SA	4	Repique 1	Tantã 3	Pandeiro 3	Chocalho 3	Samba	0:00:33
0019	M4-19-SA	4	Cuíca 3	Agogô 1	Repique 3	Chocalho 3	Samba	0:00:30

(Continued on the following page.)

Table A.1: (Continued from previous page.)

GID#	Filename	#inst.	Musician #2	Musician #4	Musician #5	Musician #6	Style	Duration
0020	M3-01-PA	3	Pandeiro 5	Pandeiro 2	Tantã 4		Partido-alto	0:00:19
0021	M3-02-SA	3	Tamborim 2	Pandeiro 2	Tantã 3		Samba	0:00:29
0022	M3-03-SA	3	Pandeiro 5	Tamborim 2	Surdo 2		Samba	0:00:33
0023	M3-04-SA	3	Pandeiro 2	Tamborim 2	Surdo 2		Samba	0:00:26
0024	M3-05-PA	3	Pandeiro 5	Chocalho 3	Tantã 3		Partido-alto	0:00:29
0025	M3-06-PA	3	Repique 2	Agogô 1	Caixa 2		Partido-alto	0:00:24
0026	M3-07-PA	3	Repique 2	Cuíca 3	Caixa 2		Partido-alto	0:00:27
0027	M3-08-SE	3	Repique 2	Tamborim 3	Caixa 2		Samba-enredo	0:00:18
0028	M3-09-MA	3	Caixa 2	Pandeiro 2	Surdo 2		Marcha	0:00:18
0029	M3-10-PA	3	Cuíca 3	Reco-reco 2	Tantã 4		Partido-alto	0:00:29
0030	M3-11-PA	3	Repique 1	Agogô 1	Surdo 2		Partido-alto	0:00:32
0031	M3-12-SA	3	Repique 1	Pandeiro 2	Tantã 4		Samba	0:00:26
0032	M3-13-SA	3	Repique 3	Tamborim 2	Reco-reco 2		Samba	0:00:27
0033	M3-14-SA	3	Repique 3	Chocalho 3	Agogô 1		Samba	0:00:29
0034	M3-15-SA	3	Surdo 2	Pandeiro 1	Reco-reco 2		Samba	0:00:24
0035	M3-16-PA	3	Caixa 2	Reco-reco 2	Tantã 4		Partido-alto	0:00:23
0036	M3-17-PA	3	Repique 1	Tamborim 2	Tantã 4		Partido-alto	0:00:18
0037	M3-18-SA	3	Agogô 1	Chocalho 3	Surdo 2		Samba	0:00:30
0038	M3-19-SA	3	Pandeiro 5	Chocalho 3	Tantã 4		Samba	0:00:25
0039	M3-20-SE	3	Caixa 2	Tamborim 3	Tantã 3		Samba-enredo	0:00:24

(Continued on the following page.)

Table A.1: (Continued from previous page.)

GID#	Filename	#inst.	Musician #2	Musician #4	Musician #5	Musician #6	Style	Duration
0040	M3-21-PA	3	Pandeiro 3	Tamborim 2	Repique 3		Partido-alto	0:00:31
0041	M3-22-SE	3	Pandeiro 2	Surdo 1	Surdo 2		Samba-enredo	0:00:23
0042	M3-23-SE	3	Pandeiro 3	Surdo 1	Surdo 2		Samba-enredo	0:00:27
0043	M3-24-SE	3	Reco-reco 2	Surdo 1	Surdo 2		Samba-enredo	0:00:24
0044	M3-25-SE	3	Agogô 1	Surdo 1	Surdo 2		Samba-enredo	0:00:22
0045	M3-26-SE	3	Cuíca 3	Surdo 1	Surdo 2		Samba-enredo	0:00:22
0046	M3-27-SE	3	Tamborim 3	Surdo 1	Surdo 2		Samba-enredo	0:00:15
0047	M3-28-SE	3	Caixa 2	Surdo 1	Surdo 2		Samba-enredo	0:00:21
0048	M3-29-SE	3	Repique 2	Surdo 1	Surdo 2		Samba-enredo	0:00:27
0049	M2-01-PA	2	Pandeiro 2		Pandeiro 3		Partido-alto	0:00:25
0050	M2-02-SA	2	Tantã 3		Tantã 4		Samba	0:00:30
0051	M2-03-SA	2	Tantã 3	Surdo 2			Samba	0:00:27
0052	M2-04-SA	2	Pandeiro 2	Surdo 2			Samba	0:00:26
0053	M2-05-SA	2	Pandeiro 3	Surdo 2			Samba	0:00:34
0054	M2-06-PA	2	Pandeiro 3	Tantã 3			Partido-alto	0:00:29
0055	M2-07-SA	2	Pandeiro 3	Tantã 4			Samba	0:00:29
0056	M2-08-MA	2	Pandeiro 2	Tantã 4			Marcha	0:00:22
0057	M2-09-SA	2	Pandeiro 3	Reco-reco 2			Samba	0:00:31
0058	M2-10-SA	2	Pandeiro 2	Reco-reco 2			Samba	0:00:36
0059	M2-11-SA	2	Reco-reco 2	Tantã 3			Samba	0:00:21

(Continued on the following page.)

Table A.1: (Continued from previous page.)

GID#	Filename	#inst.	Musician #2	Musician #4	Musician #5	Musician #6	Style	Duration
0060	M2-12-SA	2	Reco-reco 2	Surdo 2			Samba	0:00:29
0061	M2-13-SE	2	Caixa 2	Agogô 1			Samba-enredo	0:00:26
0062	M2-14-MA	2	Caixa 2	Surdo 2			Marcha	0:00:22
0063	M2-15-PA	2	Caixa 2	Pandeiro 3			Partido-alto	0:00:25
0064	M2-16-PA	2	Cuíca 3	Repique 1			Partido-alto	0:00:31
0065	M2-17-PA	2	Tamborim 2	Repique 1			Partido-alto	0:00:30
0066	M2-18-PA	2		Chocalho 2	Caixa 2		Partido-alto	0:00:23
0067	M2-19-SA	2		Cuíca 1	Agogô 1		Samba	0:00:25
0068	M2-20-SA	2		Tamborim 2	Agogô 1		Samba	0:00:25
0069	M2-21-SA	2		Tamborim 2	Pandeiro 2		Samba	0:00:26
0070	M2-22-SA	2		Tamborim 2	Tantã 4		Samba	0:00:28
0071	M2-23-PA	2		Tamborim 2	Pandeiro 3		Partido-alto	0:00:33
0072	M2-24-SA	2		Cuíca 1	Tantã 4		Samba	0:00:35
0073	M2-25-SA	2		Chocalho 2	Surdo 2		Samba	0:00:35
0074	M2-26-PA	2		Repique 3	Reco-reco 2		Partido-alto	0:00:29
0075	M2-27-PA	2		Repique 3	Tamborim 2		Partido-alto	0:00:31
0076	M2-28-SE	2		Tantã 3	Repique 2		Samba-enredo	0:00:29
0077	M2-29-SE	2		Surdo 2	Repique 2		Samba-enredo	0:00:32
0078	M2-30-PA	2		Agogô 1	Repique 2		Partido-alto	0:00:30
0079	M2-31-PA	2		Reco-reco 2	Repique 2		Partido-alto	0:00:30

(Continued on the following page.)

Table A.1: (Continued from previous page.)

GID#	Filename	#inst.	Musician #2	Musician #4	Musician #5	Musician #6	Style	Duration
0080	M2-32-PA	2		Reco-reco 4	Repique 2		Partido-alto	0:00:27
0081	M2-33-SA	2		Repique 1	Repique 3		Samba	0:00:31
0082	M2-34-SA	2		Repique 1	Repique 2		Samba	0:00:27
0083	M2-35-SE	2	Repique 2		Repique 3		Samba-enredo	0:00:37
0084	M2-36-PA	2	Repique 1		Tantã 4		Partido-alto	0:00:34
0085	M2-37-SE	2	Repique 2		Tantã 4		Samba-enredo	0:00:32
0086	M2-38-SA	2	Repique 3		Tantã 4		Samba	0:00:35
0087	M2-39-SE	2	Caixa 2		Pandeiro 2		Samba-enredo	0:00:27
0088	M2-40-SA	2	Chocalho 2		Pandeiro 2		Samba	0:00:34
0089	M2-41-SA	2	Chocalho 2		Pandeiro 5		Samba	0:00:27
0090	M2-42-SA	2	Agogô 1		Tantã 3		Samba	0:00:31
0091	M2-43-SA	2	Tamborim 2		Surdo 2		Samba	0:00:40
0092	M2-44-SA	2	Reco-reco 4		Surdo 2		Samba	0:00:34
0093	M2-45-PA	2	Pandeiro 2		Tantã 3		Partido-alto	0:00:35

Table A.2: Overview of the solo-instrument tracks in BRID. Musician #1.

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0094	S1-PD1-01-SA	Pandeiro 1	Samba	80	0:00:26
0095	S1-PD1-02-PA	Pandeiro 1	Partido-alto	100	0:00:31
0096	S1-PD1-03-SE	Pandeiro 1	Samba-enredo	130	0:00:30
0097	S1-PD1-04-MA	Pandeiro 1	Marcha	120	0:00:32
0098	S1-PD1-05-CA	Pandeiro 1	Capoeira	65	0:00:33
0099	S1-PD1-06-VPA	Pandeiro 1	Virada (Partido-alto)	100	0:00:30
0100	S1-PD1-07-VMA	Pandeiro 1	Virada (Marcha)	120	0:00:32
0101	S1-PD2-01-SA	Pandeiro 2	Samba	80	0:00:26
0102	S1-PD2-02-PA	Pandeiro 2	Partido-alto	100	0:00:31
0103	S1-PD2-03-SE	Pandeiro 2	Samba-enredo	130	0:00:30
0104	S1-PD2-04-MA	Pandeiro 2	Marcha	120	0:00:32
0105	S1-PD2-05-CA	Pandeiro 2	Capoeira	65	0:00:33
0106	S1-PD2-06-VPA	Pandeiro 2	Virada (Partido-alto)	100	0:00:28
0107	S1-PD2-07-VMA	Pandeiro 2	Virada (Marcha)	120	0:00:24
0108	S1-PD3-01-SA	Pandeiro 3	Samba	80	0:00:26
0109	S1-PD3-02-PA	Pandeiro 3	Partido-alto	100	0:00:31
0110	S1-PD3-03-SE	Pandeiro 3	Samba-enredo	130	0:00:30
0111	S1-PD3-04-MA	Pandeiro 3	Marcha	120	0:00:32
0112	S1-PD3-05-CA	Pandeiro 3	Capoeira	65	0:00:33
0113	S1-PD3-06-VPA	Pandeiro 3	Virada (Partido-alto)	100	0:00:33

(Continued on the following page.)

Table A.2: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0114	S1-PD3-07-VSA	Pandeiro 3	Virada (Samba)	75	0:00:29
0115	S1-PD4-01-SA	Pandeiro 4	Samba	80	0:00:26
0116	S1-PD4-02-PA	Pandeiro 4	Partido-alto	100	0:00:31
0117	S1-PD4-03-SE	Pandeiro 4	Samba-enredo	130	0:00:30
0118	S1-PD4-04-MA	Pandeiro 4	Marcha	120	0:00:32
0119	S1-PD4-05-CA	Pandeiro 4	Capoeira	65	0:00:33
0120	S1-PD4-06-VPA	Pandeiro 4	Virada (Partido-alto)	100	0:00:27
0121	S1-PD4-07-VSA	Pandeiro 4	Virada (Samba)	75	0:00:26
0122	S1-PD5-01-SA	Pandeiro 5	Samba	80	0:00:26
0123	S1-PD5-02-PA	Pandeiro 5	Partido-alto	100	0:00:31
0124	S1-PD5-03-SE	Pandeiro 5	Samba-enredo	130	0:00:30
0125	S1-PD5-04-MA	Pandeiro 5	Marcha	120	0:00:32
0126	S1-PD5-05-CA	Pandeiro 5	Capoeira	65	0:00:37
0127	S1-PD5-06-VPA	Pandeiro 5	Virada (Partido-alto)	100	0:00:26
0128	S1-PD5-07-VPA	Pandeiro 5	Virada (Partido-alto)	75	0:00:32
0129	S1-TB2-01-SA	Tamborim 2	Samba	80	0:00:26
0130	S1-TB2-02-PA	Tamborim 2	Partido-alto	100	0:00:31
0131	S1-TB2-03-SE	Tamborim 2	Samba-enredo	130	0:00:30
0132	S1-TB3-01-SE	Tamborim 3	Samba-enredo	130	0:00:30
0133	S1-TB3-02-VSE	Tamborim 3	Virada (Samba-enredo)	130	0:00:24

(Continued on the following page.)

Table A.2: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0134	S1-TB1-01-SA	Tamborim 1	Samba	80	0:00:26
0135	S1-TB1-02-PA	Tamborim 1	Partido-alto	100	0:00:31
0136	S1-TB1-03-SE	Tamborim 1	Samba-enredo	130	0:00:30
0137	S1-RR2-01-SA	Reco-reco 2	Samba	80	0:00:26
0138	S1-RR2-02-PA	Reco-reco 2	Partido-alto	100	0:00:31
0139	S1-RR2-03-SE	Reco-reco 2	Samba-enredo	130	0:00:30
0140	S1-RR2-04-VPA	Reco-reco 2	Virada (Partido-alto)	100	0:00:31
0141	S1-RR4-01-SA	Reco-reco 4	Samba	80	0:00:26
0142	S1-RR4-02-PA	Reco-reco 4	Partido-alto	100	0:00:31
0143	S1-RR4-03-SE	Reco-reco 4	Samba-enredo	130	0:00:30
0144	S1-RR4-04-VPA	Reco-reco 4	Virada (Partido-alto)	100	0:00:31
0145	S1-CX1-01-SA	Caixa 1	Samba	80	0:00:26
0146	S1-CX1-02-PA	Caixa 1	Partido-alto	100	0:00:31
0147	S1-CX1-03-SE	Caixa 1	Samba-enredo	130	0:00:30
0148	S1-CX1-04-MA	Caixa 1	Marcha	120	0:00:36
0149	S1-CX1-05-VSE	Caixa 1	Virada (Samba-enredo)	130	0:00:26
0150	S1-RP2-01-SA	Repique 2	Samba	80	0:00:26
0151	S1-RP2-02-PA	Repique 2	Partido-alto	100	0:00:31
0152	S1-RP2-03-SE	Repique 2	Samba-enredo	130	0:00:30
0153	S1-RP2-04-VSE	Repique 2	Virada (Samba-enredo)	130	0:00:26

(Continued on the following page.)

Table A.2: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0154	S1-RP1-01-SA	Repique 1	Samba	80	0:00:26
0155	S1-RP1-02-PA	Repique 1	Partido-alto	100	0:00:30
0156	S1-RP1-03-SE	Repique 1	Samba-enredo	130	0:00:30
0157	S1-RP1-04-VPA	Repique 1	Virada (Partido-alto)	100	0:00:33
0158	S1-RP3-01-SA	Repique 3	Samba	80	0:00:26
0159	S1-RP3-02-PA	Repique 3	Partido-alto	100	0:00:30
0160	S1-RP3-03-SE	Repique 3	Samba-enredo	130	0:00:30
0161	S1-RP3-04-VSA	Repique 3	Virada (Samba)	80	0:00:36
0162	S1-CU2-01-SA	Cuíca 2	Samba	80	0:00:27
0163	S1-CU2-02-PA	Cuíca 2	Partido-alto	100	0:00:29
0164	S1-CU2-03-SE	Cuíca 2	Samba-enredo	130	0:00:31
0165	S1-CU2-04-VSE	Cuíca 2	Virada (Samba-enredo)	130	0:00:24
0166	S1-AG1-01-SA	Agogô 1	Samba	80	0:00:27
0167	S1-AG1-02-PA	Agogô 1	Partido-alto	100	0:00:36
0168	S1-AG1-03-SE	Agogô 1	Samba-enredo	130	0:00:31
0169	S1-AG1-04-VPA	Agogô 1	Virada (Partido-alto)	100	0:00:33
0170	S1-SK1-01-SA	Chocalho 1	Samba	80	0:00:26
0171	S1-SK1-02-PA	Chocalho 1	Partido-alto	100	0:00:37
0172	S1-SK1-03-SE	Chocalho 1	Samba-enredo	130	0:00:31
0173	S1-SK2-01-SA	Chocalho 2	Samba	80	0:00:26

(Continued on the following page.)

Table A.2: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0174	S1-SK2-02-PA	Chocalho 2	Partido-alto	100	0:00:37
0175	S1-SK2-03-SE	Chocalho 2	Samba-enredo	130	0:00:31
0176	S1-TT2-01-SA	Tantã 2	Samba	80	0:00:26
0177	S1-TT2-02-PA	Tantã 2	Partido-alto	100	0:00:37
0178	S1-TT2-03-SE	Tantã 2	Samba-enredo	130	0:00:29
0179	S1-TT2-04-VPA	Tantã 2	Virada (Partido-alto)	100	0:00:22
0180	S1-SU2-01-SA	Surdo 2	Samba	80	0:00:27
0181	S1-SU2-02-PA	Surdo 2	Partido-alto	100	0:00:37
0182	S1-SU2-03-SE	Surdo 2	Samba-enredo	130	0:00:32
0183	S1-SU2-04-VPA	Surdo 2	Virada (Partido-alto)	100	0:00:34
0184	S1-SU2-05-SE	Surdo 2	Samba-enredo	130	0:00:32

Table A.3: Overview of the solo-instrument tracks in BRID. Musician #2.

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0185	S2-PD2-01-SA	Pandeiro 2	Samba	80	0:00:25
0186	S2-PD2-02-PA	Pandeiro 2	Partido-alto	100	0:00:28
0187	S2-PD2-03-SE	Pandeiro 2	Samba-enredo	130	0:00:25
0188	S2-PD2-04-MA	Pandeiro 2	Marcha	120	0:00:28
0189	S2-PD2-05-CA	Pandeiro 2	Capoeira	65	0:00:32
0190	S2-PD2-06-VPA	Pandeiro 2	Virada (Partido-alto)	100	0:00:27
0191	S2-PD2-07-VMA	Pandeiro 2	Virada (Marcha)	120	0:00:37
0192	S2-PD3-01-SA	Pandeiro 3	Samba	80	0:00:25
0193	S2-PD3-02-PA	Pandeiro 3	Partido-alto	100	0:00:28
0194	S2-PD3-03-SE	Pandeiro 3	Samba-enredo	130	0:00:25
0195	S2-PD3-04-MA	Pandeiro 3	Marcha	120	0:00:28
0196	S2-PD3-05-CA	Pandeiro 3	Capoeira	65	0:00:32
0197	S2-PD3-06-VPA	Pandeiro 3	Virada (Partido-alto)	100	0:00:27
0198	S2-PD3-07-VMA	Pandeiro 3	Virada (Marcha)	120	0:00:37
0199	S2-PD3-08-VSE	Pandeiro 3	Virada (Samba-enredo)	130	0:00:46
0200	S2-PD5-01-SA	Pandeiro 5	Samba	80	0:00:25
0201	S2-PD5-02-PA	Pandeiro 5	Partido-alto	100	0:00:28
0202	S2-PD5-03-SE	Pandeiro 5	Samba-enredo	130	0:00:25
0203	S2-PD5-04-MA	Pandeiro 5	Marcha	120	0:00:28
0204	S2-PD5-05-CA	Pandeiro 5	Capoeira	65	0:00:32

(Continued on the following page.)

Table A.3: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0205	S2-PD5-06-PA	Pandeiro 5	Partido-alto	100	0:00:27
0206	S2-PD5-07-VMA	Pandeiro 5	Virada (Marcha)	120	0:00:37
0207	S2-PD6-01-SA	Pandeiro 6	Samba	80	0:00:25
0208	S2-PD6-02-PA	Pandeiro 6	Partido-alto	100	0:00:28
0209	S2-PD6-03-SE	Pandeiro 6	Samba-enredo	130	0:00:25
0210	S2-PD6-04-MA	Pandeiro 6	Marcha	120	0:00:28
0211	S2-PD6-05-CA	Pandeiro 6	Capoeira	65	0:00:32
0212	S2-PD6-06-VPA	Pandeiro 6	Virada (Partido-alto)	100	0:00:27
0213	S2-PD6-07-VMA	Pandeiro 6	Virada (Marcha)	120	0:00:37
0214	S2-TB2-01-SA	Tamborim 2	Samba	80	0:00:25
0215	S2-TB2-02-PA	Tamborim 2	Partido-alto	100	0:00:29
0216	S2-TB2-03-SE	Tamborim 2	Samba-enredo	130	0:00:28
0217	S2-TB2-04-VPA	Tamborim 2	Virada (Partido-alto)	100	0:00:29
0218	S2-TB3-01-SE	Tamborim 3	Samba-enredo	130	0:00:28
0219	S2-TB3-02-VSE	Tamborim 3	Virada (Samba-enredo)	130	0:00:26
0220	S2-RR3-01-SA	Reco-reco 3	Samba	80	0:00:25
0221	S2-RR3-02-PA	Reco-reco 3	Partido-alto	100	0:00:29
0222	S2-RR3-03-SE	Reco-reco 3	Samba-enredo	130	0:00:26
0223	S2-RR3-04-VPA	Reco-reco 3	Virada (Partido-alto)	100	0:00:29
0224	S2-RR4-01-SA	Reco-reco 4	Samba	80	0:00:25

(Continued on the following page.)

Table A.3: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0225	S2-RR4-02-PA	Reco-reco 4	Partido-alto	100	0:00:29
0226	S2-RR4-03-SE	Reco-reco 4	Samba-enredo	130	0:00:26
0227	S2-RR4-04-VPA	Reco-reco 4	Virada (Partido-alto)	100	0:00:29
0228	S2-CX2-01-SA	Caixa 2	Samba	80	0:00:25
0229	S2-CX2-02-PA	Caixa 2	Partido-alto	100	0:00:29
0230	S2-CX2-03-SE	Caixa 2	Samba-enredo	130	0:00:26
0231	S2-CX2-04-MA	Caixa 2	Marcha	120	0:00:28
0232	S2-CX2-05-VSE	Caixa 2	Virada (Samba-enredo)	130	0:00:26
0233	S2-RP2-01-SA	Repique 2	Samba	80	0:00:25
0234	S2-RP2-02-PA	Repique 2	Partido-alto	100	0:00:28
0235	S2-RP2-03-SE	Repique 2	Samba-enredo	130	0:00:25
0236	S2-RP2-04-VSE	Repique 2	Virada (Samba-enredo)	130	0:00:33
0237	S2-RP1-01-SA	Repique 1	Samba	80	0:00:25
0238	S2-RP1-02-PA	Repique 1	Partido-alto	100	0:00:29
0239	S2-RP1-03-SE	Repique 1	Samba-enredo	130	0:00:25
0240	S2-RP1-04-VPA	Repique 1	Virada (Partido-alto)	100	0:00:30
0241	S2-RP3-01-SA	Repique 3	Samba	80	0:00:25
0242	S2-RP3-02-PA	Repique 3	Partido-alto	100	0:00:29
0243	S2-RP3-03-SE	Repique 3	Samba-enredo	130	0:00:26
0244	S2-RP3-04-VPA	Repique 3	Virada (Partido-alto)	100	0:00:32

(Continued on the following page.)

Table A.3: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0245	S2-CU3-01-SA	Cuíca 3	Samba	80	0:00:29
0246	S2-CU3-02-PA	Cuíca 3	Partido-alto	100	0:00:29
0247	S2-CU3-03-SE	Cuíca 3	Samba-enredo	130	0:00:28
0248	S2-CU3-04-VPA	Cuíca 3	Virada (Partido-alto)	100	0:00:37
0249	S2-AG1-01-SA	Agogô 1	Samba	80	0:00:26
0250	S2-AG1-02-PA	Agogô 1	Partido-alto	100	0:00:29
0251	S2-AG1-03-SE	Agogô 1	Samba-enredo	130	0:00:26
0252	S2-AG1-04-VPA	Agogô 1	Virada (Partido-alto)	100	0:00:32
0253	S2-SK3-01-SA	Chocalho 3	Samba	80	0:00:26
0254	S2-SK3-02-PA	Chocalho 3	Partido-alto	100	0:00:29
0255	S2-SK3-03-SE	Chocalho 3	Samba-enredo	130	0:00:25
0256	S2-SK3-04-MA	Chocalho 3	Marcha	120	0:00:28
0257	S2-SK2-01-SA	Chocalho 2	Samba	80	0:00:25
0258	S2-SK2-02-PA	Chocalho 2	Partido-alto	100	0:00:29
0259	S2-SK2-03-SE	Chocalho 2	Samba-enredo	130	0:00:26
0260	S2-SK2-04-MA	Chocalho 2	Marcha	120	0:00:28
0261	S2-TT3-01-SA	Tantã 3	Samba	80	0:00:27
0262	S2-TT3-02-PA	Tantã 3	Partido-alto	100	0:00:29
0263	S2-TT3-03-SE	Tantã 3	Samba-enredo	130	0:00:25
0264	S2-TT3-04-VMA	Tantã 3	Virada (Marcha)	120	0:00:37

(Continued on the following page.)

Table A.3: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0265	S2-TT4-01-SA	Tantã 4	Samba	80	0:00:25
0266	S2-TT4-02-PA	Tantã 4	Partido-alto	100	0:00:29
0267	S2-TT4-03-SE	Tantã 4	Samba-enredo	130	0:00:25
0268	S2-TT4-04-MA	Tantã 4	Marcha	120	0:00:28
0269	S2-TT4-05-VMA	Tantã 4	Virada (Marcha)	120	0:00:37
0270	S2-SU2-01-SA	Surdo 2	Samba	80	0:00:26
0271	S2-SU2-02-PA	Surdo 2	Partido-alto	100	0:00:29
0272	S2-SU2-03-MA	Surdo 2	Marcha	120	0:00:28
0273	S2-SU2-04-VPA	Surdo 2	Virada (Partido-alto)	100	0:00:28
0274	S2-SU3-01-SA	Surdo 3	Samba	80	0:00:25
0275	S2-SU3-02-PA	Surdo 3	Partido-alto	100	0:00:29
0276	S2-SU3-03-SE	Surdo 3	Samba-enredo	130	0:00:26
0277	S2-SU3-04-MA	Surdo 3	Marcha	120	0:00:28
0278	S2-SU3-05-VPA	Surdo 3	Virada (Partido-alto)	100	0:00:30
0279	S2-SU3-06-SE	Surdo 3	Samba-enredo	130	0:00:30
0280	S2-SU2-05-SE	Surdo 2	Samba-enredo	130	0:00:28

Table A.4: Overview of the solo-instrument tracks in BRID. Musician #3.

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0281	S3-PD2-01-SA	Pandeiro 2	Samba	80	0:00:26
0282	S3-PD2-02-PA	Pandeiro 2	Partido-alto	100	0:00:25
0283	S3-PD2-03-SE	Pandeiro 2	Samba-enredo	130	0:00:27
0284	S3-PD2-04-MA	Pandeiro 2	Marcha	120	0:00:32
0285	S3-PD2-05-CA	Pandeiro 2	Capoeira	65	0:00:30
0286	S3-PD2-06-VPA	Pandeiro 2	Virada (Partido-alto)	100	0:00:37
0287	S3-PD2-07-VSE	Pandeiro 2	Virada (Samba-enredo)	130	0:00:28
0288	S3-PD3-01-SA	Pandeiro 3	Samba	80	0:00:26
0289	S3-PD3-02-PA	Pandeiro 3	Partido-alto	100	0:00:25
0290	S3-PD3-03-SE	Pandeiro 3	Samba-enredo	130	0:00:27
0291	S3-PD3-04-MA	Pandeiro 3	Marcha	120	0:00:32
0292	S3-PD3-05-CA	Pandeiro 3	Capoeira	65	0:00:30
0293	S3-PD3-06-VPA	Pandeiro 3	Virada (Partido-alto)	100	0:00:37
0294	S3-PD3-07-VSE	Pandeiro 3	Virada (Samba-enredo)	130	0:00:27
0295	S3-PD5-01-SA	Pandeiro 5	Samba	80	0:00:26
0296	S3-PD5-02-PA	Pandeiro 5	Partido-alto	100	0:00:25
0297	S3-PD5-03-SE	Pandeiro 5	Samba-enredo	130	0:00:27
0298	S3-PD5-04-MA	Pandeiro 5	Marcha	120	0:00:32
0299	S3-PD5-05-CA	Pandeiro 5	Capoeira	65	0:00:30
0300	S3-PD5-06-VPA	Pandeiro 5	Virada (Partido-alto)	100	0:00:39

(Continued on the following page.)

Table A.4: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0301	S3-PD5-07-VSE	Pandeiro 5	Virada (Samba-enredo)	130	0:00:28
0302	S3-TB2-01-SA	Tamborim 2	Samba	80	0:00:26
0303	S3-TB2-02-PA	Tamborim 2	Partido-alto	100	0:00:25
0304	S3-TB2-03-SE	Tamborim 2	Samba-enredo	130	0:00:27
0305	S3-TB2-04-VPA	Tamborim 2	Virada (Partido-alto)	100	0:00:38
0306	S3-TB3-01-SE	Tamborim 3	Samba-enredo	130	0:00:27
0307	S3-TB3-02-VSE	Tamborim 3	Virada (Samba-enredo)	130	0:00:33
0308	S3-RR1-01-SA	Reco-reco 1	Samba	80	0:00:26
0309	S3-RR1-02-PA	Reco-reco 1	Partido-alto	100	0:00:25
0310	S3-RR1-03-SE	Reco-reco 1	Samba-enredo	130	0:00:27
0311	S3-RR1-04-MA	Reco-reco 1	Marcha	120	0:00:32
0312	S3-RR1-05-VPA	Reco-reco 1	Virada (Partido-alto)	100	0:00:39
0313	S3-RR1-06-OT	Reco-reco 1	Other	96	0:00:43
0314	S3-RR1-07-OT	Reco-reco 1	Other	106	0:00:39
0315	S3-RR1-08-VSE	Reco-reco 1	Virada (Samba-enredo)	130	0:00:26
0316	S3-CX1-01-SA	Caixa 1	Samba	80	0:00:26
0317	S3-CX1-02-PA	Caixa 1	Partido-alto	100	0:00:25
0318	S3-CX1-03-SE	Caixa 1	Samba-enredo	130	0:00:27
0319	S3-CX1-04-MA	Caixa 1	Marcha	120	0:00:32
0320	S3-CX1-05-MA	Caixa 1	Marcha	120	0:00:32

(Continued on the following page.)

Table A.4: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0321	S3-CX1-06-MA	Caixa 1	Marcha	120	0:00:32
0322	S3-CX1-07-VSE	Caixa 1	Virada (Samba-enredo)	130	0:00:30
0323	S3-RP2-01-SA	Repique 2	Samba	80	0:00:26
0324	S3-RP2-02-PA	Repique 2	Partido-alto	100	0:00:25
0325	S3-RP2-03-SE	Repique 2	Samba-enredo	130	0:00:27
0326	S3-RP2-04-VPA	Repique 2	Virada (Partido-alto)	100	0:00:38
0327	S3-RP2-05-VSE	Repique 2	Virada (Samba-enredo)	130	0:00:30
0328	S3-RP1-01-SA	Repique 1	Samba	80	0:00:26
0329	S3-RP1-02-PA	Repique 1	Partido-alto	100	0:00:25
0330	S3-RP1-03-SE	Repique 1	Samba-enredo	130	0:00:27
0331	S3-RP1-04-VPA	Repique 1	Virada (Partido-alto)	100	0:00:35
0332	S3-RP3-01-SA	Repique 3	Samba	80	0:00:26
0333	S3-RP3-02-PA	Repique 3	Partido-alto	100	0:00:25
0334	S3-RP3-03-SE	Repique 3	Samba-enredo	130	0:00:27
0335	S3-RP3-04-VPA	Repique 3	Virada (Partido-alto)	100	0:00:35
0336	S3-AG1-01-SA	Agogô 1	Samba	80	0:00:26
0337	S3-AG1-02-PA	Agogô 1	Partido-alto	100	0:00:25
0338	S3-AG1-03-SE	Agogô 1	Samba-enredo	130	0:00:27
0339	S3-AG1-04-VPA	Agogô 1	Virada (Partido-alto)	100	0:00:37
0340	S3-SK1-01-SA	Chocalho 1	Samba	80	0:00:26

(Continued on the following page.)

Table A.4: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0341	S3-SK1-02-PA	Chocalho 1	Partido-alto	100	0:00:25
0342	S3-SK1-03-SE	Chocalho 1	Samba-enredo	130	0:00:27
0343	S3-SK1-04-MA	Chocalho 1	Marcha	120	0:00:32
0344	S3-SK2-01-SA	Chocalho 2	Samba	80	0:00:26
0345	S3-SK2-02-PA	Chocalho 2	Partido-alto	100	0:00:25
0346	S3-SK2-03-SE	Chocalho 2	Samba-enredo	130	0:00:27
0347	S3-SK2-04-MA	Chocalho 2	Marcha	120	0:00:32
0348	S3-TT1-01-SA	Tantã 1	Samba	80	0:00:26
0349	S3-TT1-02-PA	Tantã 1	Partido-alto	100	0:00:25
0350	S3-TT1-03-SE	Tantã 1	Samba-enredo	130	0:00:27
0351	S3-TT1-04-VPA	Tantã 1	Virada (Partido-alto)	100	0:00:37
0352	S3-TT1-05-VSE	Tantã 1	Virada (Samba-enredo)	130	0:00:31
0353	S3-TT4-01-SA	Tantã 4	Samba	80	0:00:26
0354	S3-TT4-02-PA	Tantã 4	Partido-alto	100	0:00:25
0355	S3-TT4-03-SE	Tantã 4	Samba-enredo	130	0:00:27
0356	S3-TT4-04-MA	Tantã 4	Marcha	120	0:00:32
0357	S3-TT4-05-VSE	Tantã 4	Virada (Samba-enredo)	130	0:00:28
0358	S3-SU3-01-SA	Surdo 3	Samba	80	0:00:26
0359	S3-SU3-02-PA	Surdo 3	Partido-alto	100	0:00:26
0360	S3-SU3-03-SE	Surdo 3	Samba-enredo	130	0:00:27

(Continued on the following page.)

Table A.4: (Continued from previous page.)

GID#	Filename	Instrument	Style	Tempo (bpm)	Duration
0361	S3-SU3-04-MA	Surdo 3	Marcha	120	0:00:32
0362	S3-SU3-05-VPA	Surdo 3	Virada (Partido-alto)	100	0:00:39
0363	S3-SU3-06-SA	Surdo 3	Samba	80	0:00:28
0364	S3-SU3-07-PA	Surdo 3	Partido-alto	100	0:00:25
0365	S3-SU3-08-SE	Surdo 3	Samba-enredo	130	0:00:27
0366	S3-SU3-09-VPA	Surdo 3	Virada (Partido-alto)	100	0:00:39
0367	S3-SU3-10-SE	Surdo 3	Samba-enredo	130	0:00:27

Appendix B

Taxonomy of Musical Instruments

In order to clarify the terminology regarding instruments and sound production that appear in this thesis, we here give a succinct account of the developments in the field of organology, the science of the description and classification of musical instruments, with particular interest in late-eighteenth- and early-nineteenth-century European taxonomical systems. For a more in-depth look on these ideas and their evolutions, we refer the reader to [348].

Instruments around the world can be classified in several distinct ways, which are conditioned by sociocultural, religious, philosophical, and technological aspects among others. These classification systems usually distinguish instruments based on one or more characteristics, e.g., physical features (material, structure, sound quality and range), playing method, location and usage/function [3]. Examples of traditional division systems include: the Chinese material-based *bāyīn* (lit. "eight sounds") [3], which was developed during the Western Zhou dynasty (1046-771 BC) and groups instruments into eight classes[1] according to their main material component (clay, gourd, silk, leather, metal, stone, wood, and bamboo); and the ancient Indian system that is described in the *Nāṭyaśāstra* (a performing arts treatise dated between the first century BC and the third century AD), in which instruments are classified with regard to four acoustic principles [349]: "stretched" (strings), "hollow" (blown instruments), "covered" (skin-covered drums), and "solid" (bells, gongs, cymbals, rattles and other instruments).

In European and other East Asian traditions, instruments are classically divided into three sections — winds, strings, and percussion —, with many different proposals for the segmentation of each section (e.g., plucked/bowed strings, woodwinds/brass, pitched/unpitched percussion). This three-class structure, while sufficient

[1] Scholars believe that the number eight was deliberately chosen for the set of instrument classes in order to match a certain cosmological worldview [3], since this number is considered particularly auspicious in traditional Chinese numerology. It appears, for example, in Buddhism (e.g., the eight winds, one for each cardinal and ordinal points in a compass), and in the Taoist philosophy, where the eight trigrams (*bāguà*) represent different fundamental principles of reality.

in several settings (in particular, the musical practice of those cultures), contains an intrinsic inconsistency: the main division process deals with two different acoustic principles. For winds and strings, instruments are grouped according to the common vibrating substance (an air column[2] or the strings themselves), where the excitation method is the main criterium for the percussion class [3]. Also, some instruments cannot be satisfactorily classified through this system. For example, the celesta[3] is grouped together with drums in the set of percussion instruments [350], and the pianoforte could be similarly classified. However, the harpsichord, closely related to the piano, might fall into the strings category alongside the guitar, due to the plucking action of its playing.

At the end of the nineteenth century, a different classification scheme was devised by MAHILLON [351]; during the period he was entrusted with indexing the instruments in the museum of the *Conservatoire Royal de Musique*, in Brussels. Mahillon took special care in establishing a system capable of describing instruments both autochthonous and exotic to Europe and arrived at a structure much akin to that of the ancient Indian system, of which the curator was probably aware. Thus, by observing the nature of the vibrating bodies used as sound sources, he separated instruments into four classes [351]:

(I) *instruments autophones*, where sound is maintained by the vibrations of the instruments' bodies themselves, given that their rigidity and elasticity allow for periodic vibrations;

(II) *instruments à membranes*, where sound is created through the vibrations of stretched membranes;

(III) *instruments à vent*, where sound is produced by the vibration of an air column, which is excited by the interaction of the air flow with certain parts of the instrument;

(IV) and *instruments à cordes*, where sound originates from the vibration of tight strings.

In this system, each class was subdivided into families according to the playing action (e.g., friction, percussion) or, in the case of class (III), to the presence of special structural parts (e.g., reeds, mouthpieces, air reservoirs). Families were then further refined into species and sometimes into variations within a same species. In later

[2] We can also consider that the term "wind" (blowing) refers to the excitation method of wind instruments [350]. Be that as it may, this classic Western classification scheme is still governed by two different principles, since "strings" and "percussion" describe distinct aspects of sound production: the sound source and the excitation, respectively.

[3] The celesta is an instrument with the form of an upright piano. The playing action of a celesta causes small felt hammers to strike metal plates, which are suspended over wooden resonators [3].

editions of his catalog, Mahillon replaced these biology-inspired terms — "family", "species", and "variations" — for rather generic ones ("branches", "sections", and "subsections", respectively) to save "family" for those groups of instruments sharing a similar build, as recommended by François-Auguste Gevaert, then director of the *Conservatoire*.

About three decades later, HORNBOSTEL and SACHS [350], two musicologists from Austria and Germany, respectively, constructed a different four-class system. They presented in this seminal article a few of the inconsistencies in the classification scheme proposed in Mahillon's *Catalogue*, in particular regarding some choices for the subdivisions within each class. They observed that, as the museum's collection expanded, Mahillon was faced with new acquisitions that were hard to classify (especially non-European instruments) and had to accordingly make alterations in his scheme. One of the inconsistencies pointed out in [350] was the separation of autophones into instruments of untuned or tuned pitch, which ignores the fact that no pure untuned or tuned pitch instrument exists and leads to a subjective classification of the entire range of instruments that produce sound combinations in between these two extremes. Another set of discordances came from the persistent application (by Mahillon) of the division process according to playing action, which, despite being praised as highly logical by Hornbostel and Sachs, gave rise to "dubious solutions" for class (IV) of strings instruments. For example, in a similar manner to the ambiguities of the three-class system that we presented earlier, Mahillon ended up putting the pianoforte (and the clavichord, another of its relatives) in the same section, while the harpsichord remained in an entire different branch.

Striving for uniformity in intraclass divisions and aiming at the universality of the classification as a whole, Hornbostel and Sachs started fresh, with new developments and detailing of Mahillon's four original classes, which they renamed as: idiophones,[4] membranophones, aerophones, and cordophones. Unlike Mahillon, however, they refrained from naming the subsequent levels ("branches", "sections", etc.) of their ranking, for which two reasons were given. First, their system defined a larger number of levels, and the division criteria at each level followed group-specific (either morphological or playing-related) principles: uniform in a given group, but not directly comparable between different groups on the same level. More importantly, they were aware that the system might need improvements and amendments, in such way that a fixed nomenclature was inadvisable. Instead, the Hornbostel–Sachs system was devised with a decimal notation: the main classes are indicated by a single digit (from one to four) and each subsequent level is expressed by appending

[4] In a previously published glossary [352], Sachs expressed the preference for the term "idiophone" in place of "autophone", which, as the author stated, can confuse the reader, particularly the laymen, into thinking that it signifies an automatic (self-playing) instrument. The prefix "idio-" comes from the Greek *ídios* meaning "own", "distinct", "specific".

another digit to the right, adding a decimal dot at every three digits. For example, the Brazilian *tamborim* (see Figure 2.16f) can be described, down to the lowest level, as 211.311, for it is a membranophone (2) in which the membrane is struck (21) directly (211) and, furthermore, it has the form of a frame drum (211.3), without a handle (211.31), and with a single skin (211.311). This codification was influenced by the Dewey Decimal Classification, which is commonly used in libraries.

With this decimal system, the Hornbostel–Sachs classification provides great versatility in the description of instruments. First, any given code number in the system does not represent a single instrument, but an entire group of instruments that share the same characteristics.[5] Second, it is not required to exhaust all the levels in the classification, i.e., we can represent a given instrument with any desired amount of detail. As the authors showed, one can even omit certain figures (replacing them by wildcards) to indicate generic "supergroups" that bring together instrument groups separated by a few characteristics. It is also possible or combine different codes (with a "+" sign) to represent signals composed of parts coming from different groups. For example, the *pandeiro* (see Figure 2.16d) is entered as 211.311+112.122, because its shape follows that of a *tamborim* combined with jingles, which are indirectly struck. Finally, common characteristics among all divisions of a class (e.g., the playing method of chordophones) can be appended to the original code as suffixes after a dash. This solves the separation of piano (314.122–4–8) and harpsichord (314.122–6–8) we mentioned before, for example, since both cordophones have the form of a board zither with resonator box (314.122), but are further specified by the suffixes (8, for keyboard; 4, for hammers; and 6, for plectra).

The Hornbostel–Sachs system was met with a mixed reception, and many scholars (including the authors) have latter suggested modifications and revisions. One example is the introduction of a class of instruments known as electrophones, i.e., instruments that generate sound by electric/electronic means [164]. Other classification systems were also proposed, but the Hornbostel–Sachs system remains today the organological paradigm most referenced by researchers and people that work with musical instruments [353].

[5] This also holds for the elements in the *Catalogue* [351].